UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Development and Deployment of AI-enabled Digital Health Interventions in Critical Care Settings, with Applications to Sepsis Management

Permalink

https://escholarship.org/uc/item/2p3918kj

Author

Boussina, Aaron

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Development and Deployment of AI-enabled Digital Health Interventions in Critical Care Settings, with Applications to Sepsis Management

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics

by

Aaron Boussina

Committee in charge:

Professor Michael Hogarth, Chair
Professor Shamim Nemati, Co-Chair
Professor Eliah Aronoff-Spencer
Professor Robert El-Kareh
Professor Gabriel Wardi
Professor Pengtao Xie

2024

Copyright

Aaron Boussina, 2024

All rights reserved.

The Dissertation of Aaron Boussina is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To my beloved Kya, who proved you're never too old to learn new tricks.

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE....iii

DEDICATION....iv

TABLE OF CONTENTS....v

LIST OF FIGURES....vii

LIST OF TABLES....ix

LIST OF ABBREVIATIONS....xi

ACKNOWLEDGEMENTS....xiii

VITA....xvi

ABSTRACT OF THE DISSERTATION....xviii

INTRODUCTION....1

Chapter 1 Development & Deployment of a Real-time Healthcare Predictive Analytics Platform....9

Chapter 2 Impact of a deep learning sepsis prediction model on quality of care and survival....21

Chapter 3 Quantifying Healthcare Provider Perceptions of a Novel Deep Learning Algorithm to Predict Sepsis: Electronic Survey....68

Chapter 4 Representation Learning and Spectral Clustering for the Development and External Validation of Dynamic Sepsis Phenotypes: Observational Cohort Study....80

Chapter 5 The Impact of Laboratory Data Missingness on Sepsis Diagnosis Timeliness....121

Chapter 6 Machine learning based DNA melt curve profiling enables automated novel genotype detection....133

Chapter 7 Large Language Models Enable More Efficient Reporting of Hospital Quality Measures....163

CONCLUSIONS AND FUTURE WORK..... 185

SUMMARY OF CONTRIBUTIONS..... 185

REFERENCES..... 189

LIST OF FIGURES

Figure 1.1: Architecture of the predictive analytics platform..... 11
Figure 1.2: AWS implementation of the analytics platform..... 12
Figure 1.3: Dashboard of real-time sepsis risk scores for silent-trial chart reviews..... 15
Figure 1.4: Median values and 95% confidence intervals of input heart rate feature and output COMPOSER risk score relative to the training dataset. Trends of conformal rejection rate and PPV over time.....19
Figure 2.1: Acknowledgements to Each COMPOSER Best Practice Advisory alert from December 2022 until April 2023.....24
Figure 2.2: Plots of the causal impact analysis using a Bayesian structural time-series model.....28
Figure 2.3: Implementation of the COMPOSER BPA was significantly associated with an increase in sepsis bundle compliance.....29
Figure 2.4: COMPOSER Best Practice Advisory.....36
Figure S2.1: System Diagram of the Deep Learning Platform.....43
Figure S2.2: Examples of monitoring of Input features and performance outputs from the COMPOSER data quality dashboard.....45
Figure S2.3: Graphical model of the state-space model, adapted from Brodersen, et al.98.....47
Figure S2.4: Residual quantile-quantile (QQ) plot for the mortality Bayesian structural time-series model.. 49
Figure S2.5: Boxplot of autocorrelation of residuals for the mortality Bayesian structural time-series model.....50
Figure S2.6: Causal impact analysis of COMPOSER Best Practice Advisory on sepsis mortality at the La Jolla Emergency Department..... 51
Figure S2.7: Boxplot of autocorrelation of residuals for the mortality Bayesian structural time-series model for sepsis mortality at the La Jolla Emergency Department..... 52
Figure S2.8. Causal impact analysis of COMPOSER Best Practice on sepsis mortality at the Hillcrest Emergency Department.....53
Figure S2.9: Boxplot of autocorrelation of residuals for the mortality Bayesian structural time-series model for sepsis mortality at the Hillcrest Emergency Department..... 54
Figure S2.10: Causal impact analysis of COMPOSER Best Practice Advisory on sepsis bundle compliance rate at the La Jolla Emergency Department.....55
Figure S2.11: Causal impact analysis of COMPOSER Best Practice Advisory on sepsis bundle compliance rate at the Hillcrest Emergency Department.....56
Figure S2.12: Blood cultures prior to antibiotics compliance rate.....57
Figure S2.13: Rate of antibiotics administered within 24 hours prior and 3 hours after severe sepsis onset.. 58
Figure S2.14: Rate of lactate measured within 6 hours prior and 3 hours after severe sepsis onset.....59
Figure S2.15. Rate of repeat lactate measured within 6 hours after severe sepsis onset if initial lactate is elevated..... 60
Figure S2.16: Rate of administration of vasoactive medications within 6 hours of septic shock.....61

Figure S2.17: Rate of administration of 30cc/kg of fluids within 3 hours of presentation of septic shock or hypotension..... 62
Figure S2.18: Causal impact analysis of change in sequential organ failure assessment (SOFA) score within 72 hours of sepsis onset at the La Jolla Emergency Department..... 63
Figure S2.19: Causal impact analysis of change in sequential organ failure assessment (SOFA) score within 72 hours of sepsis onset at the Hillcrest Emergency Department..... 64
Figure S2.20: Causal impact analysis of sepsis intensive care unit (ICU) admissions from the emergency department..... 65
Figure S2.21: Causal impact analysis of intensive care unit (ICU)-free days..... 66
Figure S2.22: Temporal trends of covariates used in the causal impact analyses..... 67
Figure 3.1: User interface for physicians and APPs..... 70
Figure 3.2: Full Survey Results..... 73
Figure 3.3: Survey Results from Providers by Experience Level..... 74
Figure 3.4: Key Survey Results by Provider Type..... 75
Figure 3.5: Key Survey Results from Providers by Experience Level..... 76
Figure 4.1: Markov phenotype transitions..... 91
Figure S4.1: Consensus matrices at k=2-5 for consensus spectral clustering on the representations of a feed forward neural network trained to predict the onset of sepsis..... 100
Figure S4.2: Cumulative density function (CDF) for consensus spectral clustering on the representations of a feed forward neural network trained to predict the onset of sepsis..... 101
Figure S4.3: Consensus matrices at k=2-5 for consensus k-means clustering on normalized EHR data.. 102
Figure S4.4: Cumulative density function for consensus k-means clustering on normalized EHR data... 103
Figure S4.5: Phenotype Chord Diagram..... 104
Figure 5.1: Graphical representation of CMS criteria for sepsis over time..... 123
Figure 5.2: Schematic of the AIVIS system..... 125
Figure 6.1: Workflow for ML novelty detection..... 140
Figure 6.2: Overview of dHRM datasets..... 141
Figure 6.3: Leave-one-species-out cross validation to determine probability threshold..... 143
Figure 6.4: Accumulating the leave-one-group-out (LOGO) experiments results in a ‘practical’ threshold.. 144
Figure 6.5: Summary of LOGO novelty detection results..... 144
Figure 6.7: Summary of practical threshold validation results..... 147
Figure S6.1. Data preparation..... 152
Figure S6.2. Overview of experimentally obtained and synthesized melt curves..... 153
Figure S6.3. Overview of all synthesized melts. Residuals to the mean from each experimental melt curve are applied to the uMelts..... 154
Figure S6.4: Correlation between melt distance and patristic distance..... 155
Figure S6.6: Probability calibration results..... 160
Figure 7.1: System architecture for automation of hospital quality measures..... 166

LIST OF TABLES

Table 1.1: System adjustments made following real-time case review to improve clinical utility..... 16
Table 1.2: Demographics and clinical characteristics of patients processed by the real-time analytics platform..... 18
Table 2.1: Demographics and baseline characteristics of septic patients before and after COMPOSER.... 23
Table 2.2: Observed outcomes in the pre-intervention period, the expected counterfactual values from causal impact analysis, and the actual post-intervention values..... 26
Table 2.3: Associations Between BPA Response and Antibiotics Timing..... 31
Supplementary Table 2.1: Demographics and baseline characteristics of septic patients at the La Jolla (LJ) and Hillcrest (HC) emergency departments..... 48
Table 4.1: Demographics, clinical characteristics, and therapies of patients with sepsis..... 87
Table 4.2: Characteristics and therapeutics of the 4 sepsis phenotypes at the development site..... 89
Table 4.3: Odds ratios [95% CI] (P values) of interventions and physiological factors for predicting phenotype transition in the UCSDa cohort..... 92
Table 4.4: Odds ratios [95% CI] (P values) of interventions and physiological factors for predicting phenotype transition in the validation cohort..... 93
Table S4.1: List of Clinical Variables..... 105
Table S4.2: Characteristics and Therapeutics of the 4 Sepsis Phenotypes at the Validation Site..... 106
Table S4.3: Characteristics and Therapeutics of the 4 Sepsis Phenotypes at the Development Site (Hour 6) 107
Table S4.4: Characteristics and Therapeutics of the 4 Sepsis Phenotypes at the Validation Site (Hour 6)..... 109
Table S4.5: Performance (AUC) of Assigning Cluster Membership at ED Triage + 3 Hours Using Missing Value Flags vs True Clinical Values..... 110
Table S4.6: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCSD Cohort (Cluster 1)..... 111
Table S4.7: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCSD Cohort (Cluster 2)..... 112
Table S4.8: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCSD Cohort (Cluster 3)..... 113
Table S4.11: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCI Validation Cohort (Cluster 2)..... 116
Table S4.12: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCI Validation Cohort (Cluster 3)..... 117
Table S4.13: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCI Validation Cohort (Cluster 4)..... 118
Table S4.14: Odds Ratios (p-values) of Interventions and Physiological Factors for Predicting SOFA Group Transitions in the UCSD Cohort..... 119
Table S4.15: Odds Ratios (p-values) of Interventions and Physiological Factors for Predicting SOFA Group Transitions in the UCI Cohort..... 120

Table 5.1: CMS laboratory missingness (out of 5 labs) in patients with sepsis where organ failure occurred before and after Tsuspicion..... 126
Supplementary Table 5.1: Patient characteristics..... 130
Supplementary Table 5.2: Missingness in all CMS laboratory components in ED and Wards in the 6 hours prior to Tsuspicion..... 131
Supplementary Table 5.3: Missingness in contributory CMS laboratory components in ED and Wards that contributed to organ dysfunction in the 6 hours prior to TOF..... 132
Table 6.1: ML methods overview and classification results..... 142
Table 6.2: Overview of best results..... 148
Table S6.1. Overview of species..... 160
Table S6.2. C-statistics for patristic distance regression using random forest with 500 trees..... 162
Table 7.1: Study cohort demographics and SEP-1 measure results..... 170
Table 7.2: LLM system agreement with manual abstraction for SEP-1..... 171
Table 7.3: Physician review of discrepant cases..... 174
Table S7.1. Prompts for SEP-1 questions..... 179
Table S7.1. Prompts for SEP-1 questions (Continued)..... 180
Table S7.1. Prompts for SEP-1 questions (Continued)..... 181
Table S7.1. Prompts for SEP-1 questions (Continued)..... 182
Table S7.2. Example abstraction (deidentified)..... 183
Table S7.3. Example Errors & Hallucinations..... 184

LIST OF ABBREVIATIONS

ADT	Admit, Discharge, Transfer
AIVIS	AI Vigilant Information Seeking
AKI	Acute Kidney Injury
API	Application Programming Interface
APP	Advanced Practice Providers
ASG	Auto Scaling Group
AU(RO)C	Area Under the (Receiver Operating characteristic) Curve
AWS	Amazon Web Services
BPA	Best Practice Advisory
CCI	Charlson Comorbidity Index
CDS	Clinical Decision Support
CMS	Center for Medicare and Medicaid Services
COMPOSER	COnformal Multidimensional Prediction Of SEpsis Risk
CSN	Contact Serial Number
dHRM	digital High Resolution Melting
ED	Emergency Department
EHR	Electronic Health Record
EPIS	Exploration, Preparation, Implementation, Sustainment
ESS	Epic Sepsis Score
FHIR	Fast Healthcare Interoperability Resources
HIPAA	Health Insurance Portability and Accountability Act
ICU	Intensive Care Unit

INR	International Normalized Ratio
IQR	Interquartile Range
IRB	Institutional Review Board
JSON	JavaScript Object Notation
LLM	Large Language Models
NLB	Network Load Balancer
PCCP	Predetermined Change Control Plan
PCR	Polymerase Chain Reaction
PPV	Positive Predictive Value
PTT	Partial Thromboplastin Time
QQ	Quantile-quantile
RAG	Retrieval Augmented Generation
RDS	Relational Database Service
RLHF	Reinforcement Learning with Human Feedback
SD	Standard Deviation
SEP-1	Measure for Severe Sepsis and Septic Shock
SIRS	Systemic Inflammatory Response Syndrome
SOFA	Sequential Organ Failure Assessment
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
UCSD	University of California San Diego
UCI	University of California Irvine
VPC	Virtual Private Cloud

ACKNOWLEDGEMENTS

Thank you, Professor **Shamim Nemati**, for your guidance, wisdom, creativity, and trust. When I first met you nearly four years ago, I became infected with your excitement and vision for the future of healthcare. I decided then and there I needed to leave my career and learn everything that I could from you. I can't thank you enough for taking that chance on me and advising a student 7 years out of undergrad whose only machine learning experience was from Coursera. Over the years I have admired your subversiveness to the status quo, your boundless ingenuity, and your persistent focus on improving the lives of patients everywhere. You've pushed me to think critically and deeply, to synthesize concepts from disparate domains, and to conjure novelty from entropy. For that, I am sincerely grateful.

I would also like to thank Professor **Mike Hogarth** for sharing his depth of knowledge and experience with me and generously supporting my curiosity and academic growth. I would like to thank Professor **Gabe Wardi** for his energy, enthusiasm, and positivity which made our collaborations fruitful and fun. I would like to thank Professor **Rob El-Kareh** & Professor **Amy Sitapati** for teaching me about informatics at the frontline of clinical environments and opening my eyes to the challenges and opportunities.

I am further grateful to the many wonderful mentors I have had the privilege of learning so much from, including Professor **Rose Yu**, Professor **Stephanie Fraley**, Dr. **Eli Aronoff-Spencer**, Dr. **Bob Owens**, Dr. **Atul Malhotra**, and Dr. **Chris Longhurst**.

I would also like to thank the wonderful clinical quality team, especially **Kim Quintero**, **DeLee Glasser**, **Nick Hilbert**, and **Chad VanDenBerg**, for welcoming me into their world and patiently teaching me about it.

Thank you to the incredible Nemati Lab, in particular Dr. **Supreeth Shashikumar**, Dr. **Jonathan Lam**, Dr. **Fatemeh Amrollahi**, and **Hayden Pour**, for your support and camaraderie throughout the years. And thank you to the Nemati Lab collaborators who have made this educational journey so enriching, including Dr. **Kristin Mekeel**, Dr. **Linda Awdishu**, and Professor **Matteo D'Antonio**.

Finally, thank you to my family and friends. To my mother Kathy, my father Touraj, my sister Tamar, and my love Tiffany for your foundational support that made this work possible. And thank you to the incredible Bioinformatics and Systems Biology students for creating an inclusive and fun community that has gifted me so many memories.

Chapter 1, in part, is a reprint of the material as it appears in "Development & Deployment of a Real-time Healthcare Predictive Analytics Platform". Boussina, A., Shashikumar, S., Amrollahi, F., Pour, H., Hogarth, M., & Nemati, S. In 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 1-4). IEEE. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in "Impact of a deep learning sepsis prediction model on quality of care and survival". Boussina, A., Shashikumar, S. P., Malhotra, A., Owens, R. L., El-Kareh, R., Longhurst, C. A., ... & Wardi, G. (2024) npj Digital Medicine, 7(1), 14. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, has been submitted for publication of the material as it may appear in JMIR AI as "Quantifying Healthcare Provider Perceptions of a Novel Deep Learning Algorithm to Predict Sepsis: Electronic Survey". Boussina, A., Ramesh, K., …, & Wardi, G. The dissertation author was the primary investigator and author of this paper along with co-first author Karthik Ramesh.

Chapter 4, in full, is a reprint of the material as it appears in “Representation Learning and Spectral Clustering for the Development and External Validation of Dynamic Sepsis Phenotypes: Observational Cohort Study”. Boussina, A., Wardi, G., Shashikumar, S. P., Malhotra, A., Zheng, K., & Nemati, S. (2023) Journal of Medical Internet Research, 25, e45614. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in part, has been submitted for publication of the material as it may appear in JAMIA Open as “The Impact of Laboratory Data Missingness on Sepsis Diagnosis Timeliness”. Boussina, A., Lam, J., …, & Nemati, S. The dissertation author was the primary investigator and author of this paper along with co-first author Dr. Jonathan Lam.

Chapter 6, in full, is a reprint of the material as it appears in “Machine learning based DNA melt curve profiling enables automated novel genotype detection”. Langouche, L., Obirieze, A. C., Sinha, M., Mack, H., Leineweber, W., ... & Fraley, S. I. (2024). BMC bioinformatics, 25(1), 185. The dissertation author was the primary investigator and author of this paper along with authors Dr. Lennart Langouche and Augustine C. Obirieze.

Chapter 7, in full, has been submitted for publication of the material as it may appear in NEJM AI as “Large Language Models Enable More Efficient Reporting of Hospital Quality Measure”. Boussina, A., Krishnamoorthy, R., …, & Nemati, S. The dissertation author was the primary investigator and author of this paper.

VITA

2014 Bachelor of Science in Chemical Engineering, University of California Berkeley

2017 Master of Arts in Biotechnology Enterprise and Entrepreneurship, Johns Hopkins University

2024 Master of Science in Computer Science, University of California San Diego

2024 Doctor of Philosophy in Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics, University of California San Diego

PROFESSIONAL APPOINTMENTS

2015 Operations Leadership Development Program, Genentech, Inc.

2017 Senior Process Development Engineer, Illumina, Inc.

2019 Statistical Programmer & Consultant, PharmaStat, LLC.

2021 Co-Founder, Healcisio, Inc.

PUBLICATIONS

Boussina, A., Shashikumar, S. P., Malhotra, A., Owens, R. L., El-Kareh, R., Longhurst, C. A., ... & Wardi, G. (2024). Impact of a deep learning sepsis prediction model on quality of care and survival. npj Digital Medicine, 7(1), 14.

Boussina, A., Wardi, G., Shashikumar, S. P., & Nemati, S. (2024, May). A Deep Learning Sepsis Prediction Model Improves Survival and Quality of Care. Poster session presented at the Symposium on Artificial Intelligence for Learning Health Systems (SAIL), Puerto Rico.

Boussina, A., Krishnamoorthy, R., …, & Nemati, S. (2024). Large Language Models Enable More Efficient Reporting of Hospital Quality Measures. Manuscript submitted for publication at NEJM AI.

Boussina, A., Ramesh, K., …, & Wardi, G. (2024). Quantifying Healthcare Provider Perceptions of a Novel Deep Learning Algorithm to Predict Sepsis: Electronic Survey. Manuscript submitted for publication at JMIR AI.

Boussina, A., Lam, J., …, & Nemati, S. (2024). The Impact of Laboratory Data Missingness on Sepsis Diagnosis Timeliness. Manuscript submitted for publication at JAMIA Open.

Boussina, A., Langouche, L., Obirieze, A. C., Sinha, M., Mack, H., Leineweber, W., ... & Fraley, S. I. (2024). Machine learning based DNA melt curve profiling enables automated novel genotype detection. BMC bioinformatics, 25(1), 185.

Boussina, A., Wardi, G., Shashikumar, S. P., Malhotra, A., Zheng, K., & Nemati, S. (2023). Representation Learning and Spectral Clustering for the Development and External Validation of Dynamic Sepsis Phenotypes: Observational Cohort Study. Journal of Medical Internet Research, 25, e45614.

Boussina, A., Shashikumar, S., Amrollahi, F., Pour, H., Hogarth, M., & Nemati, S. (2023, July). Development & Deployment of a Real-time Healthcare Predictive Analytics Platform. In 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 1-4). IEEE.

ABSTRACT OF THE DISSERTATION

Development and Deployment of AI-enabled Digital Health Interventions in Critical Care Settings, with Applications to Sepsis Management

by

Aaron Boussina

Doctor of Philosophy in Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics

University of California San Diego, 2024

Professor Michael Hogarth, Chair
Professor Shamim Nemati, Co-Chair

Sepsis remains a persistent and severe problem in healthcare, accounting for over 1.7 million adult cases and resulting in approximately 350,000 deaths annually in the United States alone. Predictive modeling and machine learning hold immense potential to assist with the early

recognition of sepsis and other critical illnesses. Yet despite the rapid advancements in healthcare technology and data analytics, there remains a considerable disconnect between the development of predictive algorithms and their practical application at the patient bedside. The reasons are multifaceted and include (1) technical barriers to effective electronic health record (EHR) integration, (2) challenges in validating the clinical utility of algorithms, (3) 'one-size-fits-all' approaches that fail to capture the relevant context from a heterogeneous and complex patient population, (4) low EHR data quality with high rates of missingness, and (5) delayed feedback loops from reportable quality measures to assess sustained improvement to quality of care. The aim of this dissertation is to address the aforementioned challenges. First, I designed a scalable predictive analytics platform that enables modular deployment of deep learning models, including large foundation models, directly onto the EHR. I then clinically validated and implemented a model for the early prediction of sepsis onto this platform. Finally, I improved the clinical utility of this platform for the management of sepsis through digital phenotyping, active sensing, endotyping with novel bioassays, and automated assessment of the Center for Medicare and Medicaid Services (CMS) Severe Sepsis and Septic Shock (SEP-1) measure.

INTRODUCTION

BACKGROUND

Sepsis remains a persistent and severe problem in healthcare, accounting for over 1.7 million adult cases and resulting in approximately 350,000 deaths annually in the United States alone.[1,2,3] Moreover, sepsis disproportionately affects vulnerable populations such as the elderly, immunocompromised individuals, and those with chronic health conditions.[4,5] It is well established that early intervention, in particular the timely administration of antibiotics, improves patient outcomes.[6,7,8,9,10] Yet, while the timely administration of antibiotics in sepsis cases is crucial for reducing mortality rates, an alarming proportion of cases experience delays in appropriate treatment.[9,11,12] Several factors contribute to these delays, including diagnostic uncertainty, workflow inefficiencies, and suboptimal utilization of clinical decision support systems.[13,14,15] For example, traditional diagnostic methods often rely on physiological markers, which may present ambiguously, leading to delays in diagnosis and treatment.[2,16] Additionally, healthcare facilities often face systemic bottlenecks such as busy emergency departments, shortage of healthcare staff, or complex patient cases, further extending the time to administration.[17,18]

Predictive modeling and machine learning hold immense potential to ameliorate these factors and assist with the early recognition of sepsis and other critical illnesses.[19,20,21,22,23] Yet despite the rapid advancements in healthcare technology and data analytics, there remains a considerable disconnect between the development of predictive algorithms and their practical application at the patient bedside.[24] The reasons are multifaceted and include (1) technical barriers to effective electronic health record (EHR) integration,[25,26] (2) challenges in validating

the clinical utility of algorithms,[27,28] (3) 'one-size-fits-all' approaches that fail to capture the relevant context from a heterogeneous and complex patient population,[29,30,31] (4) low EHR data quality with high rates of missingness,[32] and (5) delayed feedback loops from reportable quality measures to assess sustained improvement to quality of care.[33] The aim of this thesis is to address the aforementioned challenges by designing a scalable predictive analytics platform that enables modular deployment of deep learning models, including large foundation models, directly onto the EHR.[34] Moreover, I clinically validate and implement a model for the early prediction of sepsis onto this platform.[35] Finally, I improve the clinical utility of this platform for the management of sepsis through digital phenotyping, active sensing, endotyping with novel bioassays, and automated assessment of the Center for Medicare and Medicaid Services (CMS) Severe Sepsis and Septic Shock (SEP-1) measure.[36,37]

SIGNIFICANCE

While predictive analytics hold immense potential for transforming healthcare systems and improving patient outcomes, only a fraction of these algorithms are successfully deployed prospectively and even fewer are integrated into clinical practice.[24,38,39,40] One major reason for this is the large technical barrier to accessing EHR data in real-time and providing timely results back to clinicians.[25,26,33] The challenges are multifactorial and include security, interoperability, availability, and scalability. Existing clinical decision support (CDS) solutions tend to be EHR-vendor or hospital specific and have not been generalized to new institutions. Further, customized on-premise solutions to EHR integration are susceptible to system interruptions and are difficult to scale. With the rise of large foundation models with billions of parameters, these impediments are likely to become even more pronounced. Further, even when these technical

barriers are overcome, algorithms that are developed and tested in controlled research environments often fail to demonstrate the same efficacy in diverse real-world clinical settings.[27,28,41,42]

Sepsis, a life-threatening condition triggered by an uncontrolled response to infection, exemplifies the aforementioned issues. Studies have shown that earlier identification of patients with sepsis would result not only in improved mortality but also in substantial savings for institutions and third-party payers.[43] Despite major interest in strategies to relieve the morbidity and mortality of sepsis, the majority of algorithms designed to predict sepsis never make it to the bedside.[44] Older models designed to detect sepsis were largely based on clinical criteria (i.e. SIRS criteria, hypotension or combination of these). These models were associated with occasional improvement in quality metrics (i.e. increased rates of lactate orders or time to antibiotics), but did not improve patient-centered outcomes and had poor positive predictive value (PPV).[45,46,47] More recently, several studies have implemented sophisticated models at various hospitals showing benefit to patients. Shimaburuko et al. conducted a small randomized trial of 142 patients in the ICU using a machine-learning algorithm to predict severe sepsis and found a decrease in in-hospital mortality and length of stay in the intervention group, although this study was limited to patients either in the hospital wards or intensive care units.[48] Adams et al. recently provided a prospective analysis of the TREWS model at five hospital systems in which they demonstrated a significant decrease in mortality, organ failure and length of stay in hospitalized patients when the sepsis alert was confirmed by a provider.[21] While this study was not randomized, and suffers from confounders inherent in the study design, the data are compelling that proper attention to implementation may improve patient-centered outcomes in sepsis.[49] However, a commonly used predictive model, the Epic Sepsis Score (ESS), has not

demonstrated consistent improvement in patient-centered outcomes. Importantly, researchers at the University of Michigan highlighted a substantial drop in test characteristics (sensitivity, specificity, PPV) of the ESS at their institution from what was reported by Epic, as well as an unacceptably high rate of false positives.[50]

Despite some recent promise in achieving patient benefit with early sepsis warning systems, the performance of existing systems remains lackluster. The aforementioned TREWS system, for example, has a PPV of 27% at a sensitivity of 80%.[51] Closer examination of the model during clinical implementation phase reveals an even lower PPV of 17%, after excluding patients already on antibiotics. This is significant as a low PPV can result in alarm fatigue in which healthcare providers, overwhelmed by frequent false alarms, may start to ignore or override alerts, thus compromising patient safety.[52,53] One promising avenue of research is the use of modern deep learning techniques to improve predictive performance and reduce false alarms. Work from Shashikumar et al. has shown that out-of-distribution rejection on the representations from a neural network can reduce false alarms by 75%.[20] However, as of the time of writing, no patient-centered outcomes or provider perceptions have been reported for a deep learning model in the area of sepsis prediction.

Even a true positive alert for sepsis might not be clinically actionable and may further contribute to alarm fatigue. Sepsis is a complex syndrome with multiple underlying etiologies, and the presentation can vary significantly among patients. An alert indicating high sepsis risk may not necessarily specify the root cause, such as the source of infection or infectious agent, thus complicating targeted treatment interventions. The timing of the alert may also come at a stage where sepsis is already too advanced for some interventions to be effective, or conversely, too early to warrant aggressive treatment like vasopressors. The identification of clinical

phenotypes, or subtypes, early in sepsis may allow for more precision care as providers may tailor therapies (ie, quantity of fluid resuscitation or timing of vasoactive medications) based on a particular phenotype. Seymour et al described 4 novel distinct clinical phenotypes of sepsis that correlated with host response patterns and clinical outcomes.[31] However, phenotyping using unsupervised clustering on high-dimensional features is sensitive to noise, which complicates the generalizability of the associated phenotypes.[54-56] Further, the algorithmic approaches for early prediction and early phenotyping are only as good as the quality of data that is available and are adversely impacted by the presence of outliers and missingness.[20] To address the maxim of "garbage in, garbage out," there is potential for novel uncertainty-aware systems that can mimic the diagnostic process and seek out important data elements.[19] Simultaneously, there is potential for novel bioassays that could be requested by such a system to rapidly identify the causative pathogen and inform appropriate antimicrobial therapy. Blood cultures remain the existing diagnostic standard, but have a substantial time to positivity (typically 24-72 hours), a low sensitivity, and are routinely contaminated.[57] dHRM analysis, in which the melting profile from thousands of digital PCR reactions generates a set of fluorescence-temperature melting curves, offers an emerging alternative that enables rapid, effective, and economical post-amplification nucleic acid characterization for profiling DNA sequences.[58-62] However, at the time of writing, the ability to extract features from the melt curve signatures from this technique to detect known and novel pathogens in sepsis patients remains underexplored.

Aggravating the aforementioned implementation barriers for advanced analytics in sepsis management, it remains a challenge to rapidly assess the impact of these interventions on the quality of care. The SEP-1 bundle, which all hospitals are held accountable to, was introduced by CMS to improve patient care through completion of a protocolized bundle within 3-6 hours of

sepsis onset time.[63] Abstraction of this measure, however, involves a complicated and laborious 63-step process that places a heavy reporting burden on the hospital.[64,65] Further, measure results are generated too late (e.g. quarterly) and with too few samples to inform real-time quality improvement and support a learning health system.[66] Recently, LLMs have demonstrated exceptional performance in task understanding and completion.[67] GPT-4, for example, demonstrated human-level performance on standardized medical tests without task-specific fine-tuning.[68] As such, state-of-the-art advances in LLM methodologies including in-context learning, self-consistency, and reinforcement learning with human feedback (RLHF), may enable automation of SEP-1 reporting and continuous monitoring of an algorithm's effect on clinical processes.[69,70,71]

Given the substantial challenges in deploying predictive models, personalizing their recommendations, and measuring their clinical impact, this thesis addresses an unmet need for a clinically validated, uncertainty-aware sepsis AI platform that can account for the patient's phenotype and care context to generate clinically actionable recommendations and assess their impact on the quality of care.

SUMMARY OF CONTRIBUTIONS

In Chapter 1, we design a platform for predictive analytics in healthcare that is scalable, cloud-based, fault-tolerant, and capable of extracting and processing EHR data for any patient at UC San Diego Health at any time following admission and transferring results back into the EHR.[34] The system is built in a modular plug-and-play manner such that any number of predictive modules can be added. We further prospectively deploy the sepsis model from Shashikumar, et. al. onto the platform from Chapter 1 and evaluate performance.[20] Finally, we

validate the clinical utility of the model deployed in Chapter 1 and fine-tune the target population based on implementation feedback.

In Chapter 2, we clinically implement the model and evaluate patient-centered outcomes, including sepsis-related mortality and SEP-1 bundle compliance.[30,35]

In Chapter 3, we evaluate provider perceptions of the deep learning model deployed in Chapter 2. We conduct an electronic survey and explore the perceived utility of the alert as a function of provider type and years of experience.

In Chapter 4, we deepen the capabilities of the sepsis AI platform through digital phenotyping. We identify novel sepsis phenotypes that are prognostic, generalizable, and may enable personalized treatment recommendations. We demonstrate that representation learning allows for highly consistent phenotypes and that these phenotypes are a function of patient state and the treatments they receive. We further observe that the administration of a 30 mL/kg fluid bolus may be associated with worse outcomes in certain phenotypes, whereas prompt antimicrobial therapy is associated with improved outcomes.

In Chapter 5, we explore the timeliness of sepsis diagnosis as a function of data missingness. We demonstrate that incorporating active sensing into the sepsis AI platform may support earlier recognition of sepsis. We further deploy active sensing capabilities on top of the platform from Chapter 1.

In Chapter 6, we exhibit a novel bioassay for pathogen profiling as a candidate target for the active sensing feature in Chapter 5. We show that time-series modeling of digital High Resolution Melting (dHRM) curves enables accurate differentiation of pathogens commonly indicated in neonatal sepsis. We further show that these methods can be used for novelty detection of emerging pathogens.

In Chapter 7, we expand the functionality of the platform to enable automated assessment of quality of care per the SEP-1 bundle. We deploy an LLM-based system that ingests Fast Healthcare Interoperability Resources (FHIR) and outputs a completed SEP-1 abstraction. We demonstrate that LLMs with interoperable EHR data can accurately perform abstraction of the SEP-1 quality measure, and furthermore that open-source LLMs running on consumer-grade hardware are sufficiently capable.

Chapter 1 Development & Deployment of a Real-time Healthcare Predictive Analytics Platform

The deployment of predictive analytic algorithms that can safely and seamlessly integrate into existing healthcare workflows remains a significant challenge. Here, we present a scalable, cloud-based, fault-tolerant platform that is capable of extracting and processing electronic health record (EHR) data for any patient at any time following admission and transferring results back into the EHR. This platform has been successfully deployed within the UC San Diego Health system and utilizes interoperable data standards to enable portability.

Clinical relevance— This platform is currently hosting a deep learning model for the early prediction of sepsis that is operational in two emergency departments.

INTRODUCTION

Despite the rapid growth in the number of predictive models developed for healthcare applications, there has been a relative dearth of successful implementations into clinical practice.[38,39] One major reason for this is the large technical barrier to accessing EHR data in real-time and providing timely results back to clinicians.[25,26] The challenges are multifactorial and include security, interoperability, availability, and scalability. Existing clinical decision support (CDS) solutions tend to be EHR-vendor or hospital specific and have not been generalized to new institutions.[72] Further, customized on-premise solutions to EHR integration are susceptible to system interruptions and are difficult to scale. In this chapter, we present a secure, high-availability, cloud-based platform that can process EHR data on any patient within a hospital at any point during their admission. We demonstrate how this platform can close the CDS loop and provide realtime recommendations to clinicians natively within the EHR. We

further describe how process control tooling is leveraged to ensure system availability and model fidelity. Finally, we showcase the deployment of a deep learning model for the early prediction of sepsis onto this platform and into clinical practice.[20]

This chapter is distinct from prior publications in real-time healthcare analytics in the following ways: (1) It describes a platform for real-time predictions across all inpatient settings including Intensive Care Units (ICUs), Emergency Departments (EDs), and wards. (2) It utilizes collection of data elements beyond laboratory results, vital-signs, and demographics information, including problem-list items, procedures, and clinical notes. (3) It closes the CDS loop and enables outputs from predictive models to be written directly to the EHR. (4) It demonstrates the deployment of the platform into clinical practice.[73,74]

METHODS

Platform Architecture

Figure 1.1 shows the high-level architecture for the predictive analytics platform. Data are extracted from the EHR at routine intervals and subsequently preprocessed for model consumption. An Indications-for-Use module assesses the clinical context of the patient including where the patient is in their treatment timeline to determine whether the patient is included/excluded for CDS. The predictive model is then run and outputs are sent back to the EHR to provide clinicians with relevant recommendations. The system is built in a modular plug-and-play manner such that any number of predictive modules can be added.

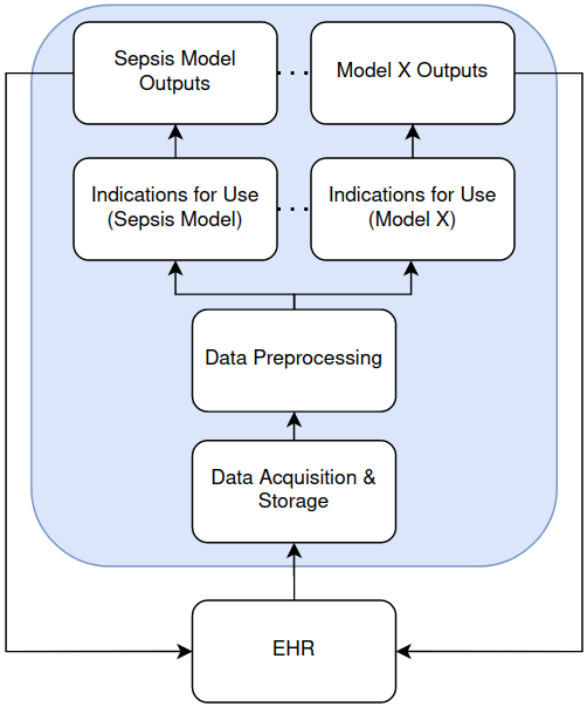


Figure 1.1: Architecture of the predictive analytics platform.

Cloud Implementation

Our predictive analytics platform is hosted within a HIPAA-compliant Amazon Web Services (AWS) environment (Figure 1.2). The environment is an isolated enclave with communications only permitted through whitelisted ports. The application layer is hosted on a single EC2 instance with data stored in a MySQL Relational Database Service (RDS). The EC2 instance is part of an Auto Scaling Group (ASG) that is connected to a Network Load Balancer (NLB). This configuration enables a new copy of the EC2 instance to be brought online immediately in the event of primary instance failure. Similarly, the RDS is configured for regular backup which enables switchover to the secondary database in the event of failure.

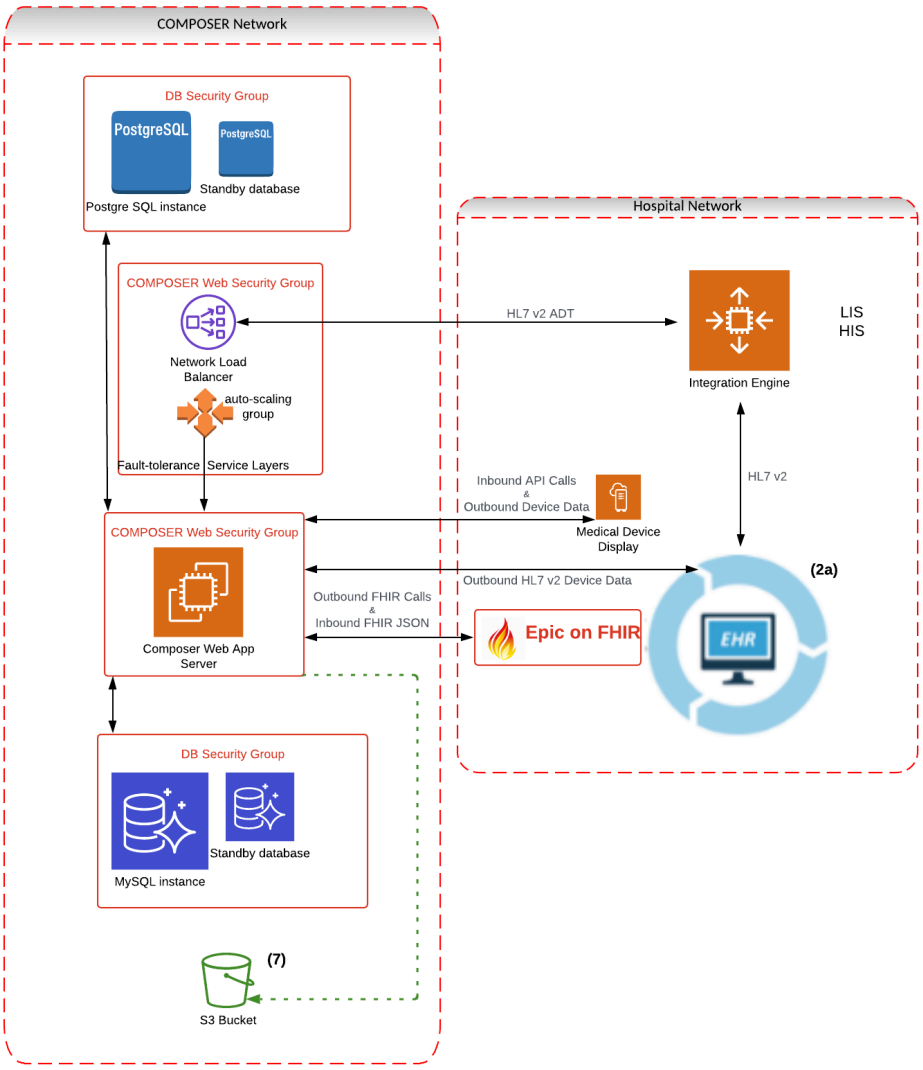


Figure 1.2: AWS implementation of the analytics platform.

For security, all system credentials such as private keys are stored within the AWS Secrets Manager and automatically updated on a routine schedule. For traceability, all system logs are preserved within S3 storage buckets. For portability, the architecture is captured within terraform scripts that automate the AWS build.

Data Pipeline

All active admitted patients within the healthcare system are identified from HL7v2 ADT messages.[75] All ADT messages are forwarded from the hospital's integration engine to MirthConnect software running within the application EC2 instance. The patient contact serial numbers (CSNs) from the PID segment are converted to patient FHIR IDs by calling the "Patient. Search" API from the hospital's Epic FHIR server. The platform application authenticates its requests to the FHIR server with OAuth 2.0 using a backend private key.

With the patient FHIR IDs retrieved for every admitted patient, the application makes regular calls to the FHIR server to retrieve updates to the Patient, ServiceRequest, Observation, MedicationRequest, Condition, and Procedure resources. The resources are returned as JSON bundles which are then parsed and preserved within the RDS as a condensed JSONB column containing all updated data for a patient within an elapsed timeframe.

The semi-structured data are then passed to the Data Preprocessing Module where they are converted into a structured format with columns for every feature and missing values imputed by a predefined sample-and-hold. The Data Preprocessing Module also enforces that the values are physiologically possible and not the result of inadvertent entry by applying upper and lower limits on the features.

Models are then directly deployed using these database tables as inputs. Model outputs such as the risk score and top contributing features are placed into OBX segments and an outbound HL7v2 message is constructed and sent back to the integration engine. The observation identifier (field OBX.3) is registered within the EHR allowing all model outputs to be filed to the flowsheet. EHR-native decision support, such as Best Practice Advisories (BPAs), then utilize these flowsheet items to generate clinician-facing alerts.

Model Deployment & Silent-Mode Evaluation

Using this platform, we deployed the COMPOSER deep learning model for the early prediction of sepsis described in Shashikumar et al.[20] The model was run in silent-mode evaluation over the course of 7 months in which sepsis risk scores were filed to patient flowsheets, but alerts were not displayed to clinicians. During this period, the performance of the model was evaluated and routine chart reviews with a panel of experts were conducted to assess the clinical utility of the silent alerts (Figure 1.3). COMPOSER alerts were reviewed by our team of physicians and data scientists, who then systematically underwent a chart review procedure to determine whether the algorithm's alert firings were appropriate (true positive, concern for sepsis), inappropriate (false positive, sepsis-look alike), or missed detection (false negative, failure to alert in case of sepsis). Chart review entailed reviewing vital signs and laboratory results to assess if SIRS/sepsis/septic shock criteria was met, medications to assess if the clinical team started antibiotics, progress notes to understand the clinical team's decision/thought process, imaging studies relevant to infection, and problem list to contextualize the patient's current presentation.

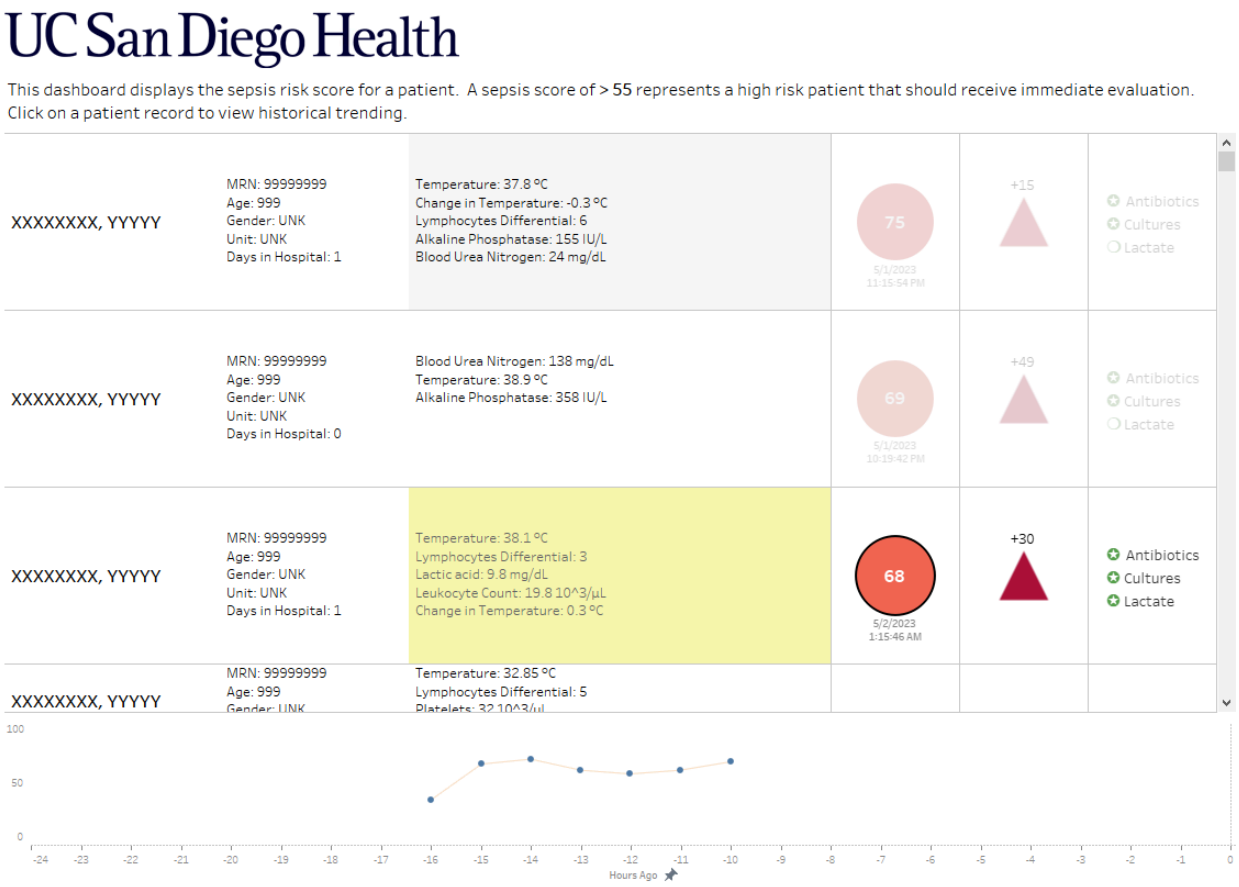


Figure 1.3: Dashboard of real-time sepsis risk scores for silent-trial chart reviews. Patient demographics, model top causes, the risk score, the change in risk score, and the completed elements of the SEP-1 bundle are displayed to the panel to facilitate review. Historical trending of the risk score is displayed in a bottom panel.

Once chart review was completed, the team of physicians came to a collective consensus about whether the algorithm's alert was appropriate. Alerts were categorized as appropriate (true positive) or inappropriate (false positive). For patients with a calculated sepsis score that did not meet threshold to fire an alert, the algorithm's lack of alert firing was categorized as appropriate (true negative) or inappropriate (false negative).

We made adjustments to the system based on the data from false positives and false negatives, including temporary alert suppression for post-operative patients (0-36 hour post-op

window), a delta logic component that monitors for changes in a patient's clinical status (based on calculated sepsis risk score), possible alert suppression for cancer/transplant/immunocompromised patients, and incorporation of medical history (i.e. transplant history, immunosuppression, history of cirrhosis) into COMPOSER (Table 1.1).

Table 1.1: System adjustments made following real-time case review to improve clinical utility

Problem	**Changes to Algorithm**	**Rationale**
Alerts for Patients with Sepsis Already Receiving Treatment	Suppression logic for patients already started on antibiotics	Patients quickly identified with sepsis and treated accordingly do not need alert; alert would not change management – excessive alert → alert fatigue
Redundant Alerts in Post-operative Patients	Alert suppression during 0-12 hour post-operative window	Minimize repeat firing for classic sepsis mimicker (minimize false positives → more judicious firing & higher PPV)
Repeat Alerts for Patients	Delta logic (> 10% change in sepsis risk score)	If patient's sepsis risk score is high but unchanged, don't want to resend alert; however, if sepsis risk score increases by > 10%, indicative that either current antibiotic regimen is providing inadequate coverage for infection, lack of source control, or new/brewing infection present

Process Control

To ensure high-availability of the platform, AWS Cloud-Watch alerts were created to notify the development team of any event that brought the system out of a state of control (e.g. if a service was unreachable). These CloudWatch alerts were integrated with PagerDuty to ensure 24/7 front-line support. Further, the platform was registered with healthcare IT within ServiceNow to enable end-users to escalate issues directly to the development team.

In addition to system interruptions, deployed models are at risk of model drift due to changes in the data distribution over time.[76,77,78] To ensure detection of possible model drift we implemented a quality dashboard that automatically tracks a model's inputs, outputs, and

performance. Specifically, the median values of measurement inputs and risk score outputs are monitored to ensure they don't pass the upper or lower quartiles from the training cohort. If any value falls outside of those limits, it is flagged for review by the development team. Similarly, model performance metrics such as the positive predictive value (PPV) and sensitivity are tracked on a weekly basis. Finally, the quality dashboard tracks the rate of rejection from conformal prediction.[20] Conformal prediction is a method for detecting out-of-distribution data within the lower dimensional representations of a neural network. Changes in rejection rates, therefore, are expected to correspond to changes in data distribution that affect the model's predictions.

RESULTS

Clinical Population

Table 1.2 shows the patient population processed by the platform from June 1st, 2022 to January 1st, 2023. During this 7-month period, 63,133 patients and 1,368,763 patient-hours were processed across 63 care units.

Table 1.2: Demographics and clinical characteristics of patients processed by the real-time analytics platform.

Number of Patients, N	63,133
Age, Mean (SD)	52.0 (19.08)
Males, N (%)	31,187 (49.4%)
Females, N (%)	31,860 (50.5%)
Unknown Gender, N (%)	86 (0.14%)
ED Patients, N	47,993
ICU Patients, N	3,565
Ward Patients, N	14,185
Length of Stay (hours), Median (IQR)	6.7 [4.03 - 8.47]

Sepsis Model Performance

Figure 1.4 shows outputs exported from the quality dashboard for the median values of a sample input feature and COMPOSER's sepsis risk score relative to the training set following clinical deployment on 2022-12-07. Also shown are the model's PPV and rate of conformal rejection. The dashboard demonstrates that the input feature distributions and model outputs had not drifted substantially post-deployment. Further, the model's performance in real-time using the analytics platform did not differ significantly from retrospective performance.

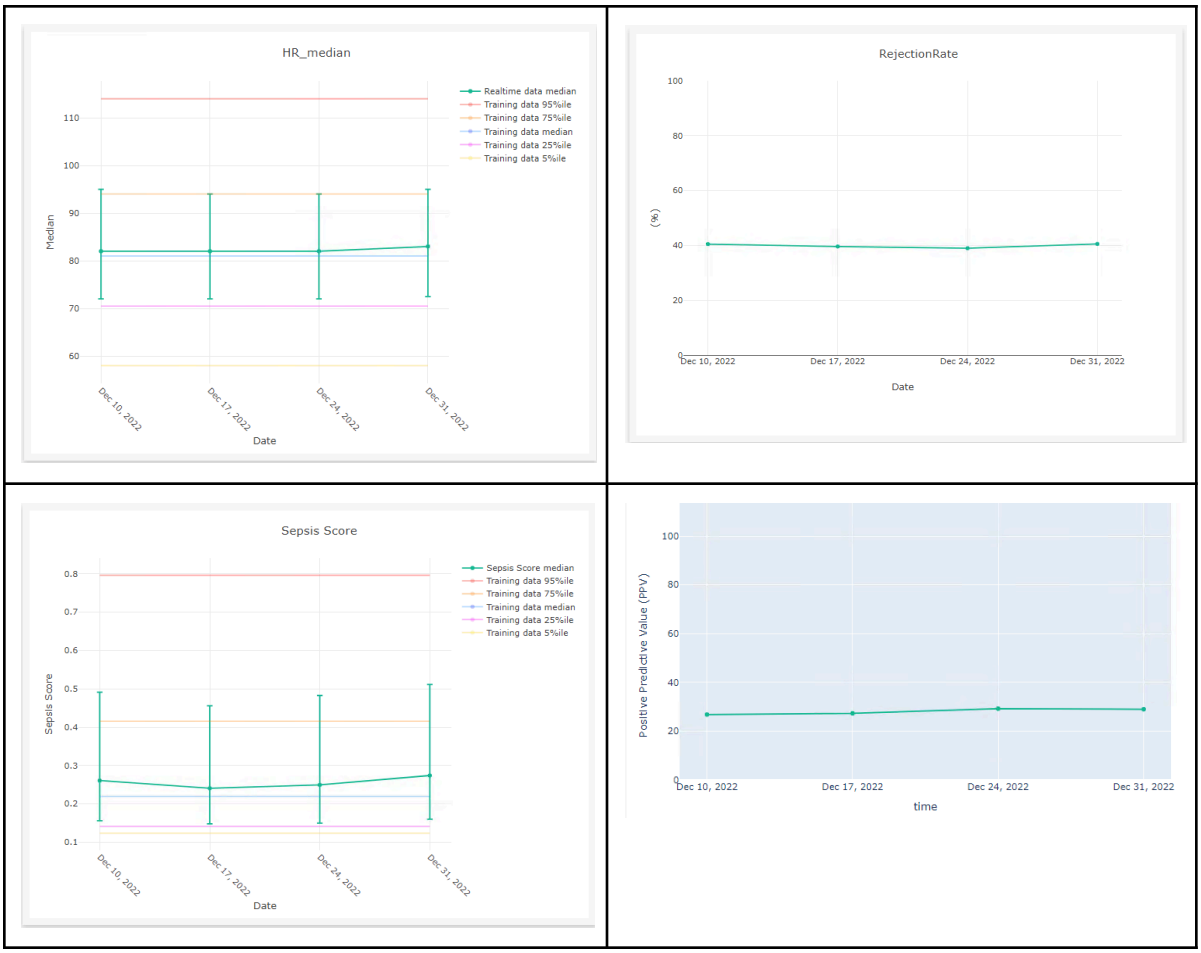


Figure 1.4: Median values and 95% confidence intervals of input heart rate feature and output COMPOSER risk score relative to the training dataset. Trends of conformal rejection rate and PPV over time.

System Availability

From June 1st, 2022 to January 1st, 2023 the platform experienced a total of 0 hours of system downtime and 28 hours of inter-connectivity interruptions. This corresponds to an overall uptime of 99.44%. The single largest instance of downtime (10 hours) was related to a significant update to the FHIR API protocol which resulted in a difference in version parity with the platform.

CONCLUSION

Using cloud architecture and interoperable data standards we have built a production-grade system to enable safe, rapid deployment of predictive analytics into the clinic. We have leveraged best-practices in software engineering and process control to ensure that the platform is sustainable and robust. We have designed the platform together with our clinical collaborators to ensure that model predictions are relevant and clinically actionable. This work aims to address the growing divide between the abundance of new deep learning models and the relative paucity of predictive models in clinical practice. While in this chapter we have only showcased a single model at a single institution, we have developed the system with portability and scalability in mind. We are currently developing prediction models for other clinical use cases on this platform and targeting additional institutions for deployment.

ACKNOWLEDGEMENTS

Chapter 1, in part, is a reprint of the material as it appears in "Development & Deployment of a Real-time Healthcare Predictive Analytics Platform". Boussina, A., Shashikumar, S., Amrollahi, F., Pour, H., Hogarth, M., & Nemati, S. In 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 1-4). IEEE. The dissertation author was the primary investigator and author of this paper.

Chapter 2 Impact of a deep learning sepsis prediction model on quality of care and survival

Sepsis remains a major cause of mortality and morbidity worldwide. Algorithms that assist with the early recognition of sepsis may improve outcomes, but relatively few studies have examined their impact on real-world patient outcomes. Our objective was to assess the impact of a deep-learning model (COMPOSER) for the early prediction of sepsis on patient outcomes. We completed a before-and-after quasi-experimental study at two distinct Emergency Departments (EDs) within the UC San Diego Health System. We included 6217 adult septic patients from 1/1/2021 through 4/30/2023. The exposure tested was a nurse-facing Best Practice Advisory (BPA) triggered by COMPOSER. In-hospital mortality, sepsis bundle compliance, 72-h change in sequential organ failure assessment (SOFA) score following sepsis onset, ICU-free days, and the number of ICU encounters were evaluated in the pre-intervention period (705 days) and the post-intervention period (145 days). The causal impact analysis was performed using a Bayesian structural time-series approach with confounder adjustments to assess the significance of the exposure at the 95% confidence level. The deployment of COMPOSER was significantly associated with a 1.9% absolute reduction (17% relative decrease) in in-hospital sepsis mortality (95% CI, 0.3%–3.5%), a 5.0% absolute increase (10% relative increase) in sepsis bundle compliance (95% CI, 2.4%–8.0%), and a 4% (95% CI, 1.1%–7.1%) reduction in 72-h SOFA change after sepsis onset in causal inference analysis. This study suggests that the deployment of COMPOSER for early prediction of sepsis was associated with a significant reduction in mortality and a significant increase in sepsis bundle compliance.

INTRODUCTION

Few sepsis algorithms have been rigorously tested at the bedside or evaluated with regard to patient outcomes.[21,48,79,80] Existing algorithms within electronic health records (EHRs) have demonstrated relatively poor positive predictive value (PPV) and may contribute to provider mistrust of predictive models.[50,81] Of note, false positive alerts from such models often lead to alarm fatigue and provider burnout/mistrust. COMPOSER was specifically designed to reduce false alarms by flagging outliers and out-of-distribution samples as indeterminate.[20]

Based on this conceptual framework, we integrated the COMPOSER algorithm in two emergency departments (ED) at UC San Diego (UCSD) Health via our EHR (Epic Systems, Verona, WI). We seek to test the hypothesis that our algorithm-based intervention was feasible in real-time and that the additional information would help to guide clinicians to earlier sepsis recognition and result in improved patient outcomes. To accomplish this goal, we conduct a quasi-experimental study in which we track outcomes before and after deployment, with historical control data to account for baseline acuity, comorbidities, seasonal effects, and secular trends over time.

RESULTS

During the study period, January 1st, 2021 through April 30th, 2023, 6,340 ED encounters met the Sepsis-3 consensus sepsis definition, of which 123 were excluded because they were transitioned to comfort measures before sepsis onset. The final study included 6217 patients, 5065 in the pre-intervention phase, and 1152 in the post-intervention phase. Table 2.1 shows baseline characteristics and summary statistics for the study cohort. Baseline characteristics from each emergency department are compared in Supplementary Table 2.1. Most septic patients exhibited some level of chronic comorbidity (median Elixhauser of 5) and the

median SOFA score at the time of sepsis was 2. We did not observe significant differences in the baseline characteristics between the pre-intervention cases and post-intervention cases.

Table 2.1: Demographics and baseline characteristics of septic patients before and after COMPOSER.

	Total	**Pre-Intervention**	**Post-Intervention**	***P* Value[a]**
Characteristic				
Number of Patients, N (%)	6,217 (100%)	5,065 (81.5%)	1,152 (18.5%)	-
Age, Mean (SD)	63 (17.1)	63 (17.0)	64 (17.3)	0.08
Sex, N (%)				
Male	3,592 (57.8%)	2,966 (58.6%)	626 (54.3%)	-
Female	2,625 (42.2%)	2,099 (41.4%)	526 (45.7%)	-
Race				
Asian	530 (8.5%)	404 (8%)	126 (10.9%)	-
Black or African American	639 (10.3%)	519 (10.2%)	120 (10.4%)	-
White	2,983 (48%)	2,440 (48.2%)	543 (47.1%)	-
Other[b]	2,065 (33.2%)	1,702 (33.6%)	363 (31.5%)	-
Ethnic Group				-
Hispanic/Latino	1,756 (28.2%)	1,449 (28.6%)	307 (26.6%)	-
Not Hispanic/Latino	4,461 (71.8%)	3,616 (71.4%)	845 (73.4%)	-
Organ Dysfunction				
Elixhauser Comorbidity Index, Median (IQR)	5 (0-13)	5 (0-13)	5 (0-14)	0.64
SOFA Score at Time of Sepsis, Median (IQR)	2 (1 - 3)	2 (1 - 3)	2 (1 - 3)	0.99
Lab Values				
Lactate at Time of Sepsis	2.4 (1.6 - 4.3)	2.4 (1.6 - 4.3)	2.4 (1.6 - 4.3)	0.76
Interventions				
Mechanical Ventilation, N (%)[c]	1,035 (16.6%)	849 (16.8%)	186 (16.1%)	0.64
Administration of Vasoactive Medications, N (%)[c]	424 (6.8%)	345 (6.8%)	79 (6.9%)	1.0

[a]P values for continuous variables are based on Kruskal-Wallis rank sum tests. P values for categorical variables are based on Pearson's chi-squared tests.

[b]Other race corresponds to Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, Other Race or Mixed Race, or Unknown.

COMPOSER alerts

During the post-intervention period, an average of 235 alerts were generated per month corresponding to 1.65 alerts per nurse per month. Alerts by acknowledgement reason are visualized in Fig. 2.1. The most common acknowledgement reason was "Will Notify MD Immediately" which comprised over half of all acknowledgement reasons. Only about 5.9% of BPAs were exited without acknowledgement and responses to the BPA remained consistent across the 5-month intervention period.

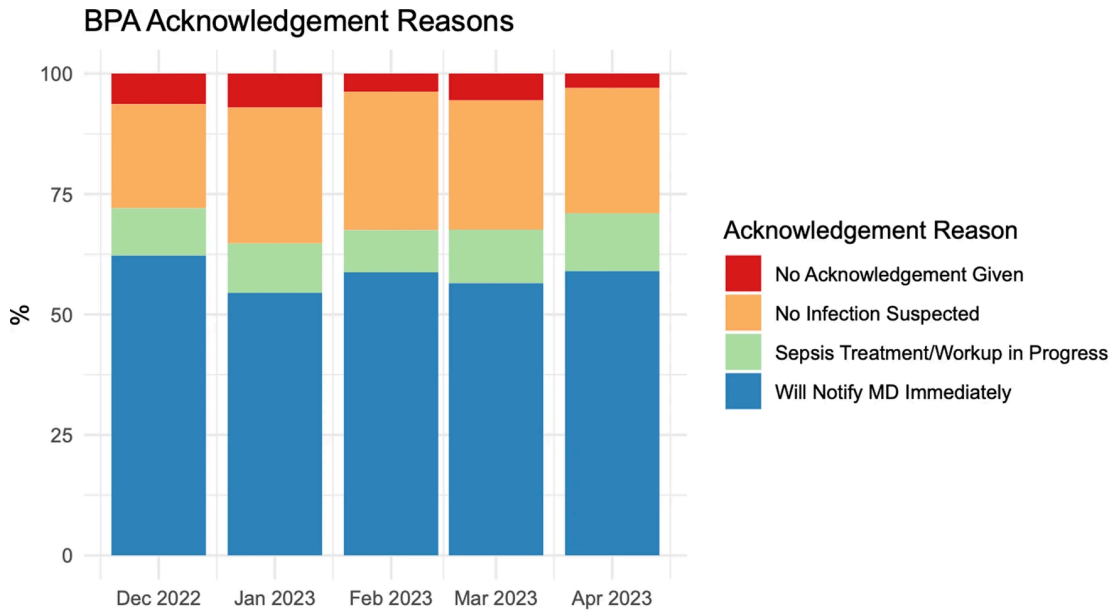


Figure 2.1: Acknowledgements to Each COMPOSER Best Practice Advisory alert from December 2022 until April 2023.

Interventions and patient outcomes

The results from causal impact analysis on our primary and secondary outcomes are summarized in Table 2.2. The observed in-hospital mortality rate and the corresponding predictions from the Bayesian structural time-series model are shown in Fig. 2.2a. The residual quantile-quantile and autocorrelation plots are described in Figs. S2.4 and S2.5. The average sepsis mortality rate during the post-intervention period was 9.49%. If the COMPOSER algorithm had not been deployed, the expected counterfactual mortality rate would have been 11.39% with a 95% confidence interval of [9.79%, 13.00%], corresponding to a 1.9% absolute decrease in sepsis-related in-hospital mortality. This value corresponds to a 17% relative decrease in in-hospital mortality among patients with sepsis and 22 additional patients who survived during the 5-month intervention period. The probability of this occurring by chance is determined from the Bayesian one-sided tail-area probability, $p = 0.014$. Additional data regarding the difference in mortality at our two hospitals are provided in Figs. S2.6–S2.9. We found at one site (the "safety net" hospital) we had a significant decrease in mortality in the post-intervention period, but we did not observe a significant change at the other clinical site (the quaternary care facility).

Table 2.2: Observed outcomes in the pre-intervention period, the expected counterfactual values from causal impact analysis, and the actual post-intervention values.

Outcome	**Pre-Intervention Value**	**Expected Post-Intervention Value (95% CI)**	**Actual Post-Intervention Value**
In-hospital mortality %	10.3%	11.4% (9.8% - 13.0%)	**9.5%**
Average 72 hour Change in SOFA	3.71	3.71 (3.6 - 3.8)	**3.56**
Sepsis bundle compliance rate	48.3%	48.4% (45.5% - 51.0%)	**53.4%**
Blood cultures prior to antibiotics compliance rate	71.1%	72.0% (69.9% - 73.9%)	73.9%
Rate of antibiotics administered within 24 hours prior and 3 hours after severe sepsis onset.	82.8%	82.8% (81.3% - 84.4%)	**84.6%**
Rate of lactate measured within 6 hours prior and 3 hours after severe sepsis onset	83.5%	83.4% (81.3% - 85.8%)	85.6%
Rate of repeat lactate measured within 6 hours after severe sepsis onset if initial lactate is elevated	97.8%	97.3% (96.2% - 98.4%)	**98.6%**
Rate of administration of vasoactive medications within 6 hours of septic shock	58.0%	57.5% (46.7% - 68.2%)	55.5%
Rate of administration of 30cc/kg of fluids within 3 hours of presentation of septic shock or hypotension	54.2%	53.9% (48.9% - 58.8%)	**59.3%**
ICU transfer rate	32.6%	32.5% (30.7% - 34.2%)	31.8%
Average ICU-free days	25.4	25.1 (24.6 - 25.6)	25.6

Significant post-intervention values against the 95% confidence interval are bolded.

The results from causal impact analysis on our primary and secondary outcomes are summarized in Table 2.2. The observed in-hospital mortality rate and the corresponding

predictions from the Bayesian structural time-series model are shown in Fig. 2.2a. The residual quantile-quantile and autocorrelation plots are described in Figs. S2.4 and S2.5. The average sepsis mortality rate during the post-intervention period was 9.49%. If the COMPOSER algorithm had not been deployed, the expected counterfactual mortality rate would have been 11.39% with a 95% confidence interval of [9.79%, 13.00%], corresponding to a 1.9% absolute decrease in sepsis-related in-hospital mortality. This value corresponds to a 17% relative decrease in in-hospital mortality among patients with sepsis and 22 additional patients who survived during the 5-month intervention period. The probability of this occurring by chance is determined from the Bayesian one-sided tail-area probability, $p = 0.014$. Additional data regarding the difference in mortality at our two hospitals are provided in Figs. S2.6–S2.9. We found at one site (the "safety net" hospital) we had a significant decrease in mortality in the post-intervention period, but we did not observe a significant change at the other clinical site (the quaternary care facility).

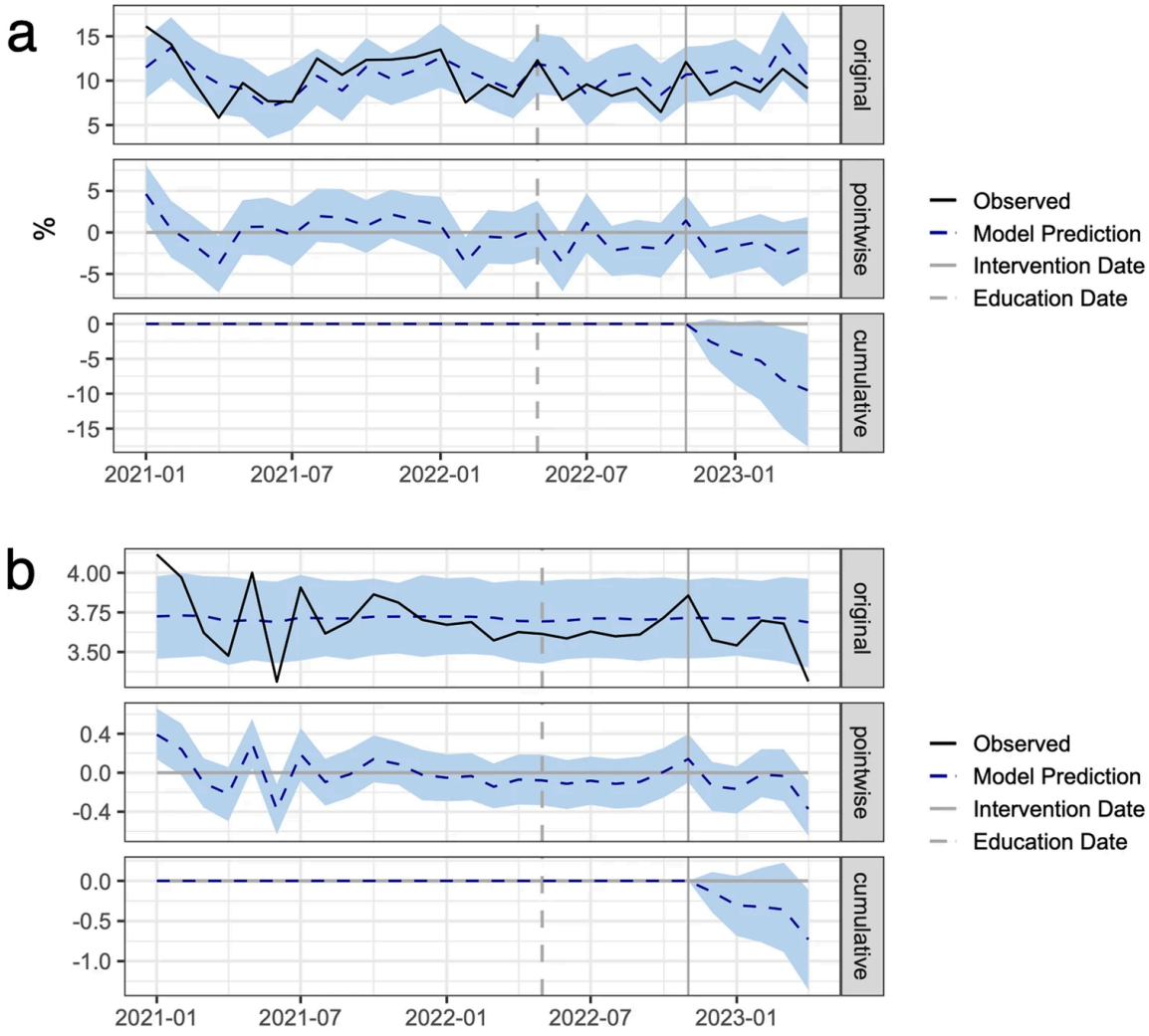


Figure 2.2: Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period. Preparation for the implementation of COMPOSER began in May 2022 approximately 6 months prior to the go-live date of the model. **a** The cumulative post-intervention in-hospital sepsis mortality rate is below the 95% confidence limit. **b** The cumulative post-intervention 72-h change in SOFA score is below the 95% confidence limit.

The average compliance rate during the post-intervention period was 53.42% while, in the absence of the COMPOSER intervention, the expected compliance rate would have been 48.38% (95% CI, 45.46%–51.01%; Fig. 2.3), or a 5.0% (95% CI, 2.4%–8.0%) increase in

compliance. This corresponds to a 10% (95% CI, 5%–16%) relative increase in sepsis bundle compliance following the implementation of COMPOSER. As shown in Figs. S2.10 and S2.11, compliance with our sepsis bundle increased at both EDs during the intervention period. Compliance with specific bundle elements is shown in Table 2.2 and Figs. S2.12–S2.17. We observe significant improvements in antibiotic compliance, repeat lactate compliance, and administration of fluids compliance.

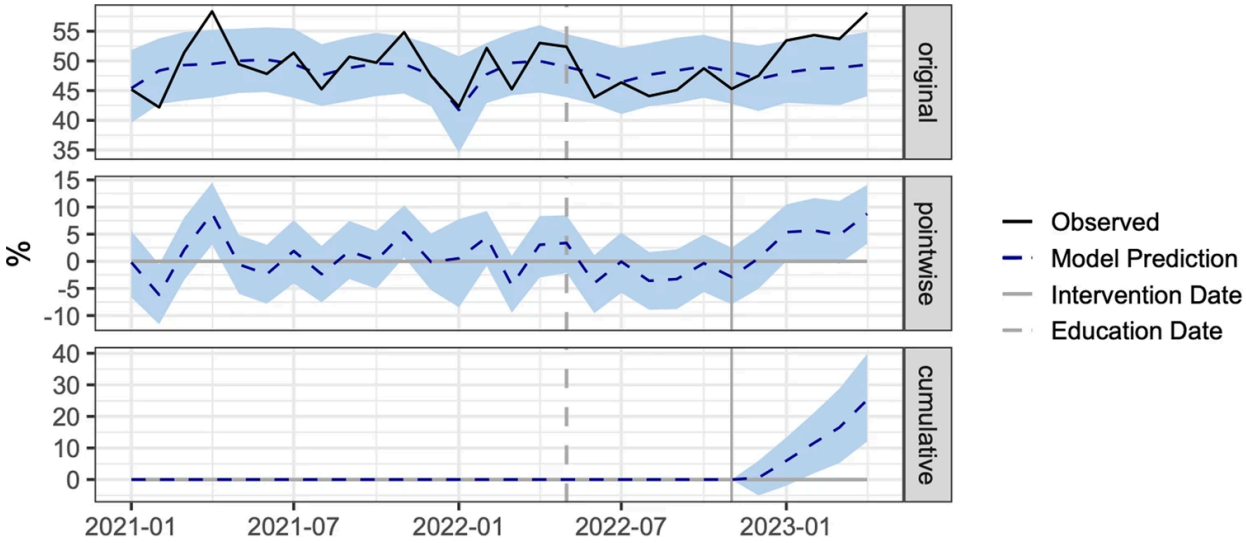


Figure 2.3: Implementation of the COMPOSER BPA was significantly associated with an increase in sepsis bundle compliance.

We also observed a reduction in the 72-h change in SOFA score following sepsis onset (Fig. 2.2b). The average change in SOFA score during the post-intervention period was 3.56. In the absence of the COMPOSER intervention, however, the expected counterfactual average change in SOFA score would have been about 3.71 with a 95% confidence interval of [3.58, 3.83]. This corresponds to a 4% decrease in the average 72-h change in SOFA score following sepsis onset. The probability of this occurring by chance is $p = 0.013$. Additional data on the change in SOFA score at each emergency department are provided in Figs. S2.18 and S2.19. We

further observe a downward trend in our secondary endpoint of ICU admissions (Fig. S2.20) and an upward trend in ICU-free days (Fig. S2.21) although neither reaches statistical significance. The temporal trends of all covariates used in the Bayesian structural time-series models are provided in Fig. S2.22.

Associations between time-to-antibiotics in septic patients and acknowledgement reasons are shown in Table 2.3. We observe that in cases where nurses indicated that they would notify physicians immediately, there was a significant reduction in time to antibiotic administration ($p = 0.002$; two-sided t-test with adjustments for ED volume, sex, baseline SOFA, Elixhauser comorbidity score, and age).ibiotic compliance, repeat lactate compliance, and administration of fluids compliance.

Table 2.3: Associations Between BPA Response and Antibiotics Timing.
Coefficients and *P* values are tabulated for a linear regression model with adjustments. If the nurse selected "No Infection Suspected" on an 80 year old male with an Elixhauser of 4 and a baseline SOFA of 1, the expected "time from ED triage to antibiotics administration" would be 14.5 hours. However, if the nurse had selected "Will Notify MD Immediately" the expected "time from ED triage to antibiotics administration" would have been 7.7 hours.

Variable	**Coefficient [95% CI]**	***P* Value[a]**
(Intercept)	22.42 [-45.27 – 90.11]	0.513
Age	0.01 [-0.15 – 0.18]	0.875
Male	-0.53 [-5.99 – 4.93]	0.848
Elixhauser Comorbidity Index	0.21 [-0.10 – 0.52]	0.180
Baseline SOFA	1.12 [-2.89 – 5.13]	0.581
Monthly ED Volume (in thousands)	0.62 [-13.33 – 14.58]	0.930
BPA Acknowledgement: No Infection Suspected	-13.24 [-26.60 – 0.127]	0.052
BPA Acknowledgement: Sepsis Treatment/Workup In Progress	-19.16 [-33.73 – -4.60]	0.010
BPA Acknowledgement: Will Notify MD Immediately	-19.95 [-32.39 – -7.51]	0.002

[a]*P* values are based on two-sided t-tests with adjustments for ED volume, sex, baseline SOFA, Elixhauser comorbidity score, and age.

DISCUSSION

In this before-and-after quasi-experimental study, we demonstrated that the implementation of a real-time deep-learning model to predict sepsis in two EDs was associated with a 5.0% absolute increase in sepsis bundle compliance and a 1.9% absolute decrease in-hospital sepsis-related mortality. This finding represents, to our knowledge, the first instance of prospective use of a deep-learning model demonstrating an association with improved patient-centered outcomes in sepsis. Our findings also suggest that the utilization of such models in clinical care was also associated with improvements in intermediate outcomes, such as less

organ injury at 72 h from the time of sepsis and improvements in elements of sepsis bundles which may explain the mortality benefit described. Importantly, we show in scenarios where nursing staff reported notification of the provider with concern for sepsis (approximately 55% of cases) that antibiotics were administered sooner, providing a plausible mechanism for the lower-than-expected in-hospital mortality we report.

Despite major interest in strategies to relieve the morbidity and mortality of sepsis, novel therapeutics have failed to translate into meaningful patient-centered outcomes. The potential to improve care through the use of artificial intelligence is attractive, particularly with advances in machine learning in the past decade.[82,83] Unfortunately, the majority of algorithms designed to predict sepsis never make it to the bedside.[44] To the best of our knowledge, the only deep-learning model previously tested in an ED setting is the Sepsis Watch by investigators at Duke; however, no patient-centered outcomes have been reported thus far.[84] As such, the present study is the first reporting of improvement in patient-centered outcomes attributable to the deployment of a deep-learning-based sepsis prediction model.

The use of deep learning for early prediction of sepsis is significant since such models are capable of modeling temporal, nonlinear, and complex correlations among risk factors, thus enabling them to solve more difficult problems. Moreover, deep-learning models are capable of handling large quantities of multimodal data from radiology imaging, clinical notes, and wearable sensors, among other.[85,86,87,88] Additionally, this class of models provides a flexible framework for transfer learning and continual learning to enable the adoption of such models to local healthcare settings.[89,90,91]

Importantly, the COMPOSER deep-learning model was designed to minimize false alarms via the conformal prediction framework. This approach imposes a boundary around the algorithm, which enables the model to identify whether it has enough prior knowledge of similar cases to determine reliably whether a patient is at risk for sepsis. If the algorithm finds the data non-conformant to the training samples, it will flag the case as 'indeterminate'. The resulting reduction in false alarms, previously reported to be 75%, greatly reduces the burden of resources or time spent on false diagnoses.

There are various potential reasons that may explain the reduction in mortality described above. First, we noted a high percentage (~55%) of alerts were transmitted by nursing staff to physicians. In this scenario, we found that these patients were more likely to receive timely antibiotics, thus providing a potential mechanism to decrease mortality and mitigate organ dysfunction at 72 h. The use of artificial intelligence to facilitate a shared mental model of risk between nursing staff and providers has demonstrated good acceptance and improved use of these models in other clinical areas.[92] In our system, for instance, we chose to have the nurses receive the alert and determine if escalation to the provider was appropriate. While the ideal target population for such an intervention is unclear, we felt that our nurses would be the ideal candidate for this alert because of the high frequency of nurses opening patients' charts. In the author's collective experience, physicians in the ED may have up to 15–20 patients at a time and may not receive a BPA that requires a chart to be open to receive the notification. Given the high rate of provider notification, we suspect that this approach was beneficial to patient care while additionally minimizing unnecessary alerts. Finally, although speculative, it is possible that the implementation of the alert improved situational awareness of sepsis care within our ED staff. This finding has been reported in other sepsis clinical decision tools as well.[93]

Despite our study's strengths, we acknowledge several limitations. First, our study was not randomized and thus our findings do not allow definitive causal inferences or mechanistic insights. We performed a causal impact analysis with common confounders which revealed that the implementation was significantly associated with positive outcomes. Regardless, we view the findings as important and believe that they provide a strong rationale for further research. Second, our study was conducted at two EDs in a large academic center that has a major interest in sepsis and clinical informatics. Although we had a large sample size and a diverse population of patients (racial, ethnic, socioeconomic status, etc.), we acknowledge the need for external validation in other healthcare settings (e.g., community hospitals, different demographics, hospitals without robust IT infrastructure, etc.). Third, one could argue that an abrupt intervention has important immediate benefits raising awareness and helping to prioritize the care of a specific group of patients. Conversely, the sustainability of the intervention could be questioned, emphasizing the need for longer-term follow-up. Although human interventions are subject to fatigue and complacency, we anticipate our automated algorithms will improve over time with increasing experience and larger data sets, which will likely result in improvements in end-user satisfaction. However, we certainly recognize the importance of continuous education as a component of care optimization. Finally, we did not evaluate the impact of this alert on patients who ultimately did not have sepsis, such as the potential adverse effects of inappropriate use of antibiotics and healthcare costs associated with this. We also acknowledge that we did not have any comparison data from the same time period as all of our EDs used this model. However, we did not have any other quality improvement initiatives during the same time period. Despite these limitations, we view our new findings as actionable and important.

In the before-and-after quasi-experimental design study conducted at two EDs, we demonstrate that the implementation of a real-time deep-learning model to predict sepsis was associated with a significant increase in bundle compliance, a significant reduction in in-hospital mortality, less organ dysfunction at 72 h, and improved timeliness to antibiotics when nurses notified the physician of the BPA. To our knowledge, this is the first time that the improvement of patient outcomes due to the use of a deep-learning model for sepsis prediction has been reported. Future multicenter randomized trials are indicated to validate these findings across a diverse hospital and patient population.

METHODS

Study design and cohort

We conducted a prospective before-and-after quasi-experimental study to evaluate the impact of a sepsis Best Practice Advisory (BPA; Fig. 2.4) powered by the COMPOSER deep-learning model on patient outcomes and process measures. The University of California San Diego Institutional review board (IRB) approval was obtained with the waiver of informed consent (#805726) and additional approval was obtained from the Aligning and Coordinating QUality Improvement, Research, and Evaluation (ACQUIRE) Committee (project #609). Our study was completed in accordance with STROBE guidelines.[94] These EDs have a total volume of approximately 100,000 patients annually with one serving at a quaternary academic center and the other at an urban "safety net" hospital.

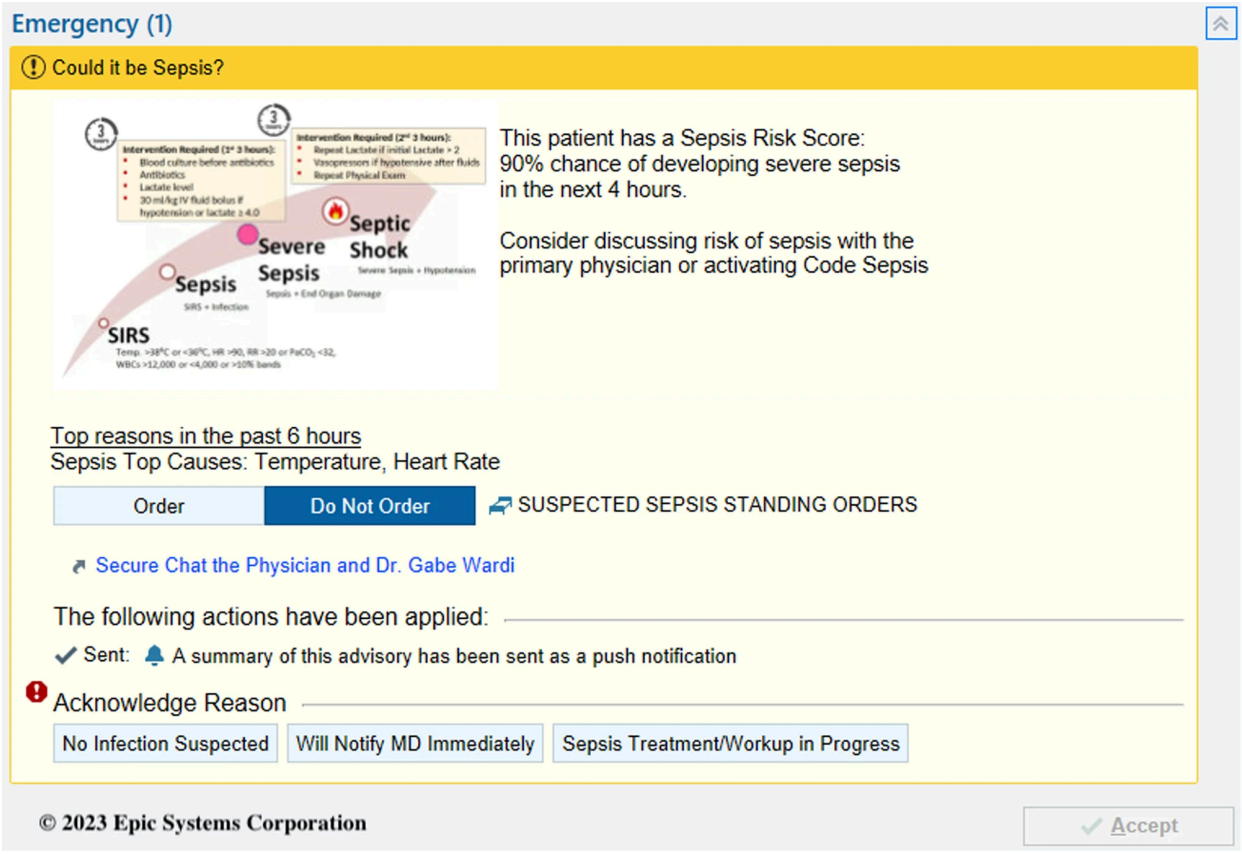


Figure 2.4: COMPOSER Best Practice Advisory.

Patients were identified as septic according to the latest international consensus definitions for sepsis ("Sepsis-3").[3,6] The onset time of sepsis was established by following previously published methodology, using evidence of organ dysfunction and suspicion of clinical infection.[3,20,95] Clinical suspicion of infection was defined by a blood culture draw and at least 4 days of non-prophylactic intravenous antibiotic therapy satisfying either of the following conditions: (1) if a blood culture draw was ordered first, then an antibiotics order had to occur within the following 72 h, or (2) if an antibiotics order occurred first, then a blood culture draw had to occur within the next 24 h. Evidence of organ dysfunction was defined as an increase in the Sequential Organ Failure Assessment (SOFA) score by two or more points. In particular,

evidence of organ dysfunction occurring 48 h before to 24 h after the time of suspected infection was considered, as suggested in Seymour et al.[3] Finally, the time of onset of sepsis was taken as the time of clinical suspicion of infection. The inclusion of 4 days of non-prophylactic antibiotics to improve the specificity of sepsis is similar to what Rhee et al. proposed as a surveillance approach for identifying sepsis from electronic health records which have outperformed reliance on administrative coding of sepsis. We included all adult patients (age ≥18 years old) who met the criteria for the above-described Sepsis-3 definition within the first 12 h of their ED stay. We excluded patients who were transitioned to comfort measures prior to their time of sepsis and patients who developed sepsis after 12 h of hospital admission. All data used to derive the onset time of sepsis and patient outcomes were extracted via SQL queries on Epic Clarity.

Sepsis algorithm and platform

The COMPOSER algorithm for the early prediction of sepsis is described in Shashikumar et al.[20] It is a feed-forward neural network model that incorporates routinely collected laboratory and vital signs as well as patient demographics (age and sex), comorbidities, and concomitant medications to output a risk score for the onset of sepsis within the next 4 h. Importantly, the model utilizes the conformal prediction method to reject out-of-distribution samples that may arise due to data entry error or unfamiliar cases. The model achieves an area under the receiver operating characteristic curve (AUROC) of 0.938–0.945 within ED settings. We fixed the score threshold to achieve an 80% sensitivity level. Prior work demonstrated that at this sensitivity, the PPV was 20.1%.

The COMPOSER algorithm is hosted on the cloud-based healthcare analytics platform described in Chapter 1 that enables access to data elements in real-time by leveraging the FHIR

and HL7v2 standards (Fig. S2.1).[34] Specifically, the Amazon Web Services (AWS)-hosted infrastructure receives a continuous stream of Admit, Discharge, Transfer (ADT) messages from the hospital's integration engine to determine the active patients and map their journey through care units. The platform extracts data at an hourly resolution for these patients using FHIR APIs with OAuth2.0 authentication and passes the feature set to COMPOSER. Hourly frequency was selected to ensure adequate data availability for a prediction. The resulting sepsis risk score and the top features driving the recommendation are then written to a flowsheet within the EHR using an HL7v2 outbound message. This flowsheet triggers a nurse-facing BPA (Fig. 2.4) on the chart open which alerts the caregiver that the patient is at risk of developing severe sepsis and provides the model's top reasons. The nurse could acknowledge the alert by selecting one of the four options: (i) no infection suspected, (ii) sepsis treatment/workup in progress, or (iii) will notify MD immediately. If a nurse exited the patient's chart without selecting an option, we recorded this as "no acknowledgement". If the "will notify MD immediately" option is selected, the nurse can use a 'secure chat' feature to contact the provider from within the BPA to discuss the care of the patient.

Prospectively deployed algorithms are susceptible to model drift in which their performance degrades overtime due to changes in the patient population or treatment practices.[76] To detect this possibility, we implemented a data quality dashboard that tracks the median values of all input features to ensure they are within their upper and lower process control limits (based on the upper and lower quantiles from the training cohort). We further evaluate model performance, such as sensitivity and positive predictive value (PPV), biweekly to ensure there is no degradation in COMPOSER performance (Supplementary Note 2.2). We established a Predetermined Change Control Plan (PCCP) to trigger model retraining if the performance drops

below predetermined thresholds, although this has not been required as of the time of this reporting.

Implementation of COMPOSER into our electronic health record

Implementation was completed in various stages according to the EPIS (exploration, preparation, implementation, sustainment) framework, with frequent feedback provided to nursing staff during the implementation and sustainment periods.[96] Early stages in the Exploration stage began 2–3 years prior to model deployment with significant institutional support at the departmental and health system level. Preparation began approximately 6 months prior to the go-live date. Included in this was the creation of a multidisciplinary team to guide implementation, surveys of the nursing staff to identify specific needs, educational sessions, and iterative changes to the BPA from end-users. We employed a "silent mode trial" in which COMPOSER outputs were reviewed in real-time by a team of physicians to assess the accuracy and usefulness of the alerts as described in Chapter 1. Adjustments to the algorithm were iteratively made based on these reviews to improve the timeliness and appropriateness of the alerts. During the Implementation phase, frequent feedback and education were provided to nursing staff on the COMPOSER model. In final form, our BPA would fire for all adult (at least 18 years old) patients who were receiving care in the ED with a score above the threshold and the following exclusions: patient discharged or deceased, comfort care measures initiated, patient no longer under the care of ED nurses, or a sepsis bundle had previously been instituted during their stay. Nurses would have to have the patient's chart open for it to fire. The lockout periods for each acknowledgement reason were: "No infection suspected" 8 h; "Will notify MD immediately" 12 h; Sepsis treatment/work-up in progress" 12 h. We defined our pre-intervention

time period from January 1st, 2021 to December 6th, 2022. COMPOSER went live December 7th, 2022. Our post-implementation phase was from December 7th, 2022 until April 30th, 2023.

Primary and secondary outcomes

Our primary outcome was in-hospital mortality. Secondary outcomes included: compliance with our sepsis bundle (initial and repeat serum lactate if initial lactate >2 mmol/L, initiation and completion of a 30 mL/kg crystalloid fluid bolus, checking of blood cultures prior to antibiotics, and initiation of intravenous antibiotics within 3 h of time of sepsis), 72-h change in sequential organ failure assessment (SOFA) score following sepsis onset, ICU admission, and ICU-free days. ICU-free days were calculated as 30 less the number of ICU days with in-hospital death and stays longer than 30 days were fixed at 0. For patients who died in the ICU, this value is 0. For example, a patient who is in the ICU for 4 days and survives would have a value of 26. Patients who either die in the ICU or in the ICU for > 29 days have a value of 0.

Statistical methods

Descriptive statistics were provided as indicated. Differences between the pre-intervention and post-intervention cohort were assessed with Kruskal–Wallis rank sum tests on continuous variables and Pearson's chi-squared tests on categorical variables and significance was assessed at a P-value of 0.05. All statistical analyses were performed using the R statistical software version 4.0.4 and the CausalImpact package version 1.2.7.[97,98]

To estimate the causal effect of the COMPOSER BPA intervention, we performed causal inference using a Bayesian structural time-series model.[98] This approach, pioneered by Brodersen et al. from Google Inc., has been widely used to assess the impact of advertisement

campaigns on product sales and the effect of economic changes on markets.[99,100,101] Here, we apply it to patient outcomes data to assess the impact of the COMPOSER algorithm adjusted for confounders. Briefly, a state-space model is trained on the control time-series prior to the intervention of interest. The observed outcome is modeled as a function of the latent state and Gaussian noise, with the latent state modeled by a local linear trend, in addition to a linear regression on the model covariates. In this work, we assume the effect of the regression coefficients on the outcome of interest is independent of time and the static regression parameters are sampled from a spike-and-slab prior distribution. Posterior inference is then performed on the post-intervention time-series to estimate the counterfactual outcome if the intervention had not been introduced. Under the assumption that the response variable in the control time-series is independent of the intervention, the difference between the model prediction and the observed value is a probability density of the causal impact of the intervention over time (See Supplementary Note 2.3 for more details). Additionally, model residuals were evaluated using quantile-quantile and autocorrelation plots to ensure adequate modeling of the time-series information.

Emergency department volume, sex, baseline SOFA, comorbidity burden via the Elixhauser comorbidity score, age, COVID-19 infection status, ED location (La Jolla or Hillcrest), local trends, and season were included as covariates in the Bayesian structural time-series model and 1000 samples of Markov Chain Monte Carlo were used for posterior inference. No imputation was performed since all covariates were fully observed. There were no missing values present in the covariates. We included seasons and local trends (e.g., ED volume) as prior data suggest outcomes of sepsis patients are worse in the winter and may be impacted by high patient volumes.[102.103] Outcomes were predicted at a monthly resolution to reduce the

influence of random fluctuations on the outcome variable. We plotted the model predictions and true outcomes data, the pointwise difference between the two, and the cumulative difference across the post-intervention period and assessed significance against the 95% confidence intervals.

We further evaluated the association of alert acknowledgement and sepsis intervention, measured by the time from ED triage to the administration of antibiotics. We adjusted for the aforementioned confounders and performed a two-sided t-test on time-to-antibiotics as a function of acknowledgement reason.

ACKNOWLEDGEMENTS

Chapter 2, in full, is a reprint of the material as it appears in “Impact of a deep learning sepsis prediction model on quality of care and survival”. Boussina, A., Shashikumar, S. P., Malhotra, A., Owens, R. L., El-Kareh, R., Longhurst, C. A., ... & Wardi, G. (2024) npj Digital Medicine, 7(1), 14. The dissertation author was the primary investigator and author of this paper.

SUPPLEMENTARY MATERIAL

Supplementary Note 2.1. Cloud-based infrastructure for model deployment.

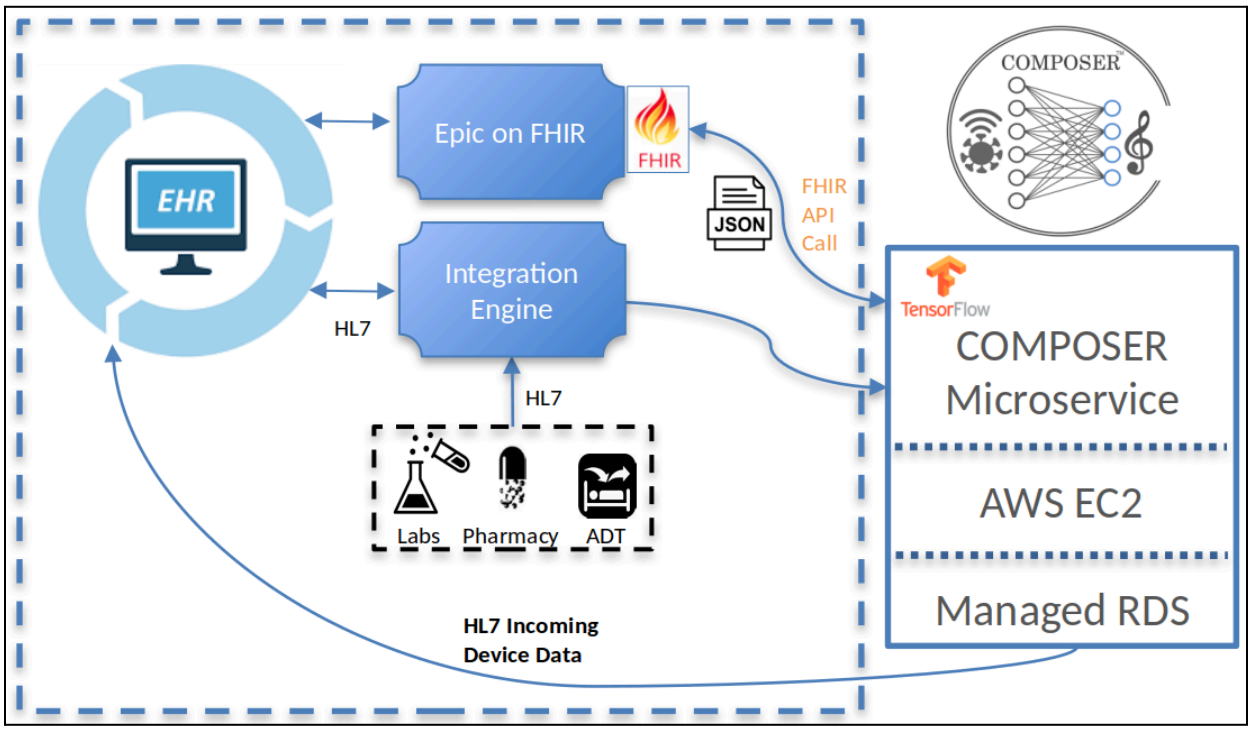


Figure S2.1: System Diagram of the Deep Learning Platform.
The core application runs on an elastic computing virtual machine with an auto-scaling group to ensure fault tolerance and high availability. HL7 message feeds and FHIR JSON responses are stored in managed relational databases. Model predictions are sent back to the EHR through an HL7 device data interface. HL7 = health level 7; AWS = Amazon Web Services; FHIR = Fast Healthcare Interoperability Resources; ADT = admission discharge transfer; RDS = relational database service; EC2 = Elastic Compute Cloud.

Supplementary Note 2.2. Algorithm Performance Monitoring

The cloud application described in Supplementary Note 2 captures the HL7v2 and FHIR feeds for a patient through discharge. This enables identification of patients meeting Sepsis-3 criteria as described in the main text. The COMPOSER performance monitoring system tags these patients for Sepsis-3 on a biweekly schedule and compares these labels against the COMPOSER BPA firings, allowing for automated monitoring of the algorithm sensitivity, specificity, and PPV. We visualize these metrics, as well as the input feature distributions, in a data quality dashboard (Figure S2.2) and generate an AWS CloudWatch alert for algorithm retraining when the PPV drops below a threshold of 20%. Additionally, the system generates PagerDuty® alerts whenever the FHIR connection is interrupted, an outbound HL7v2 message fails to be received, or the inbound ADT connection is severed.

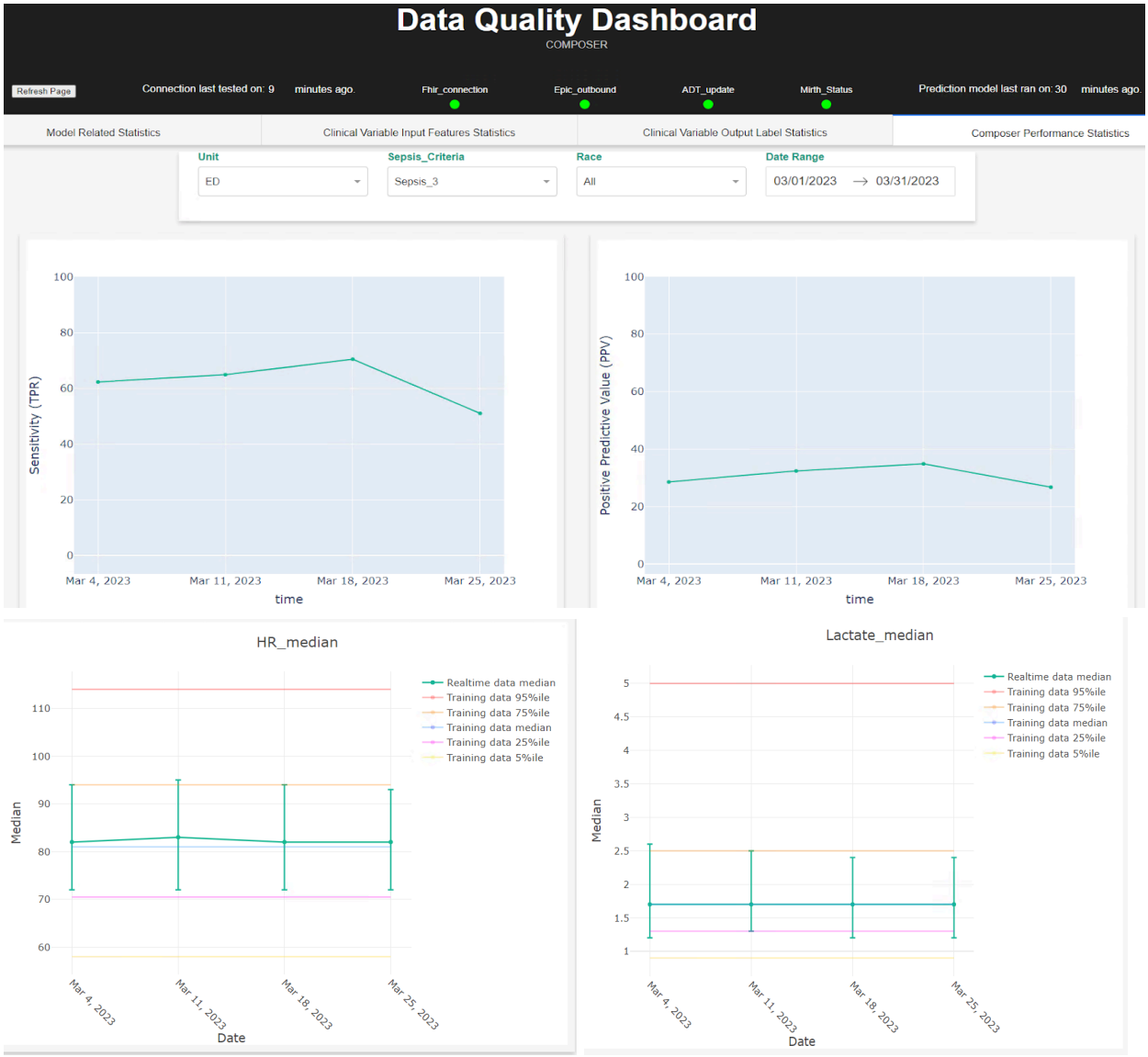


Figure S2.2: Examples of monitoring of Input features and performance outputs from the COMPOSER data quality dashboard.

Supplementary Note 2.3. Causal impact analysis with Bayesian structural time-series models. The Bayesian structural time-series model used in this work is described by Supplementary Equations 1 - 7 and Fig S2. The observed outcome y_t is modeled as the combination of a local state (μ_t), covariate regression ($x_t^T\beta$), seasonality (γ_t), and Gaussian noise (ε_t). The state components (μ_t, δ_t) are modeled by a local linear trend which evolve according to Gaussian random walks. A spike-and-slab prior is used on the regression parameters (β) which are integrated out, enabling the uncertainty of covariate influence to be accounted for in the model inference. Additionally, a hyperparameter search is performed on the order of seasonality parameter (S) to ensure whiteness of the autocorrelation of the residuals (i.e., any remaining unexplained residuals are uncorrelated gaussian noise). Posterior inference is carried out by drawing 1000 samples using Markov Chain Monte Carlo, where model parameters are sampled from their joint posterior distribution given the data. For each set of sampled parameters, a prediction y_i is generated. The distribution of these predictions across all 1000 samples provides an estimation of the prediction uncertainty.

$$(1)\ y_t = x_t^T\beta + \mu_t + \gamma_t + \varepsilon_t$$
$$(2)\ \mu_t = \mu_{t-1} + \delta_{t-1} + \eta_{\mu,t}$$
$$(3)\ \delta_t = \delta_{t-1} + \eta_{\delta,t}$$
$$(4)\ \varepsilon_t \sim N(0, \sigma_t^2)$$
$$(5)\ \gamma_{t+1} = -\sum_{s=0}^{S-2} \gamma_{t-s} + \eta_{\gamma,t}$$
$$(6)\ \eta_{\delta,t} \sim N(0, \sigma_\delta^2)$$
$$(7)\ \eta_{\mu,t} \sim N(0, \sigma_\mu^2)$$

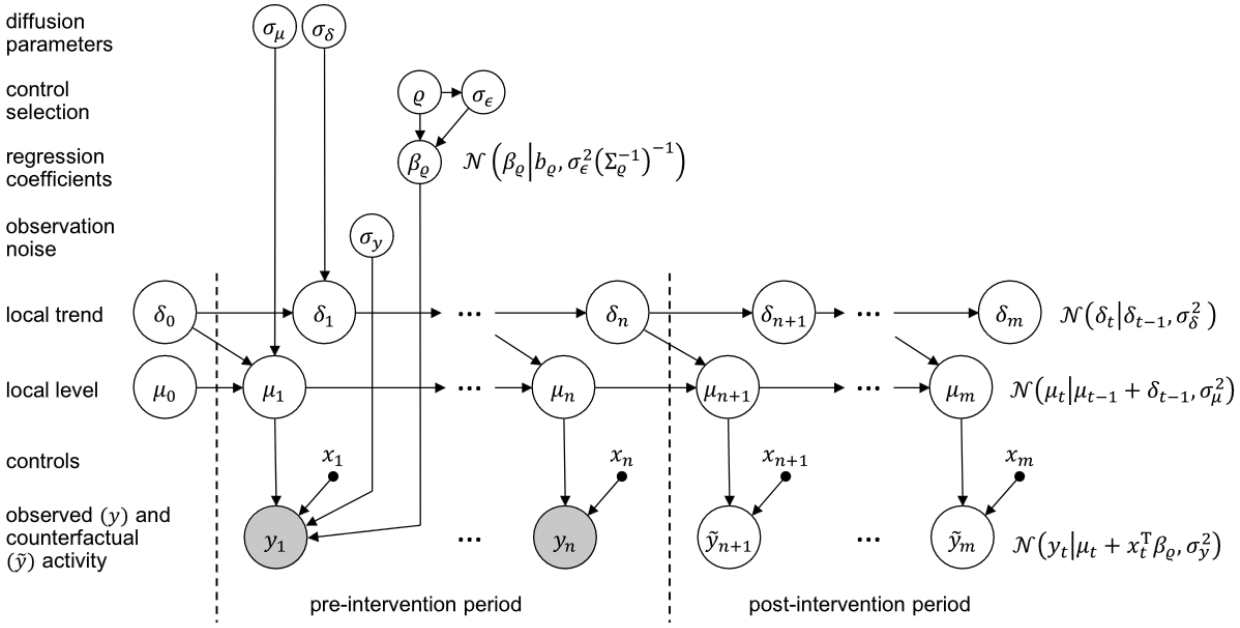


Figure S2.3: Graphical model of the state-space model, adapted from Brodersen, et al.[98]

Supplementary Note 2.4. Comparison between the two emergency departments.

Supplementary Table 2.1: Demographics and baseline characteristics of septic patients at the La Jolla (LJ) and Hillcrest (HC) emergency departments.

	Total	LJ Emergency Department	HC Emergency Department	*P* Value[a]
Characteristic				
Number of Patients, N (%)	6217 (100%)	3002 (48.3%)	3215 (51.7%)	-
Age, Mean (SD)	63 (17.1)	65 (17.1)	61 (16.8)	< .001
Sex, N (%)				
Male	3592 (57.8%)	1586 (52.8%)	2006 (62.4%)	-
Female	2625 (42.2%)	1416 (47.2%)	1209 (37.6%)	-
Ethnicity				
Asian	530 (8.5%)	365 (12.2%)	165 (5.1%)	-
Black or African American	639 (10.3%)	141 (4.7%)	498 (15.5%)	-
White	2983 (48%)	1564 (52.1%)	1419 (44.1%)	-
Other[c]	2065 (33.2%)	932 (31%)	1133 (35.2%)	-
Ethnic Group				-
Hispanic/Latino	1756 (28.2%)	805 (26.8%)	951 (29.6%)	-
Not Hispanic/Latino	4461 (71.8%)	2197 (73.2%)	2264 (70.4%)	-
Cancer Patients, N (%)	1602 (26.6%)	1211 (40.3%)	391 (12.1%)	< .001
Organ Dysfunction				
Elixhauser Comorbidity Index, Median (IQR)	5 (0-13)	7 (0-16)	0 (0-11)	< .001
SOFA Score at Time of Sepsis, Median (IQR)	0.7 (1.3)	0.5 (0.98)	0.9 (1.52)	< .001
Lab Values				
Lactate at Time of Sepsis	2.4 (1.6 - 4.3)	2.15 (1.5 - 3.5)	2.7 (1.7 - 5.0)	< .001
Interventions				
Mechanical Ventilation, N (%)[d]	1,035 (16.6%)	849 (16.8%)	186 (16.1%)	< .001
Administration of Vasoactive Medications, N (%)[d]	424 (6.8%)	345 (6.8%)	79 (6.9%)	0.008

[a]P values for continuous variables are based on Kruskal-Wallis rank sum tests. P values for categorical variables are based on Pearson's chi-squared tests.
[b]Ethnicity is self-reported.
[c]Other ethnicity corresponds to respondant selecting Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, Other Race or Mixed Race, or Unknown.
[d]Within 72-hours of ED arrival.

Supplementary Note 2.5. Diagnostic plots.

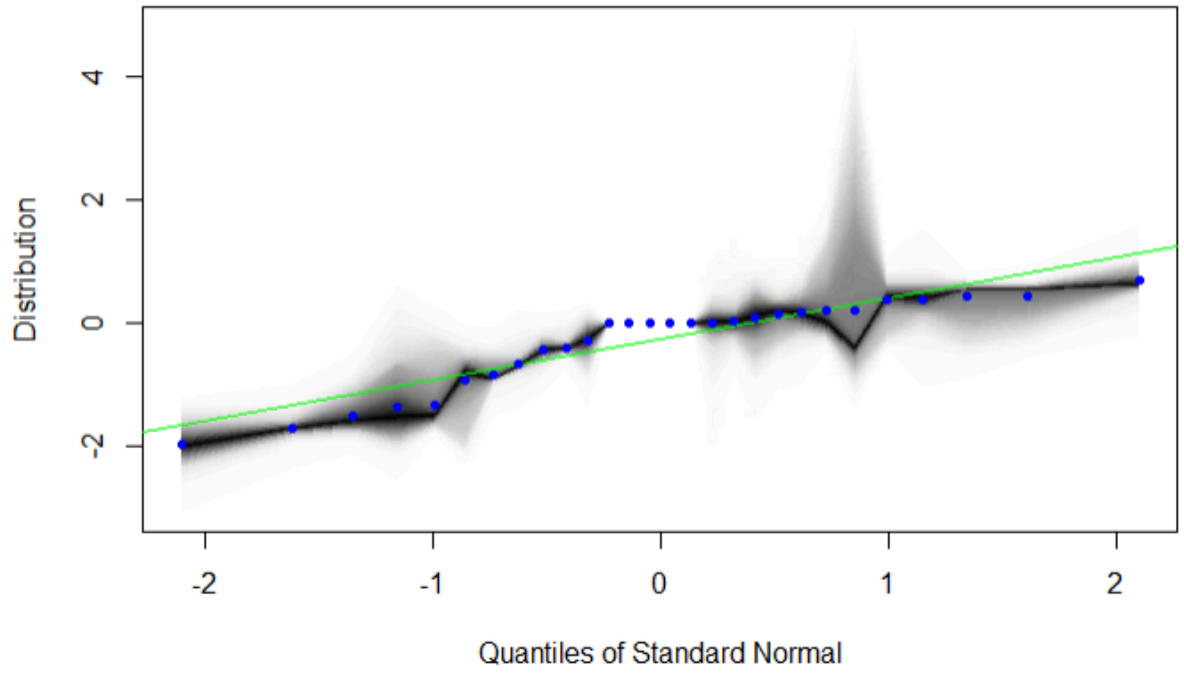


Figure S2.4: Residual quantile-quantile (QQ) plot for the mortality Bayesian structural time-series model.
Residuals from a well-fitted model are expected to be random Gaussian noise. The observed residuals for the sepsis mortality time-series model (Figure 2.2A) are well approximated by a normal distribution.

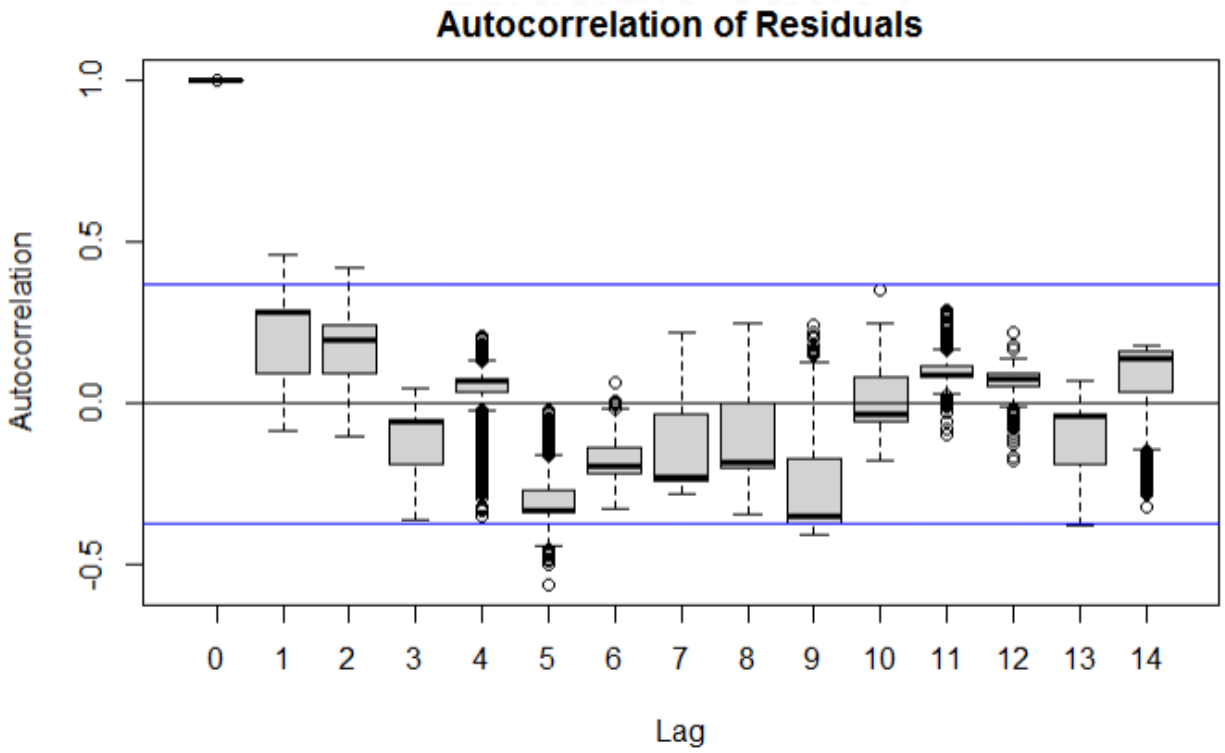


Figure S2.5: Boxplot of autocorrelation of residuals for the mortality Bayesian structural time-series model.
The center line of each box represents the median autocorrelation across 1000 runs of the Markov chain Monte Carlo algorithm. The box limits represent the upper and lower quantiles and the whiskers extend up to 1.5x the IQR with outliers displayed as individual points. There is no significant correlation between the residuals suggesting adequate fit of the time-series model.

Supplementary Note 2.6. Causal impact analysis of patient mortality by emergency department.

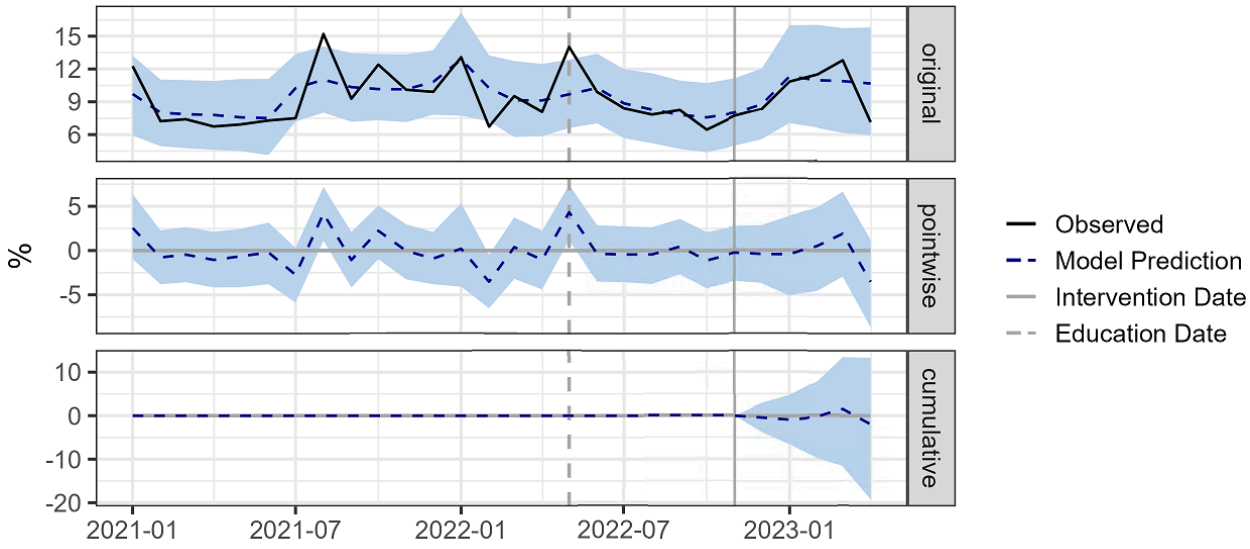


Figure S2.6: Causal impact analysis of COMPOSER Best Practice Advisory on sepsis mortality at the La Jolla Emergency Department.
Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

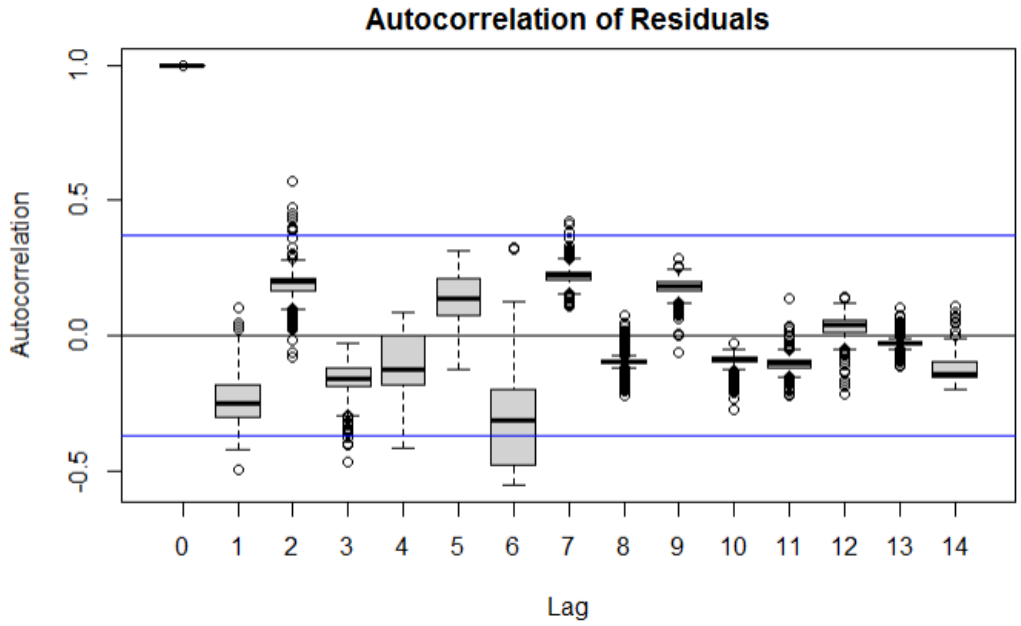


Figure S2.7: Boxplot of autocorrelation of residuals for the mortality Bayesian structural time-series model for sepsis mortality at the La Jolla Emergency Department.
The center line of each box represents the median autocorrelation across 1000 runs of the Markov chain Monte Carlo algorithm. The box limits represent the upper and lower quantiles and the whiskers extend up to 1.5x the IQR with outliers displayed as individual points. There is no significant correlation between the residuals suggesting adequate fit of the time-series model.

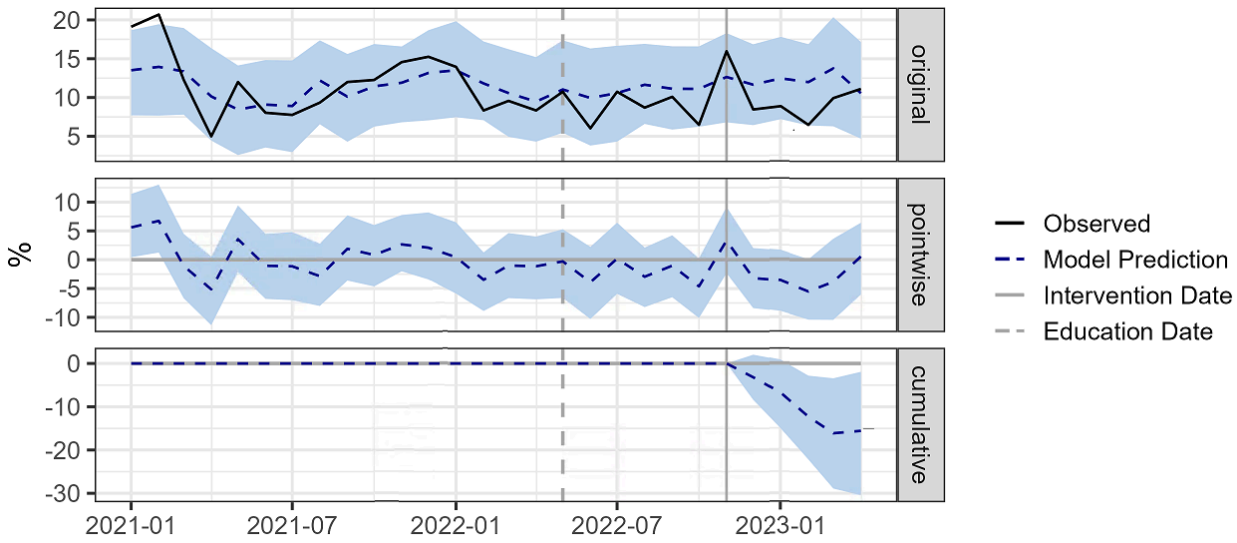


Figure S2.8. Causal impact analysis of COMPOSER Best Practice on sepsis mortality at the Hillcrest Emergency Department.

Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

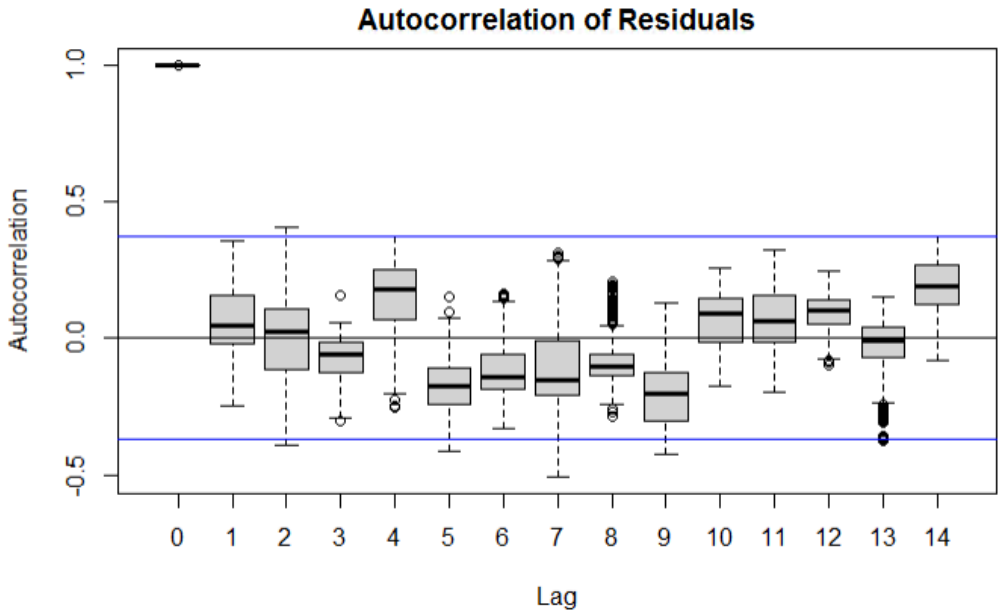


Figure S2.9: Boxplot of autocorrelation of residuals for the mortality Bayesian structural time-series model for sepsis mortality at the Hillcrest Emergency Department.
The center line of each box represents the median autocorrelation across 1000 runs of the Markov chain Monte Carlo algorithm. The box limits represent the upper and lower quantiles and the whiskers extend up to 1.5x the IQR with outliers displayed as individual points. There is no significant correlation between the residuals suggesting adequate fit of the time-series model.

Supplementary Note 2.7. Causal impact analysis of compliance with the individual sepsis bundle components.

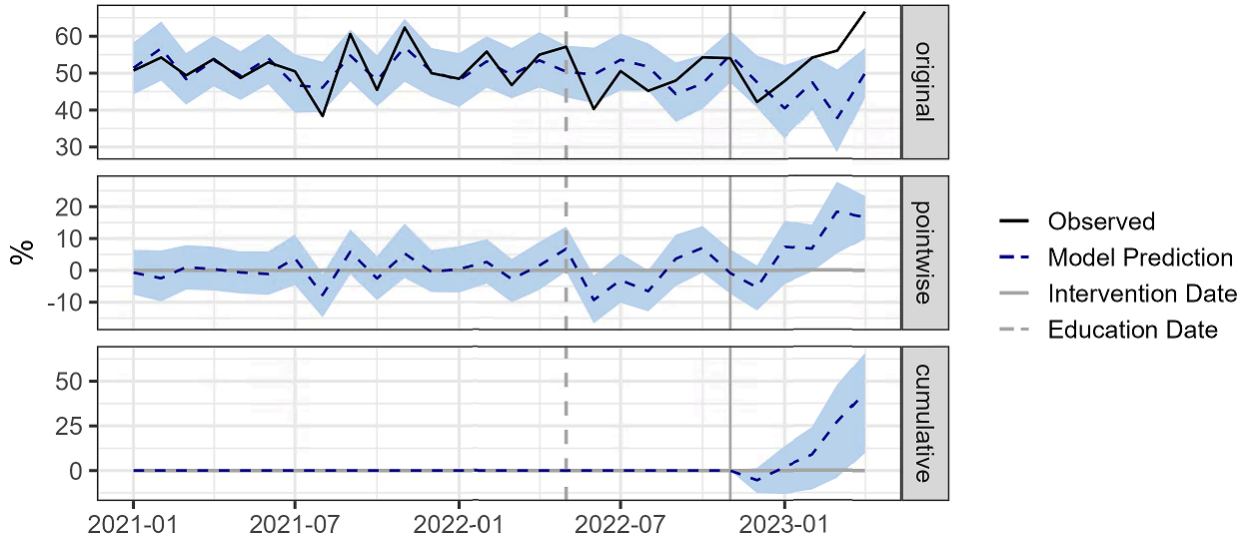


Figure S2.10: Causal impact analysis of COMPOSER Best Practice Advisory on sepsis bundle compliance rate at the La Jolla Emergency Department.
Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

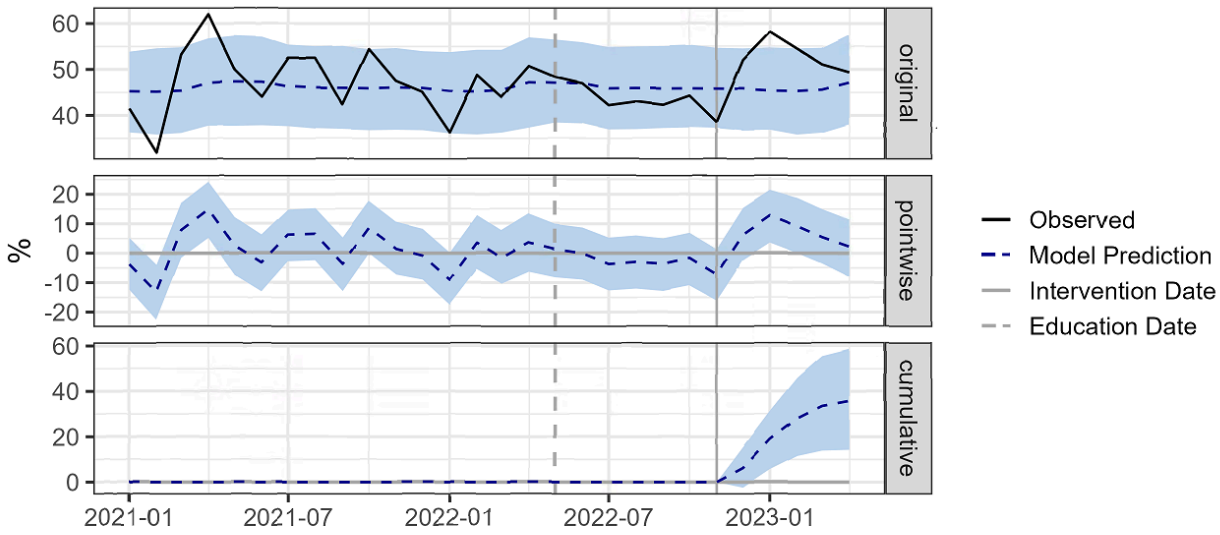


Figure S2.11: Causal impact analysis of COMPOSER Best Practice Advisory on sepsis bundle compliance rate at the Hillcrest Emergency Department.
Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

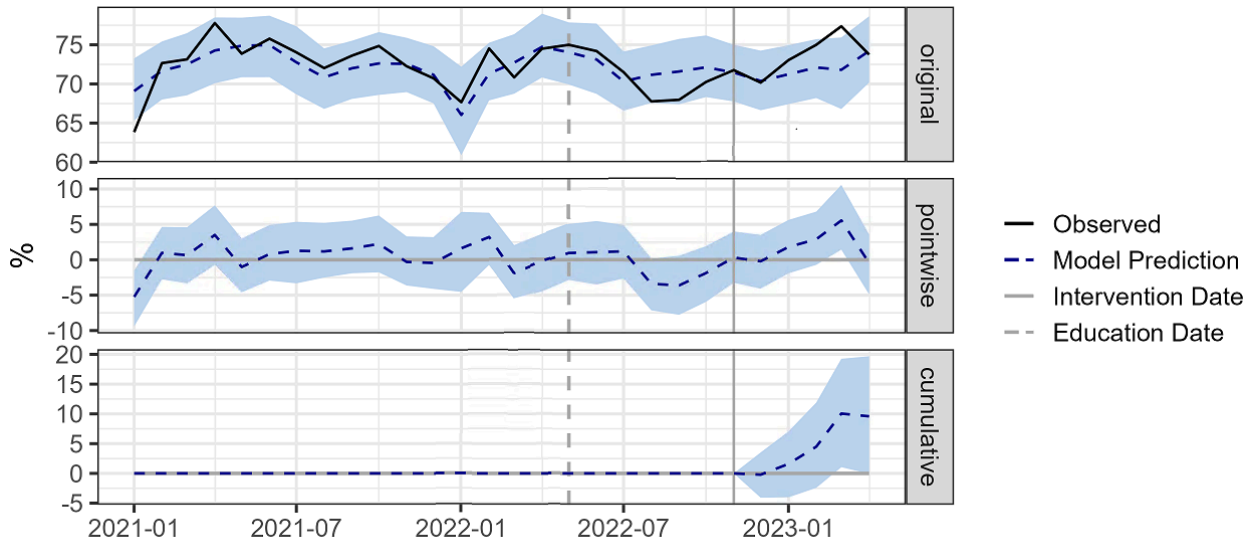


Figure S2.12: Blood cultures prior to antibiotics compliance rate.
Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

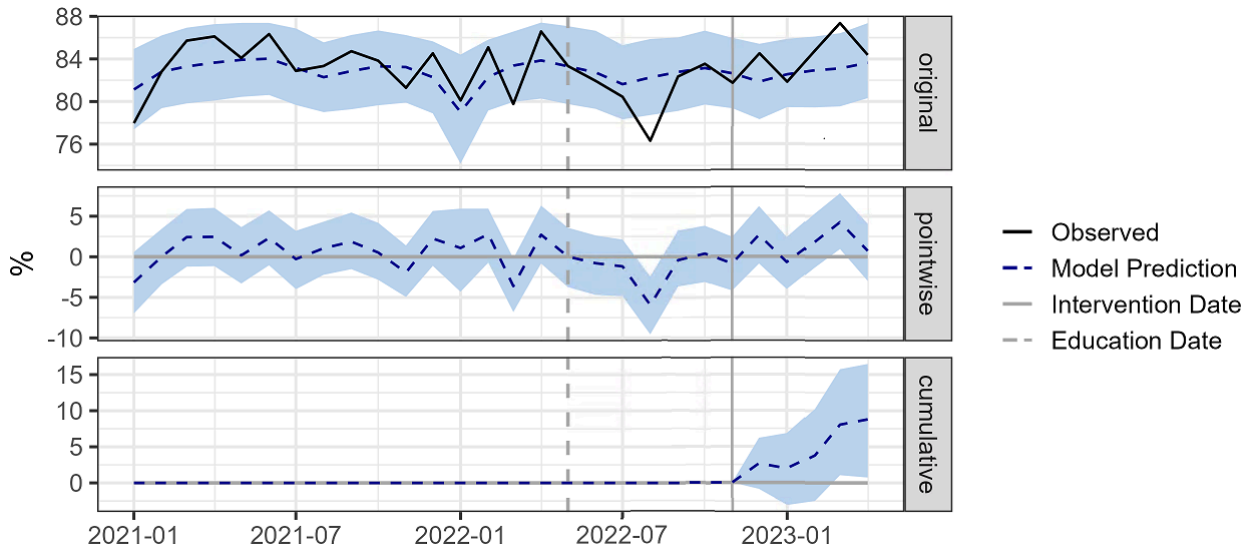


Figure S2.13: Rate of antibiotics administered within 24 hours prior and 3 hours after severe sepsis onset.
Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

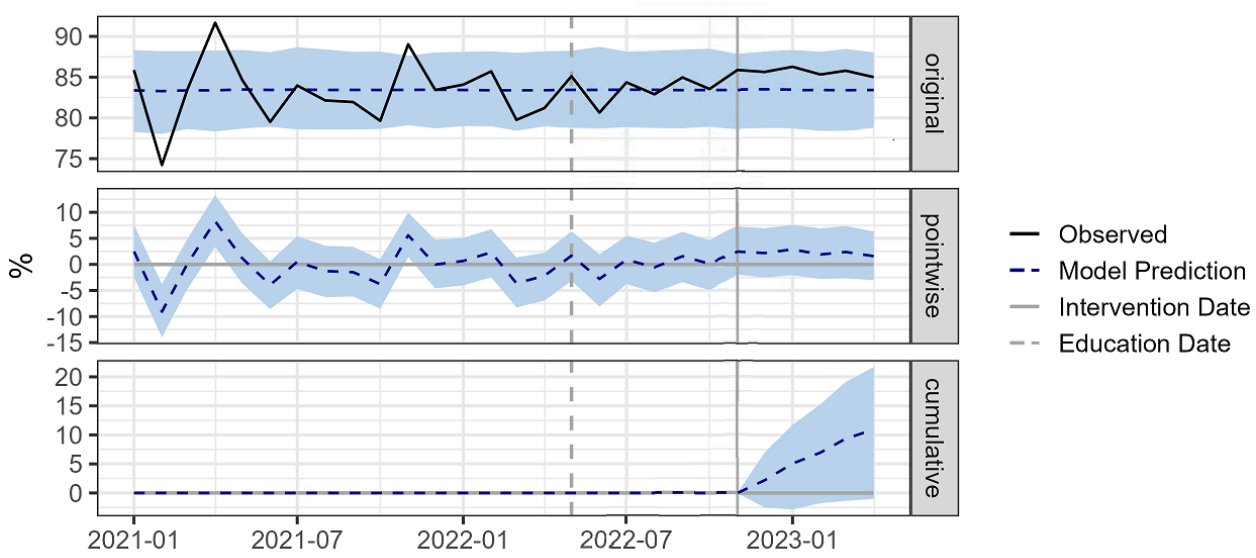


Figure S2.14: Rate of lactate measured within 6 hours prior and 3 hours after severe sepsis onset. Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

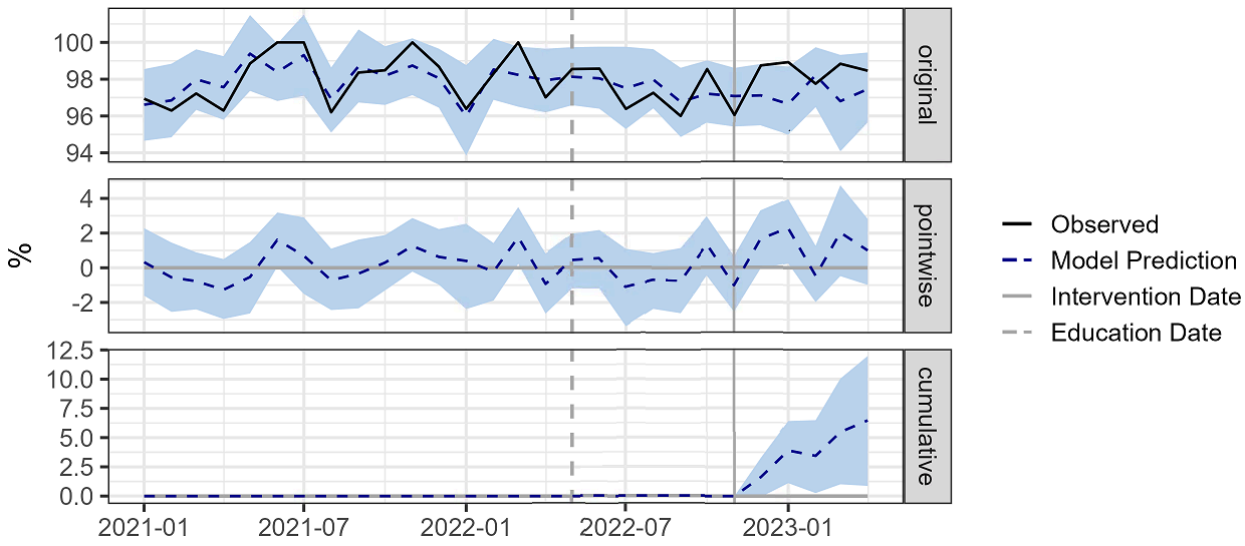


Figure S2.15. Rate of repeat lactate measured within 6 hours after severe sepsis onset if initial lactate is elevated.

Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

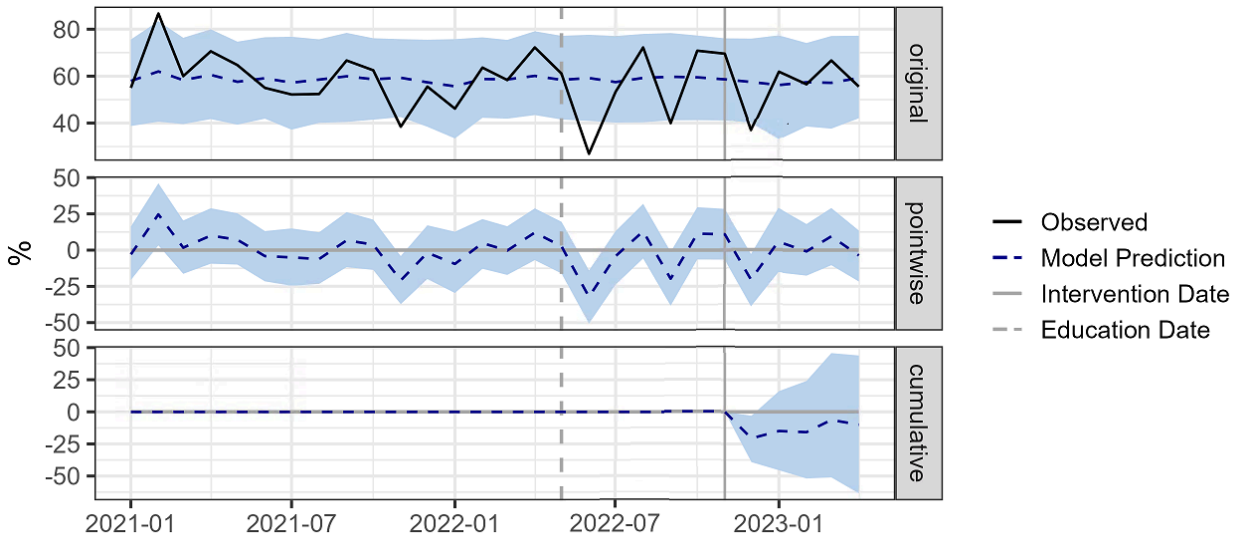


Figure S2.16: Rate of administration of vasoactive medications within 6 hours of septic shock. Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

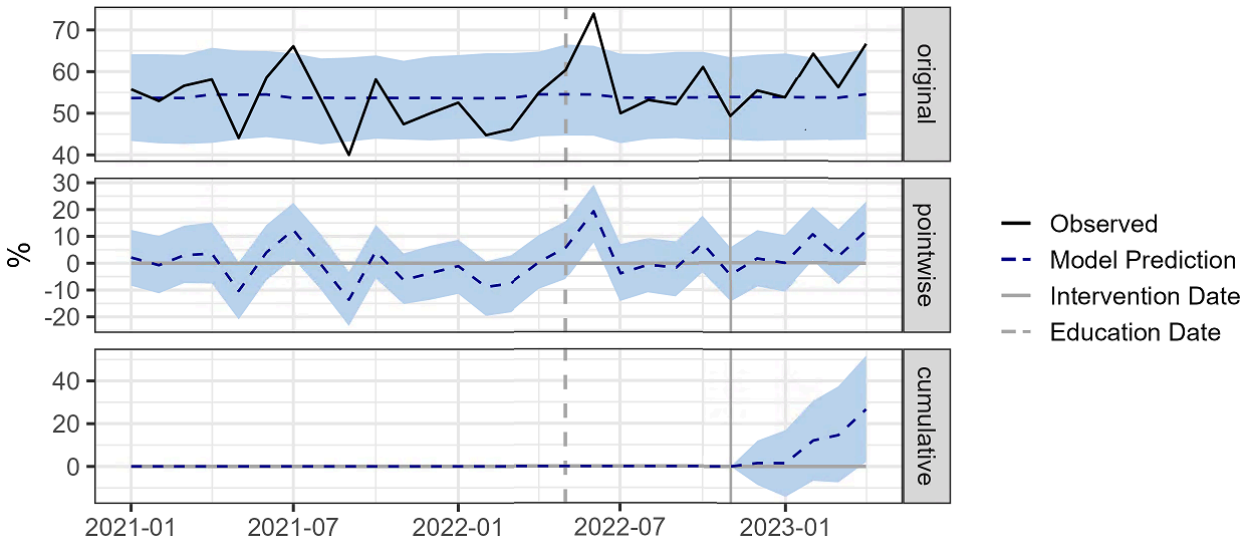


Figure S2.17: Rate of administration of 30cc/kg of fluids within 3 hours of presentation of septic shock or hypotension.
Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

Supplementary Note 2.8. Causal impact analysis of change in sequential organ failure assessment (SOFA) score by emergency department.

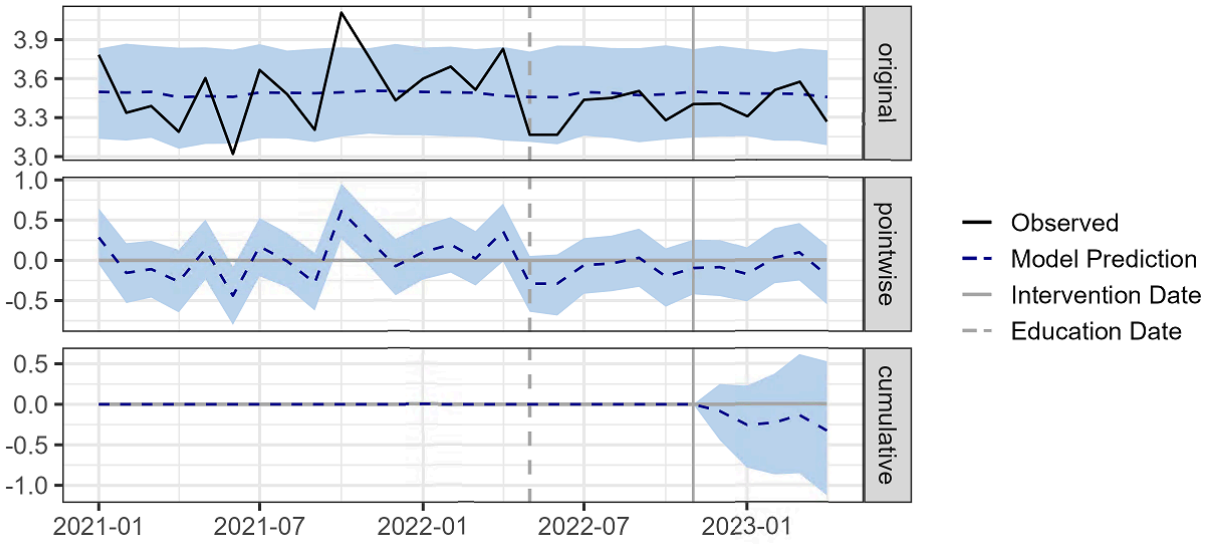


Figure S2.18: Causal impact analysis of change in sequential organ failure assessment (SOFA) score within 72 hours of sepsis onset at the La Jolla Emergency Department.
Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

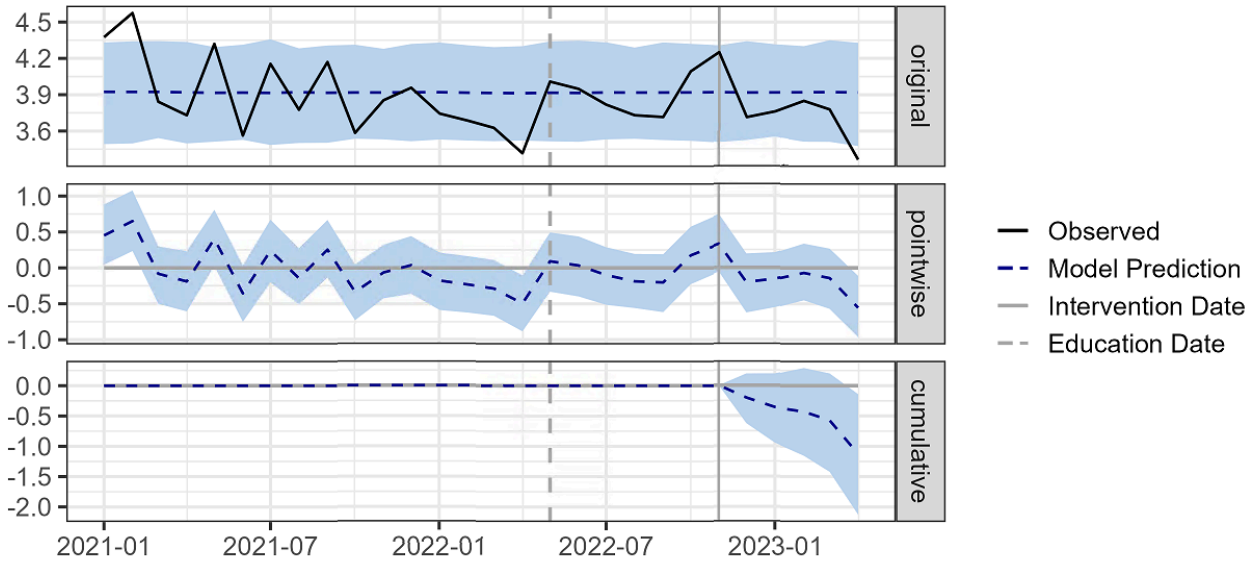


Figure S2.19: Causal impact analysis of change in sequential organ failure assessment (SOFA) score within 72 hours of sepsis onset at the Hillcrest Emergency Department.
Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

Supplementary Note 2.9. Causal impact analysis of secondary outcomes.

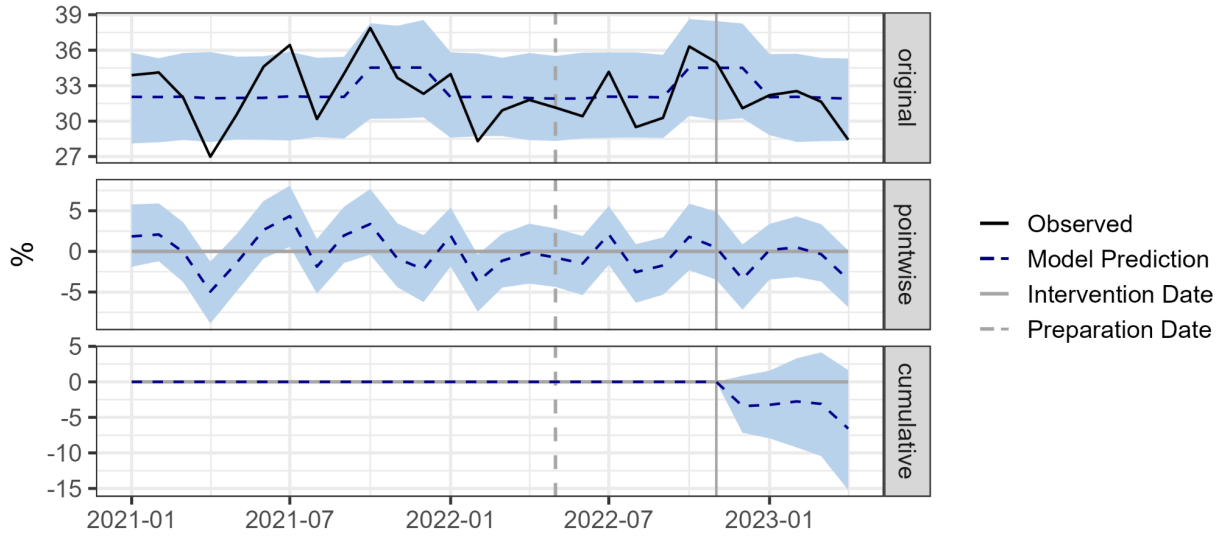


Figure S2.20: Causal impact analysis of sepsis intensive care unit (ICU) admissions from the emergency department.

Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

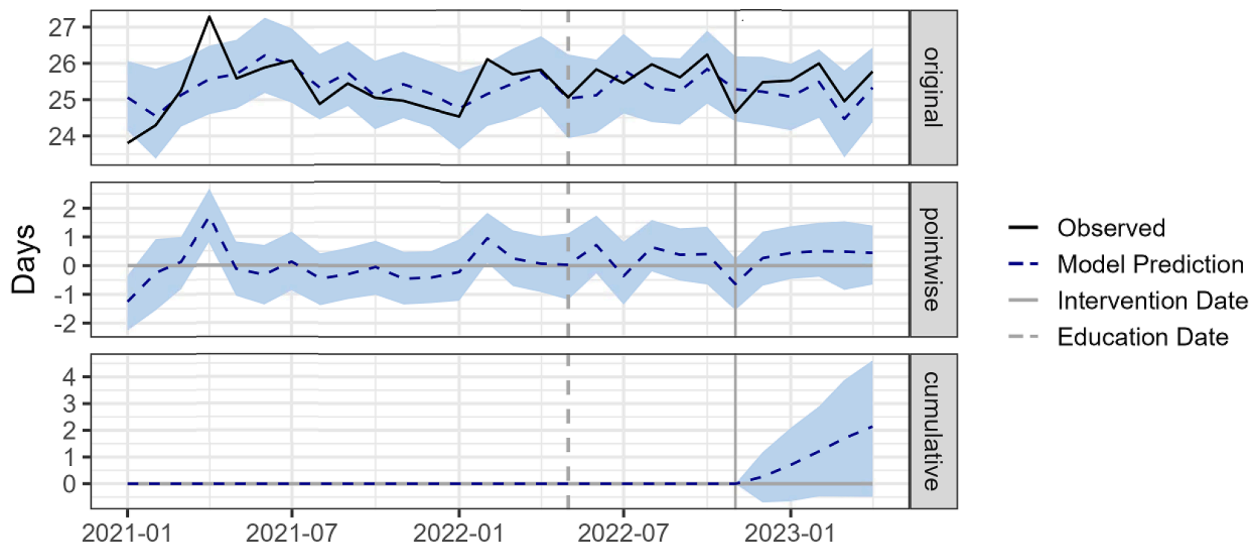


Figure S2.21: Causal impact analysis of intensive care unit (ICU)-free days.
Plots of the causal impact analysis using a Bayesian structural time-series model. The top subpanel ("original") shows the actual outcome (black) and the average model predictions (dashed blue) and 95% confidence limits (shaded blue) during the pre-intervention and post-intervention periods, indicated by the solid gray vertical line. The middle subpanel ("pointwise") shows the difference between the model predictions and the observed outcome. The bottom subpanel ("cumulative") shows the sum of the pointwise differences during the post-intervention period.

Supplementary Note 2.10. Trends of covariates used in causal impact analyses.

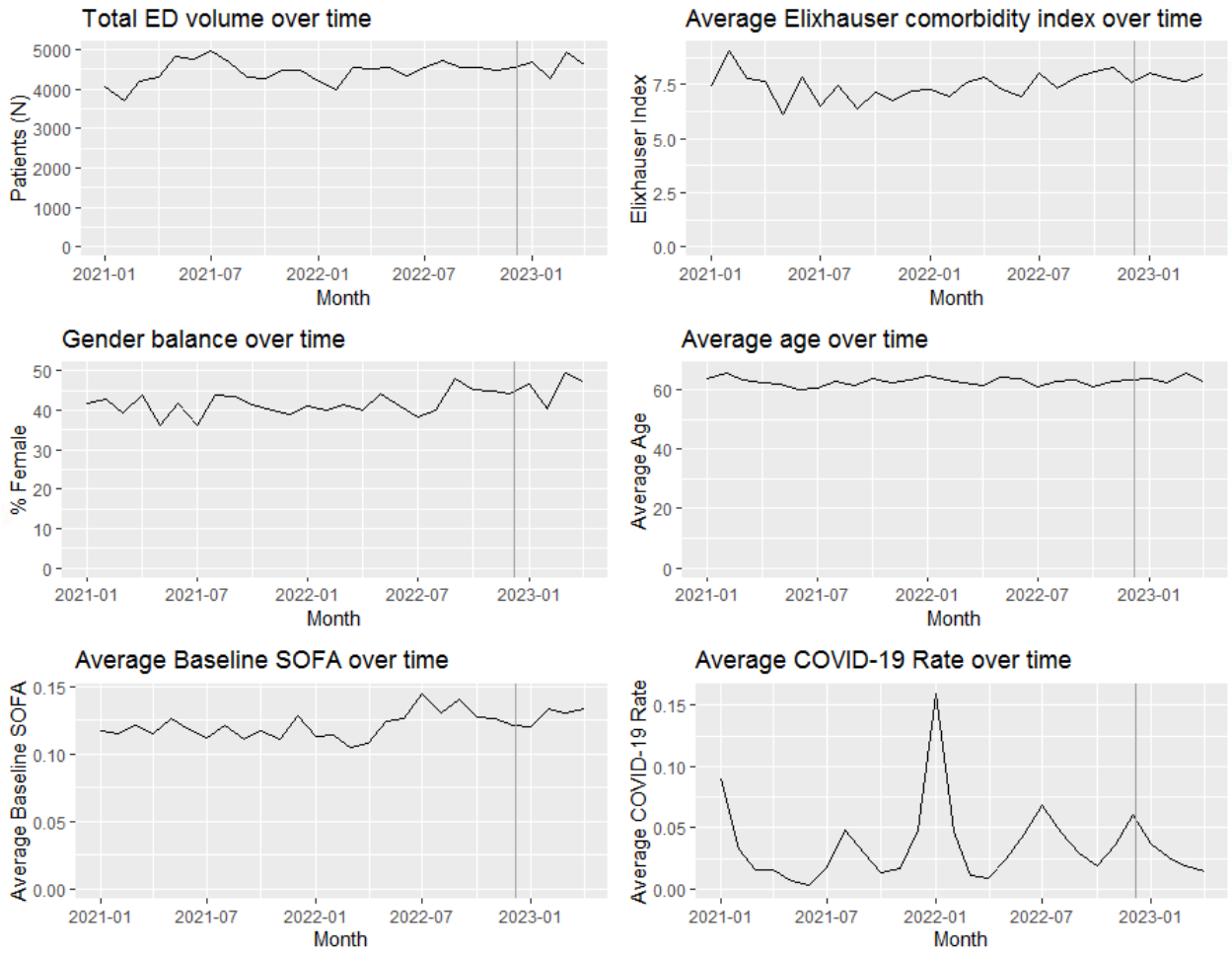


Figure S2.22: Temporal trends of covariates used in the causal impact analyses.
ED = emergency department. SOFA = sequential organ failure assessment score.

Chapter 3 Quantifying Healthcare Provider Perceptions of a Novel Deep Learning Algorithm to Predict Sepsis: Electronic Survey

INTRODUCTION

The clinical adoption of predictive models, particularly in sepsis, has faced resistance due to concerns about the accuracy and relevance of the alerts, alongside a general skepticism towards algorithmic decision-making.[50,104] Poorly designed interfaces and inferior implementation strategies may lead to disregard of the system-generated advice.[105.106,107] Despite the importance of user experience for the adoption and efficacy of clinical predictive models, there are relatively few studies focused on provider acceptance and feedback on the deployment of such models. Henry et al. found that emergency department (ED) providers and providers who had previously seen an alert were more likely to perform a follow-up evaluation, but did not evaluate perceived provider or nurse diagnostic utility.[51] Ginestra et al. directly evaluated provider perception of a sepsis predictive model and found significant differences in acceptance between nurses and physicians for a machine learning-based algorithm for inpatient sepsis .[108] In this chapter, we aim to evaluate healthcare worker perception and acceptance of the deep learning model for the prediction of sepsis in the ED described in Chapter 2. We test the hypothesis that nurses and less experienced providers perceive greater utility in the sepsis model than more experienced physicians.

METHODS

Development of the Survey Tool

To assess perceptions of COMPOSER by providers and nurses, we developed and internally validated a survey based on Ginestra et al. in accordance with the CHERRIES checklist.[109] This was a closed survey sent to a convenience sample of team members taking care

of a patient for whom a COMPOSER alert crossed the pre-defined threshold. For nurses, this corresponded to a recent Best Practice Advisory (BPA) (Figure 2.4). For physicians and advanced practice providers (APP), COMPOSER flowsheet entries were displayed on the ED trackboard (Figure 3.1). The UC San Diego Institutional review board approval was obtained with waiver of informed consent (#805726) and additional approval was obtained from the Aligning and Coordinating QUality Improvement, Research, and Evaluation (ACQUIRE) Committee (Project #609). Participants were informed of the primary investigator and the purpose of the study during recruitment in the electronic health record (Epic Systems, Verona, WI) via secure chat message during pre-selected times from several study authors. No identifiable personal data were collected. The survey was developed and tested by key stakeholders across provider types prior to distribution to respondents. This was a closed survey where participants were contacted via secure message, and was not otherwise advertised. Completion was voluntary, and not otherwise compensated. Recruitment occurred between May and September 2023 and was administered using a vendor survey tool (Qualtrics). All items except optional commentary feedback were mandatory. Participants were able to edit their responses until submitting, but not afterwards.

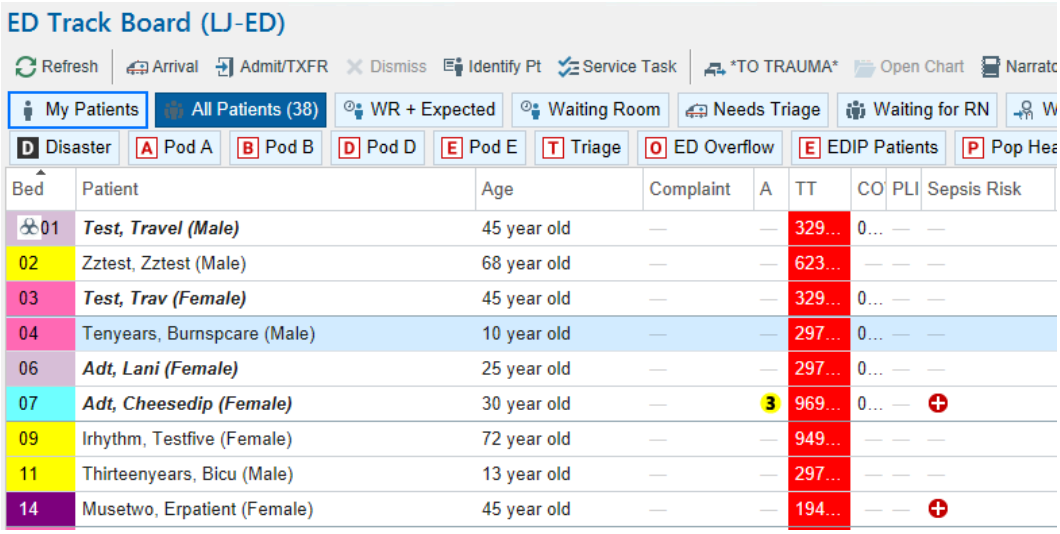


Figure 3.1: User interface for physicians and APPs.

Statistical Methods

Data were analyzed with R version 4.1.2 (R Foundation for Statistical Computing). Descriptive statistics are provided as indicated. Survey responses were analyzed using Mann-Whitney-U testing to evaluate if pre-alert sepsis suspicion, post-alert sepsis suspicion, post-alert change in management, and overall perceived utility varied by provider type and experience. Experience level was categorized a priori as either ≤5 years or > 5 years. A p value < 0.05 was considered significant in all analyses. We did not have any missing data.

RESULTS

The survey was distributed to nearly 250 providers who were exposed to the COMPOSER intervention. A total of 114 responses were received with 76 from MD/DOs, 34 from RNs, and 4 from NP/PAs with an overall response rate of roughly 50%, although the exact rate may vary slightly due to the possibility of the same provider receiving the survey more than once (Figure 3.2). 53% of respondents were from providers with less than 5 years of experience.

Of the 76 physician responses, 49 were from trainees. 38% of respondents had a positive perception of the alert's usefulness, while 39% had a neutral perception , and only 23% found it not useful. Nurses were more likely to think a patient had sepsis prior to an alert than providers (p=.007). Moreover, providers with 0-5 years of experience were more likely to have increased expectation that a patient had sepsis after the alert (p=.021), and were more likely to say that the alert was useful in care of a patient (p=.016), than those with 6+ years of experience. Finally, physicians with 0-5 years of experience were more likely to say that the alert changed their management of the patient (p=.048) than physicians with 6+ years of experience (Figure 3.3). Overall, 14% of physicians with 0-5 years of experience noted they changed their management, whereas none with 6+ years did so. Response distribution to key questions by provider type and experience are displayed in Figures 3.4 and 3.5 respectively.

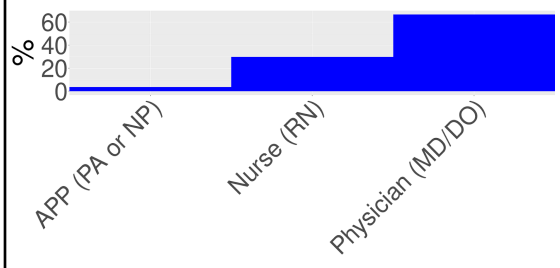
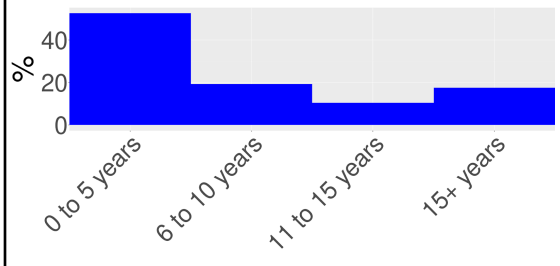
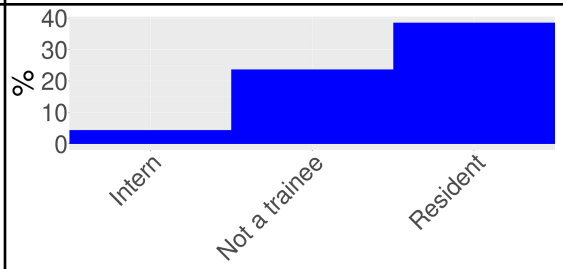
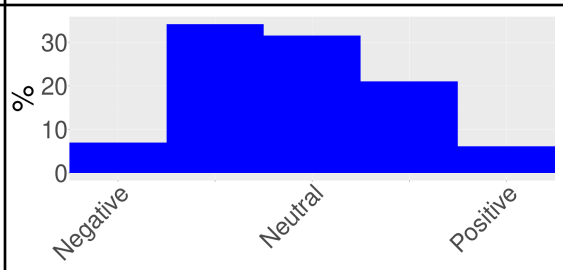
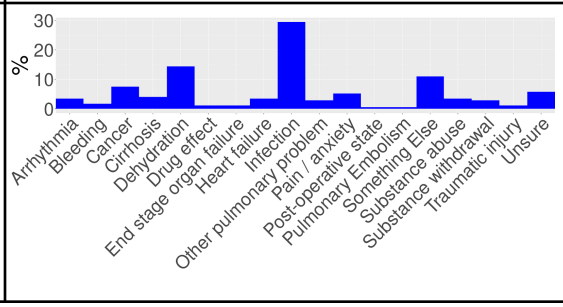
Question Number & Text	Responses
1: What is your clinical role?	
2: How long have you been working in the emergency department after receiving your terminal degree (M.D., R.N., P.A., N.P.)?	
3: If you are a physician, are you a trainee?	
4: Before the alert, did you think the patient had sepsis?	
5: What do you think caused the alert to fire?	

Figure 3.2: Full Survey Results.

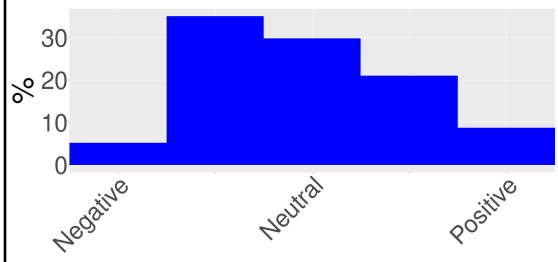
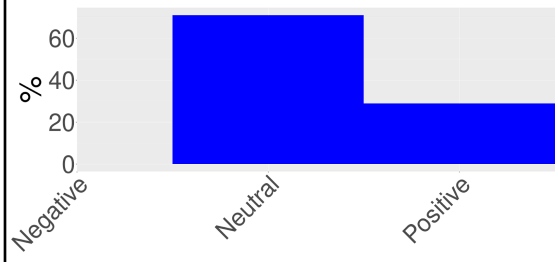
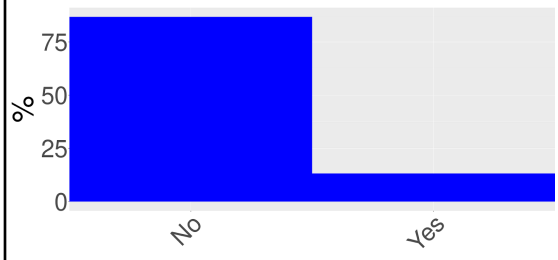
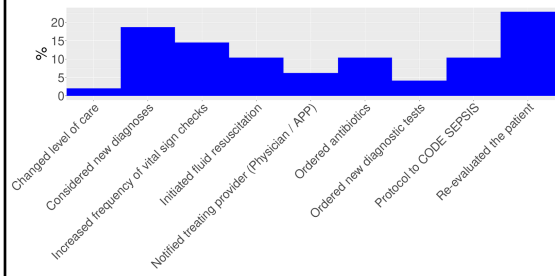
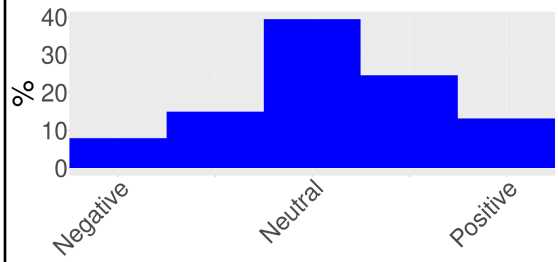
Question Number & Text	Responses
6: After the alert, did you think the patient had sepsis?	
7: Did the alert affect your expectation that the patient would develop sepsis?	
8: Did the alert change your management of the patient?	
9: How did the alert change your management of the patient?	
10: How much do you agree with this statement: "This alert was useful in the care of this patient"	

Figure 3.2: Full Survey Results (Continued).

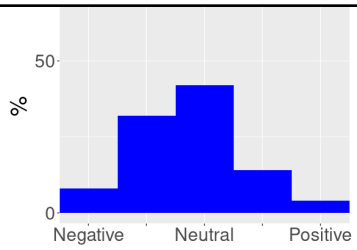
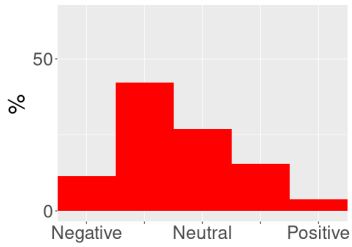
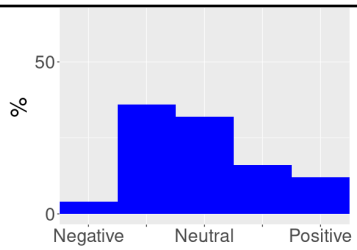
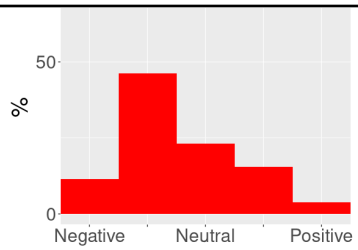
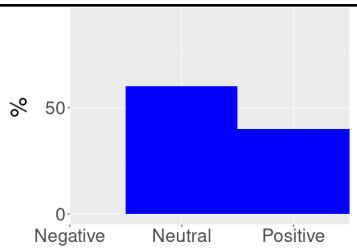
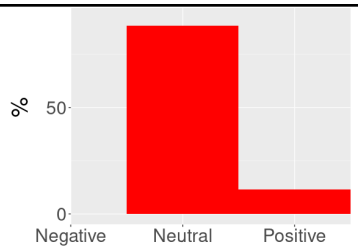
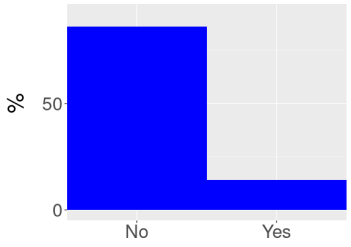
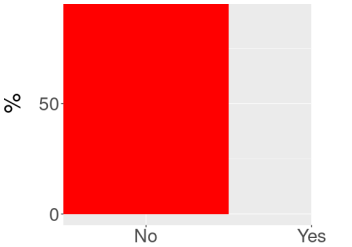
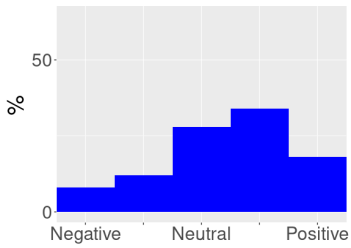
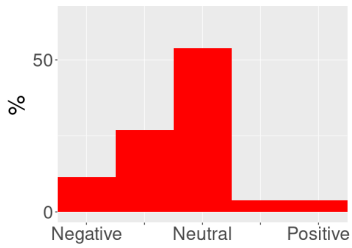
	0 - 5 Years	6+ Years	*P* value[a]
Before the alert, do you think the patient had sepsis?			.413
After the alert, did you think the patient had sepsis?			.109
Did the alert affect your expectation that the patient would develop sepsis?			**.011**
Did the alert change your management of the patient?			**.048**
How much do you agree with this statement: "This alert was useful in the care of this patient"			**.002**

Figure 3.3: Survey Results from Providers by Experience Level.
[a]*P* values are based on Mann-Whitney-U tests between groups on each row. $P < 0.05$ indicates significance.

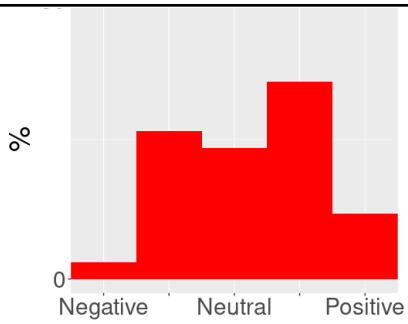
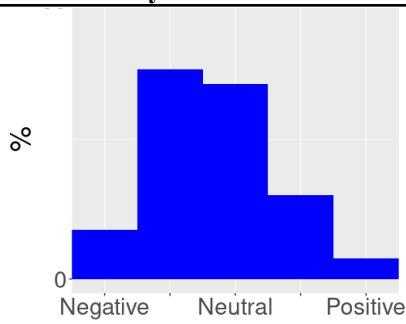
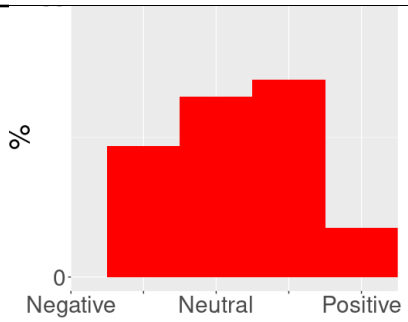
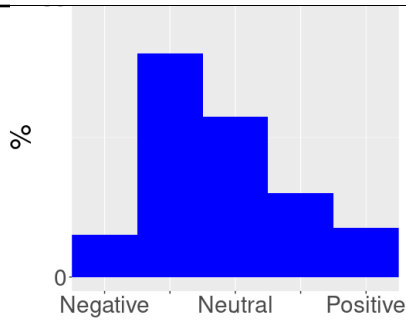
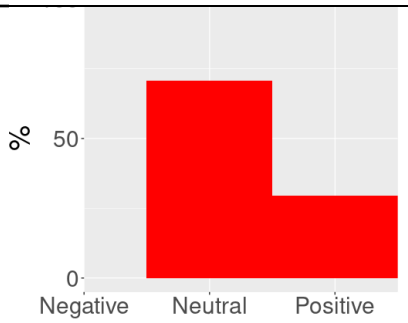
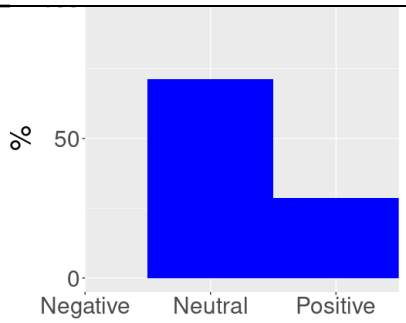
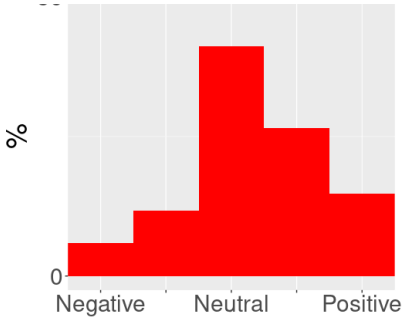
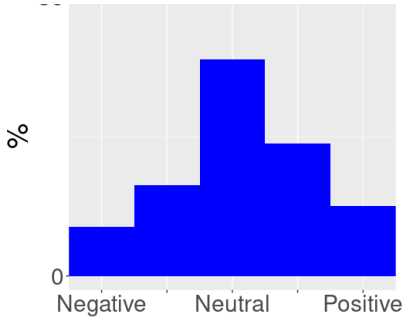
	Nurses	Physicians & APPs	*P* value[a]
Before the alert, do you think the patient had sepsis?			**.007**
After the alert, did you think the patient had sepsis?			**.011**
Did the alert affect your expectation that the patient would develop sepsis?			.947
How much do you agree with this statement: "This alert was useful in the care of this patient"			.449

Figure 3.4: Key Survey Results by Provider Type.

[a]*P* values are based on Mann-Whitney-U tests between groups on each row. $P < 0.05$ indicates significance.

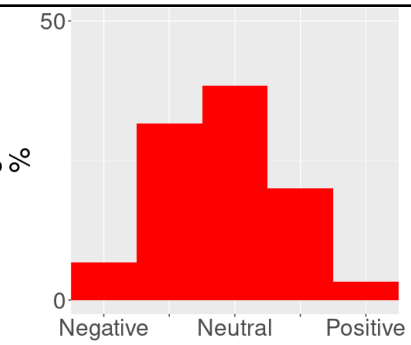
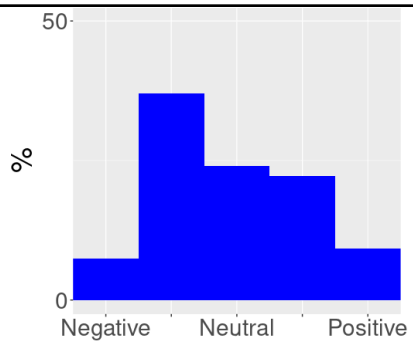
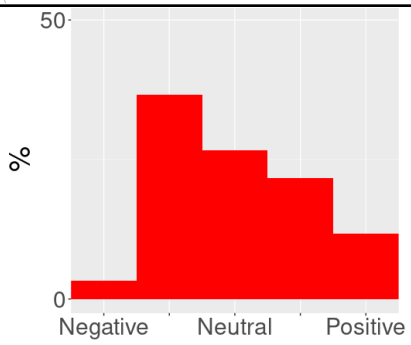
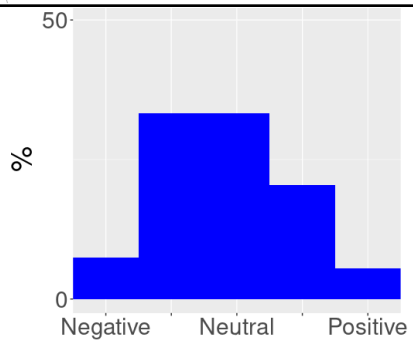
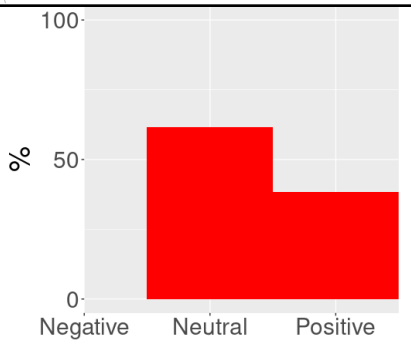
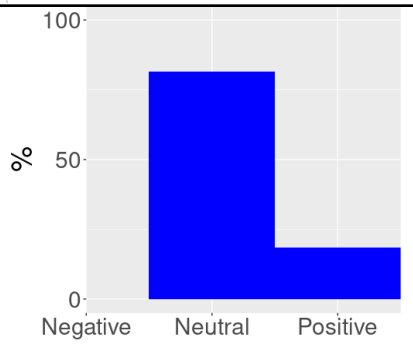
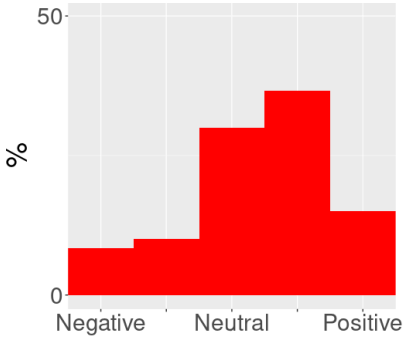
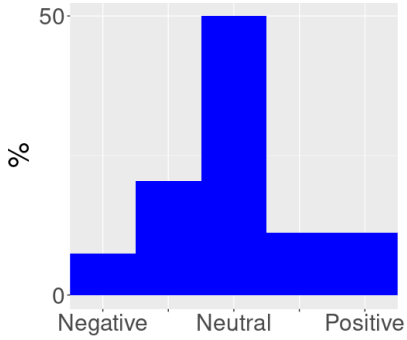
	0 - 5 Years	6+ Years	*P* value[a]
Before the alert, do you think the patient had sepsis?			.882
After the alert, did you think the patient had sepsis?			.461
Did the alert affect your expectation that the patient would develop sepsis?			**.021**
How much do you agree with this statement: "This alert was useful in the care of this patient"			**.016**

Figure 3.5: Key Survey Results from Providers by Experience Level.
[a]*P* values are based on Mann-Whitney-U tests between groups on each row. $P < 0.05$ indicates significance.

DISCUSSION

Our findings represent the first study of providers' perception of alerts generated by a deep learning sepsis model. As there are few deep learning models deployed in routine clinical practice, our findings are significant for a variety of reasons. To begin with, we found that nurses and less experienced providers (5 years of experience) derive greater perceived benefit from the alert. Providers with less experience indicated that they were more likely to believe the model was accurate in prediction of sepsis.[110] Additionally, less experienced physicians indicated they were more likely to make changes to their management based on the alert when compared to more experienced physicians. We believe that a 14% rate of indicated change in management is significant and is a potential major contributor to the decrease in mortality we have described previously.[35] Finally, we found that nearly 80% of respondents had a positive or neutral view of the alert system.

With increasing data that predictive models may be cost effective and improve patient centered outcomes, our data may inform effective implementation strategies.[21,35,111] First, recognition that perceived benefit may vary significantly amongst provider groups may impact implementation and education strategies. For example, our data suggest additional education targeted to more experienced providers and nurses may be indicated. Second, there are increasing data that shared mental models of risk between physicians and non-physicians may be an elegant strategy to improve care and minimize risk in complex medical scenarios through objective assessment.[92] Our findings demonstrate that nurses had greater perceived benefit of our model in the diagnosis of sepsis, and although speculative, this finding may further increase nurse utilization of the model for patient care. As our recent data demonstrate an improvement in patient-centered outcomes and quality metrics with this model, further empowering

non-providers to use this model to voice concerns about possible sepsis based on an objective assessment of risk may further improve patient-centered outcomes.[35]

We posit that there are important intersections with generational attitudes towards technology, as well as in the fundamental way that different types of healthcare workers interact with alerts. Provider and nurse experience is highly correlated with age and the general acceptance of a machine learning alert may align with age groups with greater native exposure to such technologies. More experienced physicians may have additional clinical experience to anticipate better which patients may develop sepsis, whereas such slight presentations may be more likely to be missed by those with less experience. Additionally, we used proactive and passive methods of alert notification (Figure 3.1). This approach may impact perceived utility since our nurses received a BPA alert, a much more visible alerting method compared to a provider-facing trackboard display which is a more passive approach to drawing attention to at-risk patients. However, while less experienced physicians and nurses overall perceived greater benefit, care should be taken when considering exclusive deployments to these groups given the possible risk of automation bias.[112] Overall, we believe that a combinatorial effect of generational exposure to machine learning and the more obvious alert to nursing staff may account for the differences observed in this study. Furthermore, future implementations may leverage this knowledge with careful design of clearly displayed and explainable decision support prompts and education to all staff regarding the alert and its purpose.

We acknowledge several limitations. First, we had a limited sample size with moderate response rate. Additionally, there may be response bias. We were not able to track changes in management based on alerts alongside survey responses to evaluate how treatment may have

been altered by the alert regardless of provider perception. Finally, different notification processes which were used for nurses and providers may impact perceptions of our alert system.

CONCLUSION

Less experienced providers and nurses were more likely to perceive benefit from the alert, and the alerts were overall received favorably. Future clinical AI model implementations might consider focused alert patterns and education to improve reception and reduce fatigue.

ACKNOWLEDGEMENTS

Chapter 3, in full, has been submitted for publication of the material as it may appear in JMIR AI as “Quantifying Healthcare Provider Perceptions of a Novel Deep Learning Algorithm to Predict Sepsis: Electronic Survey”. Boussina, A., Ramesh, K., …, & Wardi, G. The dissertation author was the primary investigator and author of this paper along with co-first author Karthik Ramesh.

Chapter 4 Representation Learning and Spectral Clustering for the Development and External Validation of Dynamic Sepsis Phenotypes: Observational Cohort Study

Recent attempts at clinical phenotyping for sepsis have shown promise in identifying groups of patients with distinct treatment responses. Nonetheless, the replicability and actionability of these phenotypes remain an issue because the patient trajectory is a function of both the patient's physiological state and the interventions they receive. In this chapter, we aim to develop a novel approach for deriving clinical phenotypes using unsupervised learning and transition modeling. Forty commonly used clinical variables from the electronic health record are used as inputs to a feed-forward neural network trained to predict the onset of sepsis. Using spectral clustering on the representations from this network, we derive and validate consistent phenotypes across a diverse cohort of patients with sepsis. We model phenotype dynamics as a Markov decision process with transitions as a function of the patient's current state and the interventions they received.

Four consistent and distinct phenotypes are derived from over 11,500 adult patients who were admitted from the University of California, San Diego emergency department (ED) with sepsis between January 1, 2016, and January 31, 2020. Over 2000 adult patients admitted from the University of California, Irvine ED with sepsis between November 4, 2017, and August 4, 2022, were involved in the external validation. We demonstrate that sepsis phenotypes are not static and evolve in response to physiological factors and based on interventions. We show that roughly 45% of patients change phenotype membership within the first 6 hours of ED arrival. We observe consistent trends in patient dynamics as a function of interventions including early administration of antibiotics.

INTRODUCTION

Attempts to improve care and decrease mortality in sepsis patients have focused on tools designed to provide timely and protocolized care. However, there is concern that this approach may be suboptimal for some patients and does not address the heterogeneity of the condition.[29,30] Clinical phenotypes, or subtypes, can be best defined as distinct groups of patients with similar laboratory abnormalities, organ dysfunction, and outcomes.[113,114,115] Efforts to identify clinical sepsis phenotypes have gained interest recently but have remained limited in clinical applications. The identification of clinical phenotypes early in sepsis may allow for more precision care as providers may tailor therapies (i.e. quantity of fluid resuscitation or timing of vasoactive medications) based on a particular phenotype. Seymour et al. described 4 novel distinct clinical phenotypes of sepsis that correlated with host response patterns and clinical outcomes.[31] However, phenotyping using unsupervised clustering on high-dimensional features is sensitive to noise, which complicates the generalizability of the associated phenotypes. [54,55,56] The common sources of variations and noise include differences in local populations, electronic medical record systems, laboratory equipment and assays, frequency of data measurement, and variations in clinical workflows and administrative practices, which collectively result in data distribution shifts. Rather than relying on arbitrary criteria (ie, quick Sequential Organ Failure Assessment [SOFA] or Systemic Inflammatory Response Syndrome criteria), representation learning provides a principled strategy for deriving informative compressed summaries or encodings from underlying clinical data.[116,117] This approach allows for dimensionality reduction and developing features that can then be used for downstream clustering and phenotype discovery.

Sepsis phenotypes may provide an opportunity for more personalized care and improving patient outcomes. The examples of individualized care that may be afforded by the identification of phenotypes may include varying quantities of crystalloid during resuscitation, earlier use of vasoactive medications, and earlier times for antibiotic administration. Differences in prognosis and treatment response between phenotypes may have a major clinical impact, and it would be important to identify such phenotypes soon after hospital presentation to guide treatment. Importantly, it is uncertain if these phenotypes are "static," namely, constant throughout a patient's stay, or "dynamic," and sepsis interventions, such as timely antibiotic administration and fluid resuscitation, may impact phenotype membership.

We seek to classify patients meeting sepsis 3 criteria in the emergency department (ED) within the first 6 hours of arrival into distinct clinical phenotypes and compare the differential impact of sepsis therapies among phenotypes on patient-centered outcomes. We also test the hypothesis that sepsis phenotypes are dynamic, namely, that transitions between phenotypes are common over time and early sepsis interventions may impact transitions between phenotypes.

METHODS

Study Design and Setting

We conducted a retrospective observational study consisting of all adults (aged ≥18 years) who were admitted from the ED with sepsis from the University of California, San Diego (UCSD) between January 1, 2016, and January 31, 2020, and the University of California, Irvine (UCI) between April 11, 2017, and April 8, 2022. UCSD consists of 2 EDs with a total annual census of 70,000 patients; one ED receives patients at a quaternary care center, whereas the other

functions in a “safety net” hospital. UCI has a single ED with a total annual census of 50,000 patients.

Ethics Approval

We followed recommendations provided by STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines to ensure appropriate reporting of this research.[94] UCSD institutional review board approval was obtained with waiver of informed consent (#800257). Data abstracted from UCI were provided under Data Use Agreement (#37533).

Selection of Participants

All patients admitted to the hospital from the ED were automatically queried for the presence of sepsis in the electronic health record (EPIC). Data, including demographics, comorbidities, vital signs, laboratory results, medications administered (including antimicrobials and intravenous fluids), length of stay, and outcomes, were abstracted into a HIPAA (Health Insurance Portability and Accountability Act)-compliant enclave for analysis. Also included in this repository was each patient’s SOFA and Charlson Comorbidity Index. Sepsis was defined according to the Third International Consensus Definition (“sepsis 3”) as clinical suspicion of infection plus end-organ damage, represented by a 2-point change in the SOFA score.[2] Clinical suspicion of infection was defined by an emergency medicine physician ordering blood cultures and administration of at least 3 days of antibiotic therapy (excluding prophylactic use). The time of sepsis was defined as the time of clinical suspicion of infection. To accomplish this, we used the earliest timestamp of either administration of intravenous antibiotics or checking blood

cultures. We included patients who met sepsis criteria within 6 hours of ED arrival and were admitted to the hospital. We excluded patients where the time of sepsis criteria was detected after evaluation in the ED; those who developed sepsis >6 hours after ED triage; those who discharged from the ED, died, or were transferred to another hospital within 6 hours of triage; and patients who had intravenous antibiotics administered for less than 3 days.

Data Abstraction, Missingness, and Processing

Data were abstracted at the time of the first measurement in the ED. Demographic and other static data (ie, sex, weight, etc) were held constant after initial abstraction. Dynamic data points, such as blood pressure and laboratory values, were sampled and updated at hourly intervals. If a bin contained multiple values within an hour, we selected the median value. We used a sample-and-hold interpolation if data were missing. If a bin did not have any value imputed during stay in the ED, we chose the global population mean value and inserted that value in the appropriate time bin. The list of input features is provided in Table S4.1 in the Supplementary Materials. All data features first underwent normality transformations and then underwent standardization by subtracting the mean and dividing by the SD. Continuous variables are provided as medians with 25% and 75% IQRs. Binary variables are provided as percentages.

Statistical Methods

Forty commonly used clinical variables including demographic data, vital signs, and laboratory results (see Table S4.1 in the Supplementary Materials for a complete list of variables) were used for cluster derivation. We considered two time points per patient: (1) 3 hours after ED triage and (2) 6 hours after ED triage. We used a spectral clustering algorithm with Gaussian

similarity due to its strong clustering performance on nonconvex regions.[118] In spectral clustering, data are represented by nodes in a graph with weighted edges corresponding to the similarity between nodes. Clustering is then performed on the top eigenvectors from a matrix derived from the edges. This approach recovers the largest connected components on the graph, and therefore a cluster corresponds to tightly connected subgraphs. A key advantage of spectral clustering is that it can recover clusters in nonlinear manifolds where simpler K-means clustering fails. We combined the spectral clustering method with the consensus clustering approach described by Seymour et al to perform the clustering algorithm in an inner loop on random subsamples of the data to determine the consistency of the clustering results across varying values of the cluster number.[31] We performed this analysis on both the preprocessed clinical features and the lower dimensional representations of these data from a feed-forward neural network trained to predict the onset of sepsis (Figures S4.1-S4.4 in the Supplementary Materials). The neural network architecture and performance are described in detail in the paper by Shashikumar et al.[20]

To investigate the dynamics of the derived phenotypes, we examined the trajectory of patients from 3 to 6 hours after ED triage. We modeled changes in phenotypes as a Markov decision process and calculated transition probabilities between phenotypes (or states) as a function of the patient's current state and a vector of actions (eg, the administration of hemodynamic resuscitation parameters and antibiotics). Specifically, we fit logistic regression models to each potential state transition using physiological and interventional features. Interventional features included the administration of antibiotics within 3 hours of ED arrival, the administration of fluids within 3 hours of ED arrival, and whether the volume of fluids administered was ≥30 mL/kg of body weight. These binary features were selected because they

imitate elements of the SEP-1 bundle and can capture nonlinearities in the underlying continuous measurements. Statistical methods are described further in Supplementary Note 4.1.

Results were validated against an external cohort of patients with sepsis from a second clinical site (UCI). Phenotype membership was assigned using the k-nearest neighbor algorithm with the same Gaussian similarity metric as used in phenotype derivation. Phenotype characteristics and dynamics were then compared between the derivation and validation set to ensure consistency of the phenotyping. All analyses were performed using R version 4.0.4 (R Foundation for Statistical Computing).

RESULTS

Characteristics of Study Subjects

We identified 11,519 patients who met criteria for sepsis within 6 hours of UCSD ED arrival between January 1, 2016, and January 31, 2020, and 2091 patients who met criteria for sepsis within 6 hours of UCI ED arrival between April 11, 2017, and April 8, 2022. Table 4.1 shows baseline characteristics and therapies administered to the patients included in this investigation. Most patients had some degree of chronic comorbidity (median Charlson Comorbidity Index of 2) and acute organ dysfunction (median SOFA score of 2). Almost all patients (10,094/11,519, 87.6%) received antibiotic therapy within 6 hours of identification of ED arrival. The median time to administration of antibiotics was 2.2 hours (range 1.12-4.05 hours) from ED triage. The median time to initiation of fluids was 0.87 hours, and the average quantity of fluid administered within 3 and 6 hours was 1084 mL and 1825 mL, respectively. Overall, in-hospital mortality and transition to hospice was 6.9% (800/11,519). Patients at the external validation site had similar characteristics (Table 4.1). Most patients received antibiotic therapy within 6 hours (median time

of 2.52 hours). The median time to fluid administration was 1.29 hours. Overall, in-hospital mortality and transition to hospice was 13.4% (280/2091).

Table 4.1: Demographics, clinical characteristics, and therapies of patients with sepsis.

Variable	UCSD (development site)	UCI (validation site)
Characteristics		
Patients, n	11,519	2091
Age (years), mean (SD)	61 (17.7)	60 (17)
Sex, n (%)		
Male	6509 (56.51)	1152 (55.09)
Female	5010 (43.49)	939 (44.91)
Organ dysfunction, median (IQR)		
Charlson Comorbidity Index	2 (0-4)	2 (0-4)
SOFA score (maximum within 3 hours)	2 (1-3)	2 (1-4)
Interventions		
Antibiotics within 3 hours, n (%)	7250 (62.94)	1228 (58.73)
Antibiotics within 6 hours, n (%)	10,094 (87.63)	1805 (86.32)
Fluids, n (%)	11,512 (99.94)	1546 (73.94)
Time to antibiotics (hours), median (IQR)	2.2 (1.12-4.05)	2.52 (1.28-4.38)
Time to fluids (hours), median (IQR)	0.87 (0.433-1.85)	1.29 (0.45-5.66)
Fluid intake (mL), first 3 hours, median (IQR)	1084.1 (266.56-2199.53)	500 (0-1000)
Fluid intake ≥30 cc/kg, n (%)	3034 (26.34)	164 (7.84)
Steroid administration, n (%)	689 (5.98)	135 (6.46)
Laboratory values, median (IQR)		
Lactate (mmol/L)	2.2 (1.5-3.1)	2.2 (1.5-3.1)
Outcomes, n (%)		
Mortality	715 (6.21)	108 (5.16)
Hospice	85 (0.74)	172 (8.23)
Septic shock	1253 (10.88)	230 (10.99)
Acute kidney injury	3479 (30.20)	819 (39.17)
Mechanical ventilation	993 (8.62)	181 (8.66)

Phenotype Characteristics

Consensus clustering outputs demonstrated more consistent clusters from spectral clustering on the lower dimensional representations than clustering on the preprocessed clinical features (Figures S4.1-S4.4). Therefore, the representation-based clusters were used to identify sepsis phenotypes. Additionally, the consensus cumulative density function for clustering on the representations determined an optimal value of k=4 (Figure S4.2). These 4 distinct sepsis phenotypes (P1, P2, P3, and P4) are shown in Table 4.2 and Table S4.3. Cluster membership 3 hours after ED arrival ranged from 13.6% (1563/11,519; P2) to 36% (4152/11,519; P4), and significant variability in the clinical characteristics of patients and rates of adverse outcomes was observed across the 4 phenotypes. P3 exhibited the worst outcomes in terms of in-hospital mortality (179/1912, 9.4%), progression to septic shock (458/1912, 24%), development of acute kidney injury (AKI; 708/1912, 37%), and need for mechanical ventilation (225/1912, 11.8%). Conversely, P2 exhibited the lowest rate of in-hospital mortality (26/1563, 1.7%), septic shock (37/1563, 2.4%), AKI (393/1563, 25.1%), and mechanical ventilation (65/1563, 4.2%). Chord diagrams of the impacted body systems across these phenotypes show that the derived phenotypes express complex differences across nearly all systems (Figure S4.5). Further, cluster membership is not solely a function of missingness (Table S4.5). The phenotype characteristics of the k-nearest neighbor assigned clusters for patients in the validation cohort demonstrate similar patterns of outcomes and clinical variables (Tables S4.3 and S4.4).

Table 4.2: Characteristics and therapeutics of the 4 sepsis phenotypes at the development site.

Variable	P1	P2	P3	P4	Total	*P* value[a]
Characteristic						
Number of patients, n (%)	3892 (33.79)	1563 (13.57)	1912 (16.60)	4152 (36.04)	11,519 (100)	N/A[b]
Age (years), mean (SD)	61 (17.5)	63 (17.7)	60 (17.9)	59 (17.7)	61 (17.7)	<.001
Sex, n (%)						<.001
Male	2164 (55.60)	819 (52.40)	1178 (61.61)	2348 (56.55)	6509 (56.51)	N/A
Female	1728 (44.40)	744 (47.60)	734 (38.39)	1804 (43.45)	5010 (43.49)	N/A
Organ dysfunction, n (%) or median (IQR)						
Charlson Comorbidity Index	2 (1-4)	2 (0-4)	3 (1-5)	2 (1-5)	2 (1-4)	<.001
Congestive heart failure component	523 (13.44)	211 (13.50)	197 (10.30)	444 (10.69)	1375 (11.94)	<.001
Moderate or severe liver disease component	170 (4.37)	45 (2.88)	121 (6.33)	184 (4.43)	520 (4.51)	<.001
Renal disease component	653 (16.78)	213 (13.63)	405 (21.18)	688 (16.57)	1959 (17.01)	<.001
SOFA[c] score (maximum within 3 hours)	2 (0-3)	1 (0-2)	3 (1-5)	2 (1-4)	2 (1-3)	<.001
SOFA score (maximum within 6 hours)	2 (1-3)	1 (0-2)	3 (1-5)	2 (1-4)	2 (1-4)	<.001
Antibiotics, n (%) or median (IQR)						
Antibiotics within 3 hours	2056 (52.83)	579 (37.04)	1652 (86.40)	2963 (71.36)	7250 (62.94)	<.001
Antibiotics within 6 hours	3252 (83.56)	1217 (77.86)	1841 (96.29)	3784 (91.14)	10,094 (87.63)	<.001
Time to antibiotics (hours)	2.82(1.483-4.85)	3.73(2.25-5.633)	1.22 (0.733-2.133)	1.87 (1.017-3.317)	2.2 (1.117-4.05)	<.001
Fluids, n (%) or median (IQR)						
Fluids	3891 (99.97)	1557 (99.62)	1912 (100)	4152 (100)	11,512 (99.99)	<.001
Time to fluids (hours)	1.05 (0.517-2.25)	1.85 (0.783-3.45)	0.63 (0.35-1)	0.75 (0.383-1.417)	0.87 (0.433-1.85)	<.001
Fluid intake (mL), first 3 hours	886.7 (98.61-1700)	233.3(0-1000)	2153.5 (1121.88-2982.53)	1380.6 (556.81-2393.94)	1084.1 (266.56-2199.53)	<.001
Vasopressors, n (%)						
Pressors	253 (6.50)	37 (2.37)	458 (23.95)	505 (12.16)	1253 (10.88)	<.001
Corticosteroids,[d] n (%)						
Steroids	120 (3.08)	27 (1.73)	96 (5.02)	139 (3.35)	382 (3.32)	<.001

Table 4.2: Characteristics and therapeutics of the 4 sepsis phenotypes at the development site. (Continued)

Variable	P1	P2	P3	P4	Total	*P* value[a]
Laboratory values, n (%) or median (IQR)						
Subjects with lactate measurements	2631 (67.60)	556 (35.57)	1863 (97.44)	3679 (88.61)	8729 (75.78)	<.001
Lactate (mmol/L)	1.9 (1.4-2.6)	1.7 (1.4-2.2)	2.6 (2-3.7)	2.2 (1.5-3.2)	2.2 (1.5-3.1)	<.001
Platelets (1000/μL)	221 (146-305)	212.5 (158-276.75)	194 (114-277)	215(137-309)	213(140-298)	<.001
Fibrinogen (mg/dL)	203(162-341)	430(267.5-430.5)	288.5 (234.5-412)	189.5 (125.5-437)	235(155.25-426.25)	.5
INR[e]	1.2 (1.1-1.5)	1.2 (1.1-1.4)	1.35 (1.2-1.7)	1.3 (1.2-1.6)	1.3 (1.1-1.6)	<.001
Outcomes, n (%)						
Mortality	186 (4.78)	26 (1.66)	179 (9.36)	324 (7.80)	715 (6.21)	<.001
Hospice	20 (0.51)	7 (0.45)	19 (0.99)	39 (0.94)	85 (0.74)	.04
Septic shock	253 (6.50)	37 (2.37)	458 (23.95)	505 (12.16)	1253 (10.88)	<.001
Acute kidney injury	1103 (28.34)	393 (25.14)	708 (37.03)	1275 (30.71)	3479 (30.20)	<.001
Mechanical ventilation	310 (7.97)	65 (4.16)	225 (11.77)	393 (9.47)	993 (8.62)	<.001

[a]*P* values for continuous variables are based on Kruskal-Wallis rank-sum tests. *P* values for categorical variables are based on Pearson χ^2 tests.
[b]N/A: not applicable.
[c]SOFA: Sequential Organ Failure Assessment.
[d]Only intravenous corticosteroids administered during the initial 6-hour time frame were included.
[e]INR: International Normalized Ratio.

Phenotype Dynamics

The examination of patient phenotypes 3 hours after ED arrival and 6 hours after ED arrival revealed that 44.9% (5170/11,519) of patients transitioned phenotypes between these time points. Further, Markov modeling of these transitions showed that certain transitions were more favored than others (Figure 4.1). Transitions between phenotypes with similar rates of adverse outcomes were more likely to occur than transitions between phenotypes with larger differences. For example, the transition from the phenotype with the lowest rate of mortality, septic shock, AKI, and mechanical ventilation (P2) to the phenotype with the highest rates for these outcomes (P3) was the least likely transition, only occurring in 1.2% (11/894) of patients.

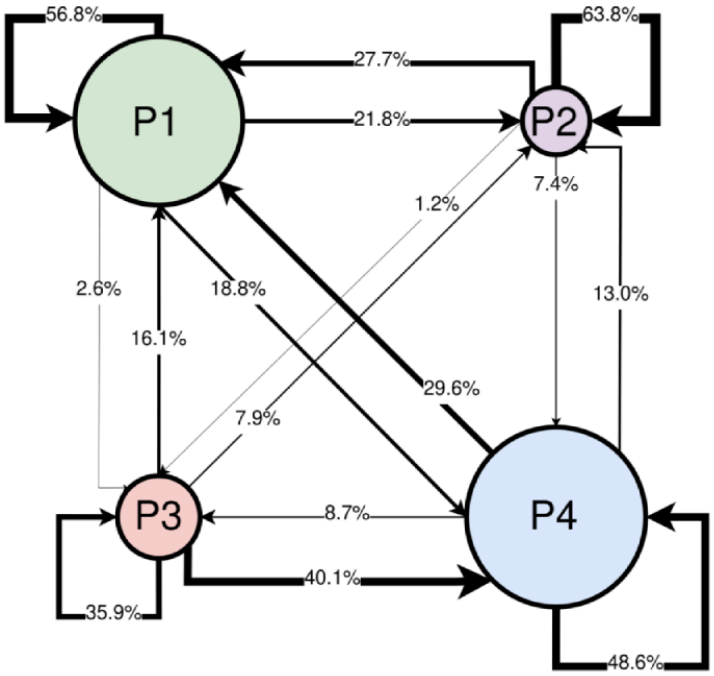


Figure 4.1: Markov phenotype transitions.
The figure shows transitions between phenotypes. P1, P2, P3, and P4 represent the initial phenotype assigned at 3 hours after emergency department arrival. Transitions are noted by arrows proportional to the transition rate. Patients who remained in the same phenotype have a line that loops around to the same initial phenotype. Mortality rates per phenotype: P1 (4.8%), P2 (1.7%), P3 (9.4%), and P4 (7.8%).

Table 4.3 and Tables S4.6-S4.9 demonstrate the impact of interventions and physiological factors on predicting phenotype transitions. The results from 4 different logistic regression models (M1-M4) predicting the transition to the 4 possible states are presented. We observed that antibiotics administered within 3 hours are significantly correlated with a higher rate of transition to the lowest mortality phenotype (P<.001) and lower rate of transition to the second highest mortality phenotype (P<.001). We also observed that the fluid volume greater than 30 mL/kg of body weight is significantly correlated with a higher rate of transition to the highest mortality phenotype (P=.005). We did not observe these trends when evaluating transitions among SOFA groups.

Table 4.4 and Tables S4.10-S4.13 present the results from this analysis on the external validation data set. We observe, as in the derivation cohort, that the timely administration of antibiotics is significantly associated with transition to a low-mortality cluster (P=.004). We do not, however, observe any significant correlations between fluid administration greater than 30 mL/kg and phenotype transitions.

Table 4.3: Odds ratios [95% CI] (*P* values) of interventions and physiological factors for predicting phenotype transition in the UCSD[a] cohort**.**

	M1[b]	**M2**	**M3**	**M4**
Cluster hospice + mortality (%)	6.2	3.2	13.8	8.3
Average SOFA[c] change	+0.2	+0.1	+0.6	+0.4
Antibiotics administered within 3 hours of ED[d] triage (true/false), odds ratio [95% CI]	1.024 [0.94-1.11] (*P*=.637)	1.438 [1.3-1.59](*P*<.001)[e]	0.955 [0.82-1.11] (*P*=.617)	0.75 [0.69-0.82](*P*<.001)[e]
Fluids administered within 3 hours of ED triage (true/false), odds ratio (95% CI)	0.876 [0.78-0.98] (*P*=.057)	1.106 [0.97-1.27](*P*=.22)	0.801 [0.62-1.03] (*P*=.149)	1.09 [0.95-1.25](*P*=.290)
Volume of fluids within 3 hours of ED triage ≥30 mL/kg (true/false), odds ratio (95% CI)	1.067 [0.98-1.16](*P*=.207)	0.938 [0.84-1.04] (*P*=.32)	1.227 [1.09-1.38] (*P*=.005)[e]	0.908 [0.84-0.99](*P*=.055)
Phenotype 2, 3 hours after ED triage (true/false),[f] odds ratio (95% CI)	0.288 [0.26-0.32](*P*<.001)	6.841 [6.11-7.65](*P*<.001)	0.444 [0.29-0.68](*P*=.001)	0.329 [0.28-0.39](*P*<.001)
Phenotype 3, 3 hours after ED triage (true/false),[f] odds ratio (95% CI)	0.144 [0.13-0.16](*P*<.001)	0.275 [0.23-0.32](*P*<.001)	21.116 [17.38-25.65] (*P*<.001)	3.264 [2.93-3.64](*P*<.001)
Phenotype 4, 3 hours after ED triage (true/false),[f] odds ratio (95% CI)	0.319 [0.29-0.35](*P*<.001)	0.501 [0.45-0.56](*P*<.001)	3.646 [3-4.43](*P*<.001)	4.374 [4.01-4.78](*P*<.001)

[a]UCSD: University of California, San Diego.
[b]M1 to M4 correspond to the 4 logistic regression models predicting the 4 possible outcomes (transition to phenotype 1, transition to phenotype 2, etc) based on the current state and interventions.
[c]SOFA: sequential organ failure assessment.
[d]ED: emergency department.
[e]*P* values <.05 represent significance.
[f]The phenotype of the patient 3 hours after ED triage.

Table 4.4: Odds ratios [95% CI] (*P* values) of interventions and physiological factors for predicting phenotype transition in the validation cohort.

	M1[a]	**M2**	**M3**	**M4**
Cluster hospice + mortality (%)	6.2	3.2	13.8	8.3
Average SOFA[b] change	+0.2	+0.1	+0.6	+0.4
Antibiotics administered within 3 hours of ED[c] triage (true/false), odds ratio [95% CI]	1.453 [1.17-1.8](*P*=.004)[d]	0.932 [0.59-1.48](*P*=.80)	0.99 [0.75-1.31](*P*=.95)	0.757 [0.63-0.91] (*P*=.015)
Fluids administered within 3 hours of ED triage (true/false), odds ratio (95% CI)	1.017 [0.82-1.26] (*P*=.895)	1.027 [0.64-1.65](*P*=.93)	1.213 [0.92-1.59](*P*=.24)	0.896 [0.74-1.08] (*P*=.337)
Volume of fluids within 3 hours of ED triage ≥30 mL/kg (true/false), odds ratio (95% CI)	0.737 [0.46-1.17] (*P*=.278)	1.461 [0.34-6.32](*P*=.67)	0.854 [0.55-1.33](*P*=.56)	1.279 [0.9-1.81](*P*=.242)
Phenotype 2, 3 hours after ED triage (true/false),[e] odds ratio (95% CI)	0.162 [0.11-0.24] (*P*<.001)	39.817 [25.03-63.34] (*P*<.001)	0.924 [0.36-2.37](*P*=.89)	0.286 [0.18-0.45] (*P*<.001)
Phenotype 3, 3 hours after ED triage (true/false),[e] odds ratio (95% CI)	0.025 [0.01-0.04](*P*<.00 1)	0.07 [0.01-0.38](*P*=.01)	32.448 [20.06-52.49](*P*<. 001)	1.984 [1.53-2.57] (*P*<.001)
Phenotype 4, 3 hours after ED triage (true/false),[e] odds ratio (95% CI)	0.131 [0.1-0.16](*P*<.001)	0 [0-Inf](*P*=.98)	4.169 [2.6-6.69] (*P*<.001)	5.539 [4.49-6.84](*P*<.00 1)

[a]M1 to M4 correspond to the 4 logistic regression models predicting the 4 possible outcomes (transition to phenotype 1, transition to phenotype 2, etc) based on the current state and interventions.
[b]SOFA: sequential organ failure assessment.
[c]ED: emergency department.
[d]*P* values <.05 represent significance.
[e]The phenotype of the patient 3 hours after ED triage.

Tables S4.6-S4.13 show the impact of interventions and physiological factors on the 16 potential phenotype transitions from one state to any other state. In the derivation and validation cohorts, we observed that patients in low-mortality phenotypes are significantly more likely to remain in that phenotype if they receive antibiotics within 3 hours of arrival. In the derivation

cohort, we observe that patients in a low-mortality phenotype are more likely to transition to a high-mortality phenotype if they receive fluids within 3 hours of arrival.

DISCUSSION

The major findings of our paper are that using a large cohort of patients with sepsis, distinct, clinical phenotypes exist and that these phenotypes may be affected by recommended therapies. The phenotypes we identified were done with data readily available to emergency physicians at the time of treatment and were distinct from traditional sepsis classifications based on comorbidities, illness severity, and laboratory results. Such results may have an impact on early sepsis management, assist with triage of sepsis patients, and potentially improve patient-centered outcomes. Importantly, we show that these phenotypes are dynamic—that is, a significant number of patients change phenotypes after their initial cluster assignment. Our results indicate that the therapies initiated by physicians may impact the trajectory of these patients and that the administration of 30 mL/kg may be associated with worse outcomes in certain patient phenotypes.

A central finding of our work is that a large number of patients may experience a transition in their initially assigned phenotype based on therapies provided in the ED. Prior work on identifying sepsis phenotypes or subtypes has been limited to static representations of this disease process, where patients are given an initial classification without continued evaluation to determine if these phenotypes change with time or intervention. Our data suggest that these phenotypes are dynamic, namely, that patients transition between phenotypes. Specifically, we found that 48.3% of patients underwent a transition in sepsis phenotype after initial classification when we reassessed for phenotype membership. We identified that certain therapeutic

interventions were associated with phenotype transition. Whether this represents the natural trajectory of patients with community-onset sepsis or host response to therapy is uncertain. However, such information may assist providers to triage patients as we show that patients rarely transition from low-mortality to high-mortality phenotypes (and vice versa) in the initial 6 hours from ED triage. It is plausible that the initiation of early sepsis therapies does indeed impact phenotypes as our data show that patients who receive prompt antimicrobials are more likely to stay in or transition to low-risk phenotypes. However, we lack biologic data (eg, gene expression and other host-response data) to establish that this change does not represent natural disease progression.

Whether a "one size fits all" approach as recommended by the current SEP-1 national quality measure is ideal remains controversial. Certain professional societies, including the American College of Emergency Medicine Physicians, have expressed concerns about the current SEP-1 bundle, specifically with regard to quantity of fluid administration and timing of antibiotics in patients without septic shock.[29,30] The current Surviving Sepsis Campaign recommends a 30 mL/kg fluid bolus for patients with sepsis-induced hypotension.[6] Prior work has described 4 distinct clinical phenotypes of patients with sepsis in the ED with distinct immune responses and clinical outcomes and showed that certain patients with sepsis may have worse outcomes with fluid resuscitation and physicians may opt for lower quantities of fluid resuscitation for patients with sepsis with certain comorbidities, such as heart failure.[31,119] Our results provide some confirmation of this, and it is possible that certain patients with sepsis may not benefit from the recommended 30 mL/kg fluid bolus currently recommended by Centers for Medicare and Medicaid Services or the Surviving Sepsis Campaign. Our data show that prompt

antimicrobial therapy is associated with significant decreases in mortality and may prevent transition to higher risk phenotypes.

We recognize that the phenotypes from the external validation site did not demonstrate a phenotype with worse outcomes due to a 30 mL/kg fluid bolus. Although speculative, we hypothesize that the observed difference in the association of high-volume fluid resuscitation with state transition is due to differences in clinical practice between the sites as demonstrated by the fluid intervention differences in Table 4.1. Specifically, 26% of patients in the derivation cohort versus 7.8% of patients in the validation cohort received fluid volumes greater than 30 mL/kg. Alternatively, this may represent the heterogeneity of sepsis syndromes. Prospective randomized trials are needed to delineate potential patient harm (or benefit) from the quantity of fluid. Indeed, we noted a strong association between transition to the highest mortality group (P3) in patients who received this quantity of fluid within their first 6 hours in the ED in the development cohort. Patients in the phenotype who had worse outcomes (P3) shared a similar clinical profile as patients who were not harmed by fluid resuscitation. This finding may assist providers in determining which patients may benefit from additional testing (ie, dynamic tests such as the passive leg raise) to assess for fluid responsiveness rather than empiric recommendations. A recent small randomized trial showed that personalized, dynamic fluid responsiveness monitoring resulted in a significant decrease in fluid administration, with lower rates of renal replacement therapy and mechanical ventilation compared with a protocol-driven group.[120]

Limitations

We acknowledge that there are limitations to our work. Because of the retrospective nature of this study, findings are associations and we cannot show causality. Data were abstracted at predefined time points, and changes in these cutoffs may impact the clinical phenotypes provided. We excluded patients who met sepsis criteria >6 hours after their initial ED triage. Inclusion of these patients may have impacted phenotype characteristics and response to therapy. However, these patients comprised a small minority of subjects with sepsis in the ED, and, thus, we felt that their exclusion would not significantly impact our findings.[121] We did not have access to host biomarkers and transcriptomic data to confirm that these clinical phenotypes have biologic validity. Certain input variables had a high degree of missingness, and the use of the population mean for imputation of missing values may have introduced bias by overestimating the severity of illness for patients with missing values. This effect is partially counteracted by the use of representations from a neural network that considers whether a value was imputed.[20]

CONCLUSION

Using a development and external validation set of 11,519 and 2091 patients with sepsis, respectively, we derive and describe 4 sepsis clinical phenotypes present within 6 hours of triage in the ED. We show that 48% of patients with sepsis change phenotype membership within the first 6 hours of ED triage, which is associated with therapies provided in the ED. We observe that the administration of a 30 mL/kg fluid bolus is associated with worse outcomes in certain phenotypes, whereas prompt antimicrobial therapy is associated with improved outcomes. Additional prospective studies are needed to validate these findings, including use of genomic

and other biologic data from sepsis patients to better quantify underlying pathophysiologic changes present.

ACKNOWLEDGEMENTS

Chapter 4, in full, is a reprint of the material as it appears in “Representation Learning and Spectral Clustering for the Development and External Validation of Dynamic Sepsis Phenotypes: Observational Cohort Study”. Boussina, A., Wardi, G., Shashikumar, S. P., Malhotra, A., Zheng, K., & Nemati, S. (2023) Journal of Medical Internet Research, 25, e45614. The dissertation author was the primary investigator and author of this paper.

SUPPLEMENTARY MATERIAL

Supplementary Note 4.1. Supplementary Methods.

Study Design and Setting:

UCSD consists of 2 EDs with a total annual census of 70,000 patients; one ED receives patients at a quaternary care center whereas the other functions in a "safety net" hospital. UC Irvine has a single ED with a total annual census of 50,000 patients. We followed recommendations provided by STROBE guidelines to ensure appropriate reporting of this research. University of California San Diego Institutional review board (IRB) approval was obtained with waiver of informed consent (#800257). Data abstracted from University of California, Irvine was provided under Data Use Agreement #37533.

Statistical Methods:

In our modeling of phenotype dynamics as a markov decision process, only interventions within the first 3 hours from ED triage were considered so that the effect of early interventions on a patient's trajectory could be examined. To ensure a clear separation between the variables related to transition and those related to the severity of illness, we excluded age, CCI, and SOFA from our transition models. However, we included the initial phenotype of the patient, which captures a set of features that includes age, comorbidities, and SOFA elements (among others), as a confounder to adjust for patient state. To evaluate the dynamics of our derived phenotypes against traditional patient categorizations, we also repeated this analysis using SOFA groups instead of phenotypes. Significant predictors of transition were determined from Wald tests on the odds ratios. Adjustment for multiple comparisons via the Bonferroni method was done by setting a conservative significance threshold of $p < 0.01$.

Supplementary Note 4.2: Outputs from Consensus Clustering on Representations.

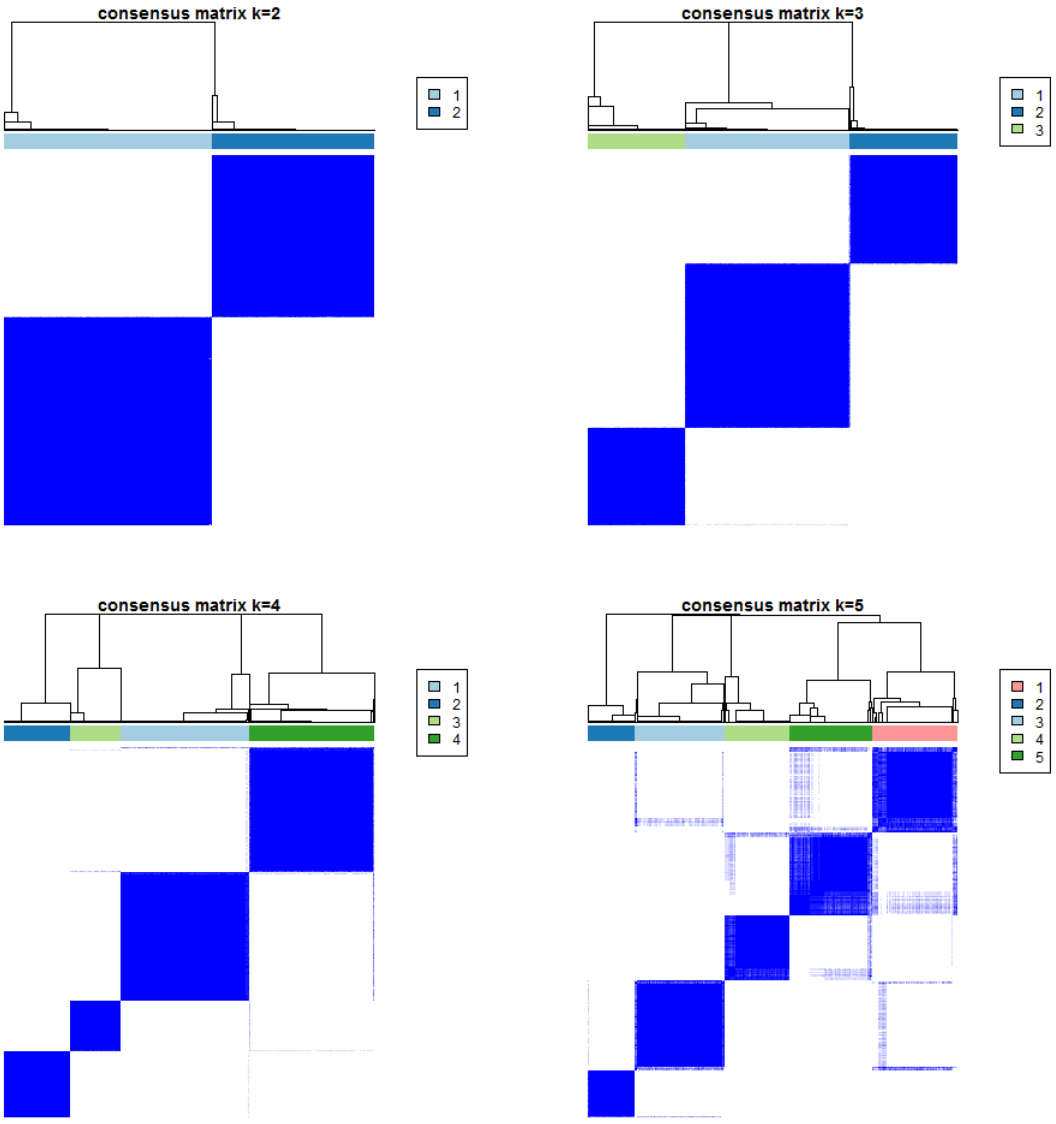


Figure S4.1: Consensus matrices at k=2-5 for consensus spectral clustering on the representations of a feed forward neural network trained to predict the onset of sepsis.

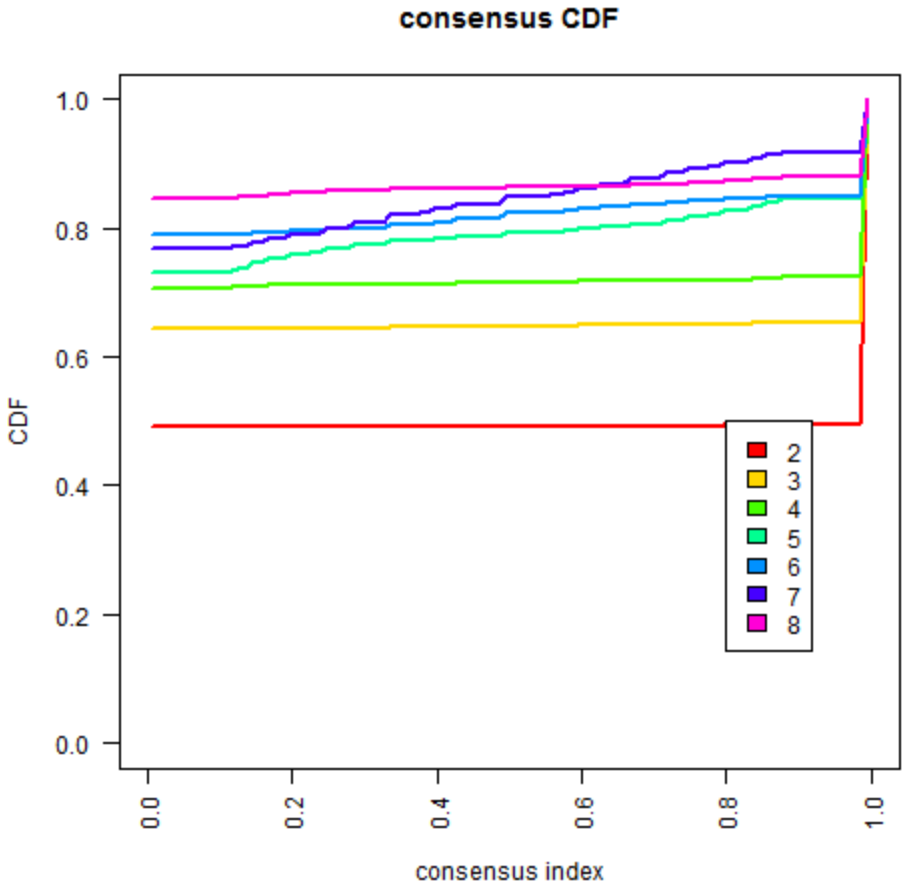


Figure S4.2: Cumulative density function (CDF) for consensus spectral clustering on the representations of a feed forward neural network trained to predict the onset of sepsis.
To determine the optimal number of clusters (N), we utilized the CDF and increased N until the consensus index was no longer primarily 0 or 1 (i.e. the CDF curve was no longer flat). The CDF shows an optimal value of k=4.

Supplementary Note 4.3: Outputs from Consensus Clustering on Raw Data.

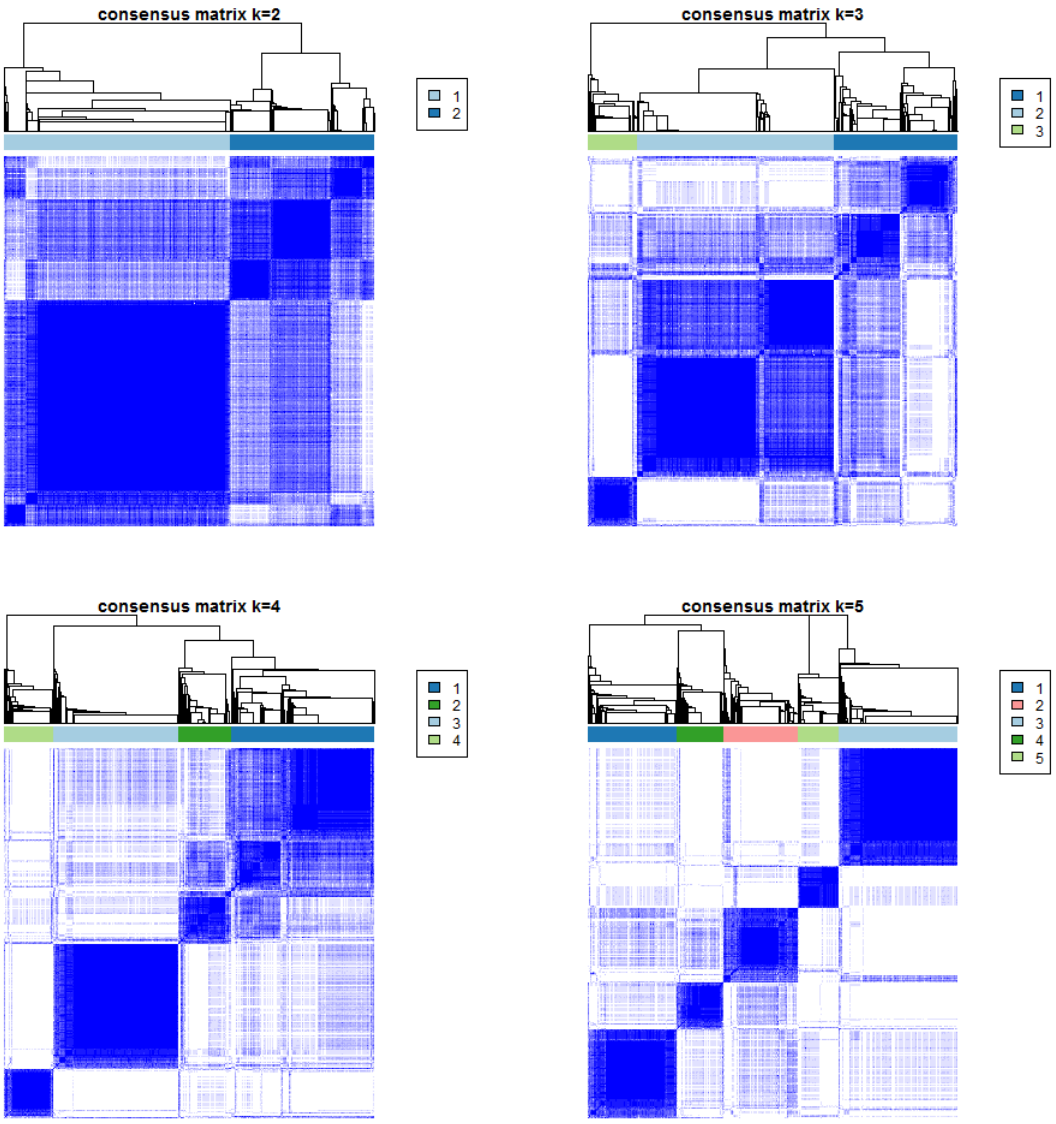


Figure S4.3: Consensus matrices at k=2-5 for consensus k-means clustering on normalized EHR data.
The presence of cluster membership overlap between consecutive runs of k-means demonstrates that this approach results in less consistent clustering than spectral clustering on representations (Figures S1-S2).

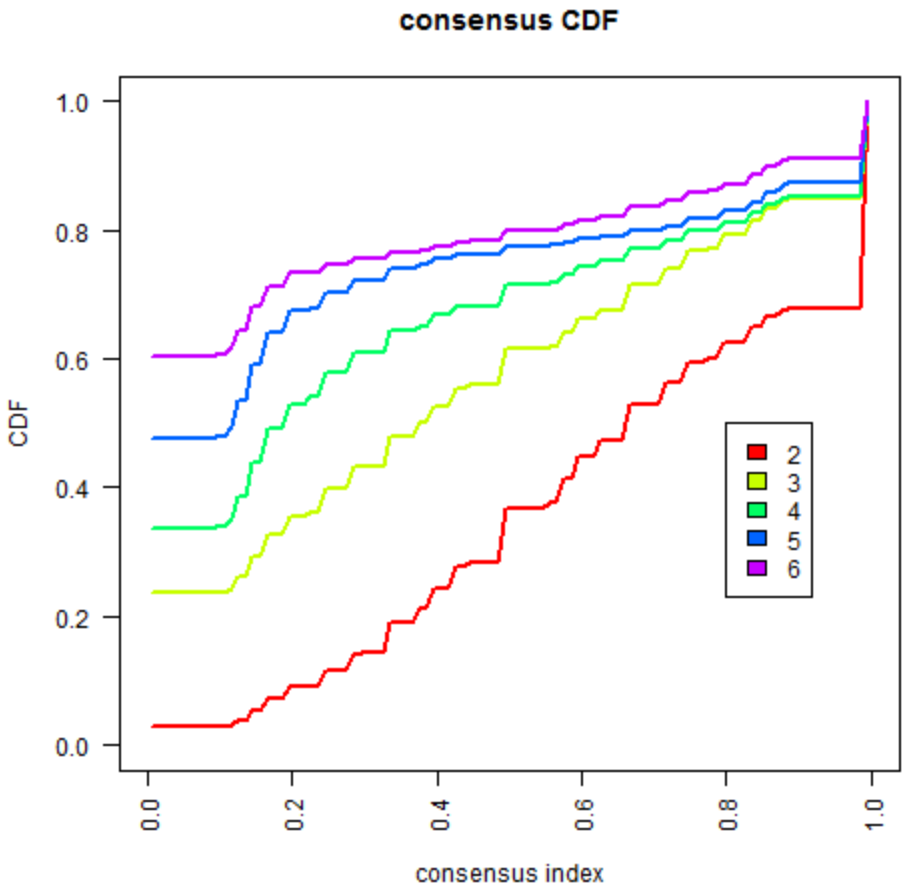


Figure S4.4: Cumulative density function for consensus k-means clustering on normalized EHR data.
The absence of a flattened curve demonstrates inconsistent clustering results.

Supplementary Note 4.4. Phenotype Characteristics.

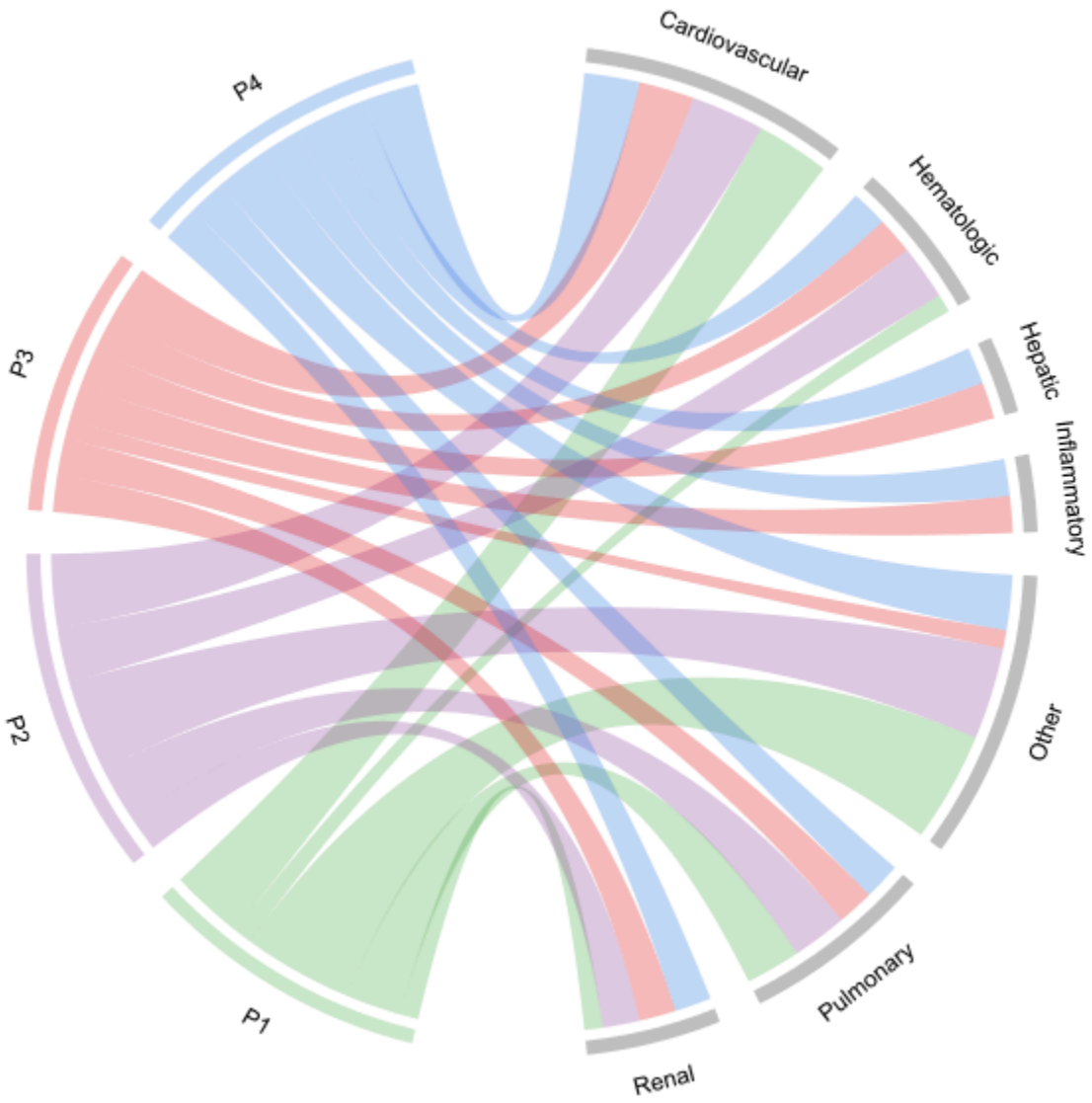


Figure S4.5: Phenotype Chord Diagram.
The chord diagram shows the clinical profile of patients in the 4 phenotypes we identified. Size of each individual ribbon represents the portion of abnormal features within that body system. For example, phenotypes 3 and 4 are more likely to have abnormal hepatic and inflammatory biomarkers compared to phenotypes 1 and 2.

Table S4.1: List of Clinical Variables

Clinical Variable
Cardiovascular
Heart Rate
Systolic Blood Pressure
Mean Arterial Pressure
Diastolic Blood Pressure
Bicarbonate
Lactic acid
Troponin I
Hematologic
Hematocrit
Hemoglobin
Partial Thromboplastin Time
Fibrinogen
Platelets
Hepatic
Aspartate Transaminase
Alkaline Phosphatase
Bilirubin Direct
Total Bilirubin
Inflammatory
Temperature
Leukocyte Count
Pulmonary
Pulse Oximetry
Respiration rate
End tidal carbon dioxide
Fraction of inspired oxygen
Partial pressure of carbon dioxide from arterial blood
Oxygen saturation from arterial blood
Renal
Measure of Excess Bicarbonate
Creatinine
Blood urea nitrogen
pH
Other
Potassium
Calcium
Chloride
Glucose
Magnesium
Phosphate
Age
Gender

Table S4.2: Characteristics and Therapeutics of the 4 Sepsis Phenotypes at the Validation Site

	P1	P2	P3	P4	Total	P Value[a]
Characteristic						
Number of Patients, N (%)	619 (29.6%)	168 (8%)	382 (18.3%)	922 (44.1%)	2091 (100%)	-
Age, Mean (SD)	61 (17.7)	61 (18.6)	62 (17.2)	60 (17.6)	61 (17.6)	0.51
Sex, N (%)						0.39
Male	370 (59.8%)	91 (54.2%)	214 (56%)	537 (58.2%)	1212 (58%)	
Female	249 (40.2%)	77 (45.8%)	168 (44%)	385 (41.8%)	879 (42%)	
Organ Dysfunction						
Charlson Comorbidity Index, Median (IQR)	2 (1-4)	2 (1-4)	3 (2-5)	3 (2-5)	2 (2-5)	<0.001
CCI, Congestive Heart Failure Component, N (%)	87 (14.1%)	17 (10.1%)	35 (9.2%)	86 (9.3%)	225 (10.8%)	0.018
CCI, Moderate/Severe Liver Disease Component, N (%)	19 (3.1%)	1 (0.6%)	13 (3.4%)	45 (4.9%)	78 (3.7%)	0.03
CCI, Renal Disease Component, N (%)	106 (17.1%)	37 (22%)	82 (21.5%)	172 (18.7%)	397 (19%)	0.26
SOFA Score (Max Within 3 Hours), Median (IQR)	2 (1-3)	1 (0-2)	3 (2-5)	3 (1-4)	2 (1-4)	<0.001
SOFA Score (Max Within 6 Hours), Median (IQR)	2 (1-4)	2 (1-2)	4 (2-5)	3 (1-4)	3 (1-4)	<0.001
Antibiotics						
Antibiotics Within 3 Hours, N (%)	307 (49.6%)	55 (32.7%)	302 (79.1%)	564 (61.2%)	1228 (58.7%)	<0.001
Antibiotics Within 6 Hours, N (%)	517 (83.5%)	124 (73.8%)	360 (94.2%)	804 (87.2%)	1805 (86.3%)	<0.001
Time to Antibiotics (Hrs), Median (IQR)	3.02 (1.617 - 5.008)	4.15 (2.333 - 6.217)	1.62 (0.883 - 2.829)	2.35 (1.283 - 4.138)	2.52 (1.283 - 4.383)	<0.001

Table S4.2: Characteristics and Therapeutics of the 4 Sepsis Phenotypes at the Validation Site (Continued)

	P1	P2	P3	P4	Total	P Value[a]
Fluids						
Fluids, N (%)	441 (71.2%)	99 (58.9%)	308 (80.6%)	698 (75.7%)	1546 (73.9%)	<0.001
Time to Fluids (Hrs), Median (IQR)	1.58 (0.583 - 5.567)	2.7 (0.625 - 6.192)	0.93 (0.417 - 3.987)	1.2 (0.433 - 5.975)	1.29 (0.45 - 5.658)	0.04
Fluid Intake (mL), First Three Hours, Median (IQR)	500 (0 - 1000)	0 (0 - 1000)	1000 (0 - 1610)	500 (0 - 1020)	500 (0 - 1000)	<0.001
Vasopressors						
Pressors, N (%)	32 (5.2%)	4 (2.4%)	89 (23.3%)	105 (11.4%)	230 (11%)	<0.001
Laboratory Values						
Subjects with Lactate Measurements, N (%)	435 (70.3%)	132 (78.6%)	276 (72.3%)	629 (68.2%)	1472 (70.4%)	0.04
Lactate (mmol/L), Median (IQR)	2.1 (1.5 - 3.2)	2.1 (1.4 - 3.425)	2.2 (1.5 - 3.1)	2.2 (1.5 - 3.1)	2.2 (1.5 - 3.1)	0.98
Outcomes						
Mortality, N (%)	18 (2.9%)	2 (1.2%)	29 (7.6%)	59 (6.4%)	108 (5.2%)	<0.001
Hospice, N (%)	40 (6.5%)	8 (4.8%)	45 (11.8%)	79 (8.6%)	172 (8.2%)	0.008
Septic Shock, N (%)	32 (5.2%)	4 (2.4%)	89 (23.3%)	105 (11.4%)	230 (11%)	<0.001
Acute Kidney Injury, N (%)	241 (38.9%)	63 (37.5%)	159 (41.6%)	356 (38.6%)	819 (39.2%)	0.008
Mechanical Ventilation, N (%)	41 (6.6%)	11 (6.5%)	43 (11.3%)	86 (9.3%)	181 (8.7%)	<0.001

[a]P values for continuous variables are based on Kruskal-Wallis rank sum tests. P values for categorical variables are based on Pearson's χ^2 tests.

Table S4.3: Characteristics and Therapeutics of the 4 Sepsis Phenotypes at the Development Site (Hour 6)

	P1	P2	P3	P4	NA	Total
Characteristic						
Number of Patients, N (%)	4122 (35.8%)	2501 (21.7%)	1154 (10%)	3587 (31.1%)	155 (1.3%)	11519 (100%)
Age, Mean (SD)	61 (17.3)	61 (17.8)	60 (17.8)	60 (17.8)	56 (20.4)	61 (17.7)
Sex, N (%)						
Male	2294 (55.7%)	1389 (55.5%)	688 (59.6%)	2063 (57.5%)	75 (48.4%)	6509 (56.5%)
Female	1828 (44.3%)	1112 (44.5%)	466 (40.4%)	1524 (42.5%)	80 (51.6%)	5010 (43.5%)
Organ Dysfunction						
Charlson Comorbidity Index, Median (IQR)	2 (1-4)	2 (1-4)	2 (1-5)	2 (1-5)	2 (1-4)	2 (1-4)
CCI, Congestive Heart Failure Component, N (%)	530 (12.9%)	313 (12.5%)	116 (10.1%)	402 (11.2%)	14 (9%)	1375 (11.9%)
CCI, Moderate/Severe Liver Disease Component, N (%)	149 (3.6%)	87 (3.5%)	91 (7.9%)	190 (5.3%)	3 (1.9%)	520 (4.5%)
CCI, Renal Disease Component, N (%)	712 (17.3%)	368 (14.7%)	242 (21%)	614 (17.1%)	23 (14.8%)	1959 (17%)
SOFA Score (Max Within 3 Hours), Median (IQR)	2 (1-3)	1 (0-3)	3 (1-5)	2 (1-4)	1 (0-3)	2 (1-3)
SOFA Score (Max Within 6 Hours), Median (IQR)	2 (1-3)	2 (1-3)	3 (2-5)	2 (1-4)	1 (0-3)	2 (1-4)
Antibiotics						
Antibiotics Within 3 Hours, N (%)	2426 (58.9%)	1435 (57.4%)	893 (77.4%)	2370 (66.1%)	126 (81.3%)	7250 (62.9%)
Antibiotics Within 6 Hours, N (%)	3548 (86.1%)	2070 (82.8%)	1094 (94.8%)	3230 (90%)	152 (98.1%)	10094 (87.6%)
Time to Antibiotics (Hrs), Median (IQR)	2.4 (1.233 - 4.383)	2.52 (1.25 - 4.65)	1.52 (0.833 - 2.783)	2.08 (1.067 - 3.767)	1.57 (0.95 - 2.592)	2.2 (1.117 - 4.05)
Fluids						
Fluids, N (%)	4120 (100%)	2498 (99.9%)	1154 (100%)	3587 (100%)	153 (98.7%)	11512 (99.9%)
Time to Fluids (Hrs), Median (IQR)	0.93 (0.467 - 2.004)	0.97 (0.367 - 2.175)	0.7 (0.4 - 1.246)	0.83 (0.45 - 1.617)	0.65 (-0.033 - 1.267)	0.87 (0.433 - 1.85)
Fluid Intake (mL), First Three Hours, Median (IQR)	1000 (177.01 - 2031.09)	858.2 (60 - 1776.84)	1956.7 (871.63 - 2859.19)	1270.6 (476.71 - 2321.93)	1050 (300 - 2034.17)	1084.1 (266.56 - 2199.53)
Vasopressors						
Pressors, N (%)	358 (8.7%)	109 (4.4%)	317 (27.5%)	459 (12.8%)	10 (6.5%)	1253 (10.9%)
Laboratory Values						
Subjects with Lactate Measurements, N (%)	3041 (73.8%)	1610 (64.4%)	1013 (87.8%)	2942 (82%)	123 (79.4%)	8729 (75.8%)
Lactate (mmol/L), Median (IQR)	2 (1.4 - 2.8)	1.9 (1.4 - 2.5)	2.8 (2 - 4.1)	2.3 (1.6 - 3.3)	2 (1.35 - 2.75)	2.2 (1.5 - 3.1)
Outcomes						
Mortality, N (%)	229 (5.6%)	67 (2.7%)	140 (12.1%)	269 (7.5%)	10 (6.5%)	715 (6.2%)
Hospice, N (%)	25 (0.6%)	14 (0.6%)	19 (1.6%)	27 (0.8%)	0 (0%)	85 (0.7%)
Septic Shock, N (%)	358 (8.7%)	109 (4.4%)	317 (27.5%)	459 (12.8%)	10 (6.5%)	1253 (10.9%)
Acute Kidney Injury, N (%)	1202 (29.2%)	665 (26.6%)	447 (38.7%)	1137 (31.7%)	28 (18.1%)	3479 (30.2%)
Mechanical Ventilation, N (%)	362 (8.8%)	122 (4.9%)	148 (12.8%)	352 (9.8%)	9 (5.8%)	993 (8.6%)

Table S4.4: Characteristics and Therapeutics of the 4 Sepsis Phenotypes at the Validation Site (Hour 6)

	P1	P2	P3	P4	NA	Total
Characteristic						
Number of Patients, N (%)	477 (22.8%)	121 (5.8%)	247 (11.8%)	830 (39.7%)	416 (19.9%)	2091 (100%)
Age, Mean (SD)	61 (17.2)	58 (19)	60 (16.5)	61 (17.7)	63 (17.9)	61 (17.6)
Sex, N (%)						
Male	286 (60%)	71 (58.7%)	138 (55.9%)	483 (58.2%)	234 (56.2%)	1212 (58%)
Female	191 (40%)	50 (41.3%)	109 (44.1%)	347 (41.8%)	182 (43.8%)	879 (42%)
Organ Dysfunction						
Charlson Comorbidity Index, Median (IQR)	2 (1-4)	2 (1-3)	3 (2-6)	2 (2-5)	2 (1-5)	2 (2-5)
CCI, Congestive Heart Failure Component, N (%)	67 (14%)	13 (10.7%)	20 (8.1%)	81 (9.8%)	44 (10.6%)	225 (10.8%)
CCI, Moderate/Severe Liver Disease Component, N (%)	15 (3.1%)	0 (0%)	11 (4.5%)	35 (4.2%)	17 (4.1%)	78 (3.7%)
CCI, Renal Disease Component, N (%)	90 (18.9%)	25 (20.7%)	58 (23.5%)	157 (18.9%)	67 (16.1%)	397 (19%)
SOFA Score (Max Within 3 Hours), Median (IQR)	2 (1-3)	1 (1-2)	3 (1-5)	2 (1-4)	3 (1-5)	2 (1-4)
SOFA Score (Max Within 6 Hours), Median (IQR)	2 (1-4)	2 (1-2)	4 (2-6)	3 (2-4)	2 (1-4)	3 (1-4)
Antibiotics						
Antibiotics Within 3 Hours, N (%)	272 (57%)	45 (37.2%)	174 (70.4%)	481 (58%)	256 (61.5%)	1228 (58.7%)
Antibiotics Within 6 Hours, N (%)	404 (84.7%)	92 (76%)	231 (93.5%)	716 (86.3%)	362 (87%)	1805 (86.3%)
Time to Antibiotics (Hrs), Median (IQR)	2.68 (1.35 - 4.65)	3.9 (2.15 - 5.717)	2.12 (1 - 3.225)	2.56 (1.317 - 4.446)	2.13 (1.2 - 4.15)	2.52 (1.283 - 4.383)
Fluids						
Fluids, N (%)	335 (70.2%)	68 (56.2%)	192 (77.7%)	643 (77.5%)	308 (74%)	1546 (73.9%)
Time to Fluids (Hrs), Median (IQR)	1.47 (0.533 - 5.6)	1.73 (0.533 - 6.458)	1 (0.446 - 4.838)	1.45 (0.517 - 6.717)	1.01 (0.262 - 3.688)	1.29 (0.45 - 5.658)
Fluid Intake (mL), First Three Hours, Median (IQR)	500 (0 - 1000)	0 (0 - 1000)	1000 (0 - 1710)	500 (0 - 1000)	500 (0 - 1000)	500 (0 - 1000)
Vasopressors						
Pressors, N (%)	19 (4%)	2 (1.7%)	71 (28.7%)	77 (9.3%)	61 (14.7%)	230 (11%)
Laboratory Values						
Subjects with Lactate Measurements, N (%)	343 (71.9%)	92 (76%)	167 (67.6%)	581 (70%)	289 (69.5%)	1472 (70.4%)
Lactate (mmol/L), Median (IQR)	2.1 (1.5 - 3.25)	2 (1.375 - 3.075)	2.4 (1.45 - 3.35)	2.2 (1.5 - 3.1)	2.1 (1.5 - 3)	2.2 (1.5 - 3.1)
Outcomes						
Mortality, N (%)	15 (3.1%)	1 (0.8%)	22 (8.9%)	47 (5.7%)	23 (5.5%)	108 (5.2%)
Hospice, N (%)	27 (5.7%)	4 (3.3%)	27 (10.9%)	82 (9.9%)	32 (7.7%)	172 (8.2%)
Septic Shock, N (%)	19 (4%)	2 (1.7%)	71 (28.7%)	77 (9.3%)	61 (14.7%)	230 (11%)
Acute Kidney Injury, N (%)	188 (39.4%)	45 (37.2%)	114 (46.2%)	327 (39.4%)	145 (34.9%)	819 (39.2%)
Mechanical Ventilation, N (%)	27 (5.7%)	4 (3.3%)	20 (8.1%)	65 (7.8%)	65 (15.6%)	181 (8.7%)

Table S4.5: Performance (AUC) of Assigning Cluster Membership at ED Triage + 3 Hours Using Missing Value Flags vs True Clinical Values

	Missing Value Flags	**Clinical Values**
M1	0.60	0.73
M2	0.80	0.91
M3	0.73	0.93
M4	0.62	0.68

Here we explore the degree to which missingness alone accounts for cluster membership. First, we train four logistic regression models to assign cluster membership at ED Triage + 3 hours based only on whether each measurement is missing or not. We then compare this against four separate models trained on the true clinical values. We report the AUCs of each of these models in the table above. As shown, severity of illness and the presence of data are indeed entangled. This is particularly apparent for M2 (the cluster 2 assignment model) which represents the healthiest phenotype. However, using information from the clinical variables significantly improves the performance of the assignment. Therefore, missingness does not fully explain cluster membership.

Supplementary Note 4.5. Phenotype Transitions.

Table S4.6: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCSD Cohort (Cluster 1)

	Antibiotics administered within 3 hours of admission (True/False)	**Fluids administered within 3 hours of admission (True/False)**	**Volume of fluids within 3 hours of admission >= 30 mL/kg (True/False)**
M1 **Cluster Mortality: 5.6%** **Cluster Hospice: 0.6%** **Cluster Hospice/Mortality: 6.2%** **Average SOFA Change: +0.2**			
	1.03 [0.91 - 1.17] (0.67)	0.82 [0.69 - 0.96] (0.04)	0.98 [0.85 - 1.14] (0.82)
M2 **Cluster Mortality: 2.7%** **Cluster Hospice: 0.6%** **Cluster Hospice/Mortality: 3.2%** **Average SOFA Change: +0.1**			
	1.35 [1.16 - 1.56] (0.0009)	1.34 [1.08 - 1.65] (0.0235)	0.82 [0.69 - 0.98] (0.0636)
M3 **Cluster Mortality: 12.1%** **Cluster Hospice: 1.6%** **Cluster Hospice/Mortality: 13.8%** **Average SOFA Change: +0.6**			
	0.35 [0.23 - 0.52] (2e-05)	0.82 [0.53 - 1.25] (0.4361)	2.19 [1.4 - 3.41] (0.0037)
M4 **Cluster Mortality: 7.5%** **Cluster Hospice: 0.8%** **Cluster Hospice/Mortality: 8.3%** **Average SOFA Change: +0.4**			
	0.79 [0.68 - 0.93] (0.014)	1.07 [0.87 - 1.31] (0.591)	1.15 [0.96 - 1.39] (0.207)

Table S4.7: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCSD Cohort (Cluster 2)

	Antibiotics administered within 3 hours of admission (True/False)	**Fluids administered within 3 hours of admission (True/False)**	**Volume of fluids within 3 hours of admission >= 30 mL/kg (True/False)**
M1 **Cluster Mortality: 5.6%** **Cluster Hospice: 0.6%** **Cluster Hospice/Mortality: 6.2%** **Average SOFA Change: +0.2**			
	0.78 [0.62 - 0.98] (0.074)	0.9 [0.72 - 1.13] (0.438)	1.45 [0.99 - 2.14] (0.112)
M2 **Cluster Mortality: 2.7%** **Cluster Hospice: 0.6%** **Cluster Hospice/Mortality: 3.2%** **Average SOFA Change: +0.1**			
	1.72 [1.39 - 2.12] (2.6e-05)	0.88 [0.71 - 1.09] (0.310)	0.58 [0.4 - 0.83] (0.013)
M3 **Cluster Mortality: 12.1%** **Cluster Hospice: 1.6%** **Cluster Hospice/Mortality: 13.8%** **Average SOFA Change: +0.6**			
	0.55 [0.21 - 1.4] (0.29)	1.41 [0.55 - 3.65] (0.55)	2.09 [0.58 - 7.56] (0.34)
M4 **Cluster Mortality: 7.5%** **Cluster Hospice: 0.8%** **Cluster Hospice/Mortality: 8.3%** **Average SOFA Change: +0.4**			
	0.38 [0.25 - 0.56] (6.2e-05)	1.9 [1.28 - 2.81] (0.0071)	1.69 [0.93 - 3.06] (0.1494)

Table S4.8: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCSD Cohort (Cluster 3)

	Antibiotics administered within 3 hours of admission (True/False)	**Fluids administered within 3 hours of admission (True/False)**	**Volume of fluids within 3 hours of admission >= 30 mL/kg (True/False)**
M1 **Cluster Mortality: 5.6%** **Cluster Hospice: 0.6%** **Cluster Hospice/Mortality: 6.2%** **Average SOFA Change: +0.2**			
	1.15 [0.82 - 1.62] (0.48777)	0.82 [0.39 - 1.74] (0.66028)	1 [0.81 - 1.24] (0.98896)
M2 **Cluster Mortality: 2.7%** **Cluster Hospice: 0.6%** **Cluster Hospice/Mortality: 3.2%** **Average SOFA Change: +0.1**			
	1.29 [0.79 - 2.09] (0.39)	1.28 [0.36 - 4.59] (0.75)	1.12 [0.84 - 1.49] (0.52)
M3 **Cluster Mortality: 12.1%** **Cluster Hospice: 1.6%** **Cluster Hospice/Mortality: 13.8%** **Average SOFA Change: +0.6**			
	1.1 [0.85 - 1.41] (0.557)	0.6 [0.34 - 1.06] (0.142)	1.19 [1.01 - 1.4] (0.078)
M4 **Cluster Mortality: 7.5%** **Cluster Hospice: 0.8%** **Cluster Hospice/Mortality: 8.3%** **Average SOFA Change: +0.4**			
	0.8 [0.63 - 1.01] (0.121)	1.75 [0.97 - 3.14] (0.118)	0.82 [0.7 - 0.96] (0.036)

Table S4.9. Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCSD Cohort (Cluster 4)

	Antibiotics administered within 3 hours of admission (True/False)	**Fluids administered within 3 hours of admission (True/False)**	**Volume of fluids within 3 hours of admission >= 30 mL/kg (True/False)**
M1 **Cluster Mortality: 5.6%** **Cluster Hospice: 0.6%** **Cluster Hospice/Mortality: 6.2%** **Average SOFA Change: +0.2**			
	1.09 [0.94 - 1.25] (0.33)	1.13 [0.88 - 1.47] (0.42)	1.09 [0.96 - 1.23] (0.26)
M2 **Cluster Mortality: 2.7%** **Cluster Hospice: 0.6%** **Cluster Hospice/Mortality: 3.2%** **Average SOFA Change: +0.1**			
	1.36 [1.11 - 1.66] (0.012)	1.43 [0.95 - 2.16] (0.156)	1.09 [0.93 - 1.29] (0.378)
M3 **Cluster Mortality: 12.1%** **Cluster Hospice: 1.6%** **Cluster Hospice/Mortality: 13.8%** **Average SOFA Change: +0.6**			
	1.22 [0.96 - 1.55] (0.17)	0.91 [0.6 - 1.39] (0.71)	1.11 [0.91 - 1.36] (0.38)
M4 **Cluster Mortality: 7.5%** **Cluster Hospice: 0.8%** **Cluster Hospice/Mortality: 8.3%** **Average SOFA Change: +0.4**			
	0.77 [0.68 - 0.88] (0.0008)	0.83 [0.66 - 1.04] (0.1733)	0.86 [0.77 - 0.97] (0.0314)

Table S4.10: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCI Validation Cohort (Cluster 1)

	Antibiotics administered within 3 hours of admission (True/False)	**Fluids administered within 3 hours of admission (True/False)**	**Volume of fluids within 3 hours of admission >= 30 mL/kg (True/False)**
M1 **Cluster Mortality: 3.1%** **Cluster Hospice: 5.7%** **Cluster Hospice/Mortality: 8.8%** **Average SOFA Change: +0.2**			
	1.639 [1.21 - 2.23] (0.0081)	1.075 [0.79 - 1.47] (0.7036)	1.256 [0.5 - 3.14] (0.6814)
M2 **Cluster Mortality: 0.8%** **Cluster Hospice: 3.3%** **Cluster Hospice/Mortality: 4.1%** **Average SOFA Change: +0.2**			
	0.64 [0.31 - 1.33] (0.32)	0.606 [0.28 - 1.31] (0.29)	2.108 [0.34 - 13.11] (0.50)
M3 **Cluster Mortality: 8.9%** **Cluster Hospice: 10.9%** **Cluster Hospice/Mortality: 19.8%** **Average SOFA Change: +0.7**			
	0.931 [0.39 - 2.22] (0.89)	0.859 [0.36 - 2.08] (0.78)	0 [0 - Inf] (0.99)
M4 **Cluster Mortality: 5.7%** **Cluster Hospice: 9.9%** **Cluster Hospice/Mortality: 15.5%** **Average SOFA Change: +0.3**			
	0.64 [0.46 - 0.88] (0.02275)	1.039 [0.75 - 1.44] (0.84822)	0.761 [0.29 - 2.03] (0.64654)

Note: NAs indicate a lack of relevant data.

Table S4.11: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCI Validation Cohort (Cluster 2)

	Antibiotics administered within 3 hours of admission (True/False)	**Fluids administered within 3 hours of admission (True/False)**	**Volume of fluids within 3 hours of admission >= 30 mL/kg (True/False)**
M1 **Cluster Mortality: 3.1%** **Cluster Hospice: 5.7%** **Cluster Hospice/Mortality: 8.8%** **Average SOFA Change: +0.2**			
	0.836 [0.39 - 1.8] (0.70)	0.565 [0.25 - 1.29] (0.26)	0 [0 - Inf] (0.99)
M2 **Cluster Mortality: 0.8%** **Cluster Hospice: 3.3%** **Cluster Hospice/Mortality: 4.1%** **Average SOFA Change: +0.2**			
	1.12 [0.6 - 2.1] (0.766)	1.596 [0.83 - 3.08] (0.241)	NA
M3 **Cluster Mortality: 8.9%** **Cluster Hospice: 10.9%** **Cluster Hospice/Mortality: 19.8%** **Average SOFA Change: +0.7**			
	0.596 [0.08 - 4.2] (0.66)	2.598 [0.48 - 14.15] (0.35)	0 [0 - Inf] (1.00)
M4 **Cluster Mortality: 5.7%** **Cluster Hospice: 9.9%** **Cluster Hospice/Mortality: 15.5%** **Average SOFA Change: +0.3**			
	1.15 [0.47 - 2.8] (0.80)	0.621 [0.23 - 1.68] (0.43)	0 [0 - Inf] (0.99)

Note: NAs indicate a lack of relevant data.

Table S4.12: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCI Validation Cohort (Cluster 3)

	Antibiotics administered within 3 hours of admission (True/False)	**Fluids administered within 3 hours of admission (True/False)**	**Volume of fluids within 3 hours of admission >= 30 mL/kg (True/False)**
M1 **Cluster Mortality: 3.1%** **Cluster Hospice: 5.7%** **Cluster Hospice/Mortality: 8.8%** **Average SOFA Change: +0.2**			
	NA	1.078 [0.37 - 3.14] (0.91)	1.161 [0.28 - 4.8] (0.86)
M2 **Cluster Mortality: 0.8%** **Cluster Hospice: 3.3%** **Cluster Hospice/Mortality: 4.1%** **Average SOFA Change: +0.2**			
	NA	0 [0 - Inf] (1)	0.898 [0 - Inf] (1)
M3 **Cluster Mortality: 8.9%** **Cluster Hospice: 10.9%** **Cluster Hospice/Mortality: 19.8%** **Average SOFA Change: +0.7**			
	1.034 [0.64 - 1.66] (0.91)	0.982 [0.65 - 1.49] (0.94)	1.015 [0.55 - 1.87] (0.97)
M4 **Cluster Mortality: 5.7%** **Cluster Hospice: 9.9%** **Cluster Hospice/Mortality: 15.5%** **Average SOFA Change: +0.3**			
	0.771 [0.48 - 1.24] (0.37)	1.039 [0.68 - 1.58] (0.88)	0.959 [0.52 - 1.78] (0.91)

Note: NAs indicate a lack of relevant data.

Table S4.13: Odds Ratios [95% confidence interval] (p-values) of Interventions and Physiological Factors for Predicting Phenotype Transition in the UCI Validation Cohort (Cluster 4)

	Antibiotics administered within 3 hours of admission (True/False)	**Fluids administered within 3 hours of admission (True/False)**	**Volume of fluids within 3 hours of admission >= 30 mL/kg (True/False)**
M1 **Cluster Mortality: 3.1%** **Cluster Hospice: 5.7%** **Cluster Hospice/Mortality: 8.8%** **Average SOFA Change: +0.2**			
	1.301 [0.93 - 1.82] (0.196)	1.077 [0.77 - 1.51] (0.715)	0.498 [0.26 - 0.97] (0.085)
M2 **Cluster Mortality: 0.8%** **Cluster Hospice: 3.3%** **Cluster Hospice/Mortality: 4.1%** **Average SOFA Change: +0.2**			
	1 [0 - Inf] (1)	1 [0 - Inf] (1)	1 [0 - Inf] (1)
M3 **Cluster Mortality: 8.9%** **Cluster Hospice: 10.9%** **Cluster Hospice/Mortality: 19.8%** **Average SOFA Change: +0.7**			
	0.986 [0.66 - 1.47] (0.954)	1.538 [1.02 - 2.31] (0.082)	0.753 [0.38 - 1.48] (0.492)
M4 **Cluster Mortality: 5.7%** **Cluster Hospice: 9.9%** **Cluster Hospice/Mortality: 15.5%** **Average SOFA Change: +0.3**			
	0.837 [0.63 - 1.11] (0.294)	0.769 [0.58 - 1.02] (0.127)	1.822 [1.09 - 3.03] (0.053)

Note: NAs indicate a lack of relevant data.

Table S4.14: Odds Ratios (p-values) of Interventions and Physiological Factors for Predicting SOFA Group Transitions in the UCSD Cohort.

	Antibiotics administered within 3 hours of ED arrival (True/False)	**Fluids administered within 3 hours of ED arrival (True/False)**	**Volume of fluids within 3 hours of ED arrival >= 30 mL/kg (True/False)**	**SOFA [3-6), 3 Hours After Admission**	**SOFA [6-10), 3 Hours After Admission**	**SOFA [10 - 18), 3 Hours After Admission**
M1 SOFA: [0 - 3)	1.042 (0.6242)	1.406 (0.0028)	0.929 (0.3798)	0.006 (<2e-16)	0.001 (<2e-16)	0 (0.9037)
M2 SOFA: [3 - 6)	0.982 (0.8003)	0.737 (0.0026)	1.088 (0.2416)	58.797 (< 2e-16)	1.651 (4.8e-06)	0.098 (0.0208)
M3 SOFA: [6 - 10)	0.942 (0.61)	1.096 (0.58)	0.955 (0.68)	15.028 (<2e-16)	589.352 (<2e-16)	81.364 (<2e-16)
M4 SOFA: [10 - 18)	1.178 (0.5673)	1.121 (0.7866)	0.866 (0.5826)	8.214 (0.0012)	159.128 (<2e-16)	4725.546 (<2e-16)

At a significance threshold of $p < 0.01$, early administration of fluids is the only intervention significantly associated with transition. Since the administration of vasopressors is one of the SOFA score criteria, this may be a result of the observed inverse correlation between the use of fluids and the administration of pressors.

Table S4.15: Odds Ratios (p-values) of Interventions and Physiological Factors for Predicting SOFA Group Transitions in the UCI Cohort.

	Antibiotics administered within 3 hours of ED arrival (True/False)	**Fluids administered within 3 hours of ED arrival (True/False)**	**Volume of fluids within 3 hours of ED arrival >= 30 mL/kg (True/False)**	**SOFA [3-6), 3 Hours After Admission**	**SOFA [6-10), 3 Hours After Admission**	**SOFA [10 - 18), 3 Hours After Admission**
M1 SOFA: [0 - 3)	1.185 (0.29)	0.87 (0.40)	1.434 (0.26)	0.008 (<2e-16)	0.002 (<2e-16)	0 (0.96)
M2 SOFA: [3 - 6)	0.922 (0.56)	1.193 (0.22)	0.939 (0.82)	52.1 (<2e-16)	1.096 (0.71)	0 (0.97)
M3 SOFA: [6 - 10)	0.953 (0.82)	0.941 (0.78)	0.716 (0.42)	8.124 (3.2e-10)	436.502 (< 2e-16)	108.058 (< 2e-16)
M4 SOFA: [10 - 18)	0.734 (0.51)	0.614 (0.33)	0.673 (0.68)	0.723 (0.79)	29.86 (1.6e-05)	553.359 (2.5e-15)

At a significance threshold of $p < 0.01$, no intervention is significantly associated with transition.

Chapter 5 The Impact of Laboratory Data Missingness on Sepsis Diagnosis Timeliness

INTRODUCTION

A cornerstone of sepsis treatment is the timely administration of appropriate antibiotics, ideally before the onset of organ failure.[122] However, in clinical practice the diagnosis of sepsis may be delayed based on the availability of laboratory results.[123] For example, there may be delay in collecting, processing, and reporting data from lab tests that are needed to meet the Centers for Medicare & Medicaid Services (CMS) defined organ failure criteria. This delay due to data missingness may worsen patient outcomes as evidence has suggested that every 1-hour delay in antibiotic administration leads to a 4-7% increase in odds of mortality risk for sepsis.[9,12124] Thus, understanding data missingness in patients with sepsis might be important for improving sepsis treatment.

The objective of this research was to measure the extent of incomplete data among hospitalized patients with sepsis, focusing on the timing of organ failure as determined by evaluations of vital signs and lab tests recorded in electronic health records (EHRs). We hypothesize that patients with sepsis with a time of organ failure after time of clinical suspicion of sepsis will have increased data missingness compared to patients where time of organ failure occurred before time of suspicion. According to the CMS SEP-1 quality measure, timelines for the 3 and 6 hour bundle should begin when a clinician makes a diagnosis of severe sepsis or septic shock or as soon as SIRS, organ failure, and infection criteria are all met within a 6 hour window.[125] If organ failure criteria are not met because of missing values, then SEP-1 bundle initiation and completion may be delayed until the time of organ failure as opposed to the earlier documented time of infection in retrospective analysis. This concept also applies to clinical implementation of the sepsis-3 criteria which requires access to laboratory results that establish

organ failure, via calculation of the Sequential Organ Failure Assessment (SOFA) score, prior to diagnosis of sepsis.[2]

MATERIALS AND METHODS

Patient Population

We conducted an observational retrospective multi-hospital cohort study of adult patients admitted to two hospitals between January 1, 2023 and October 30, 2023 within the University of California San Diego (UCSD) Health system. Patients that met the CMS criteria for sepsis were tagged and included for further analysis. The UCSD Institutional Review Board approval was obtained with waiver of informed consent (#805726, AIVIS: Next Generation Vigilant Information Seeking Artificial Intelligence-based Clinical Decision Support for Sepsis, approved November 15, 2022). All study procedures were followed in accordance with the ethical standards of the responsible committee on human experimentation and with the Helsinki Declaration of 1975.

Feature Extraction

Data were extracted from the EHR through Epic Clarity (Epic Systems). Extracted variables included patient demographics, care unit locations, timestamps of antibiotic and blood culture orders, vital sign measurements (heart rate, respiration rate, temperature, mean arterial pressure, and systolic blood pressure), and laboratory measurements (white blood cell count, bands, total bilirubin, creatinine, lactate, platelets, and partial thromboplastin time (PTT)).

We identified time of organ failure (T_{OF}) using the CMS criteria for sepsis.[126] CMS organ failure is defined as the presence of two or more systemic inflammatory response syndrome (SIRS) criteria combined with at least one element of organ dysfunction as assessed through vitals, laboratory measurements, urine output, and the presence of acute respiratory failure requiring ventilatory support (Figure 5.1).[127] Time of suspected infection ($T_{suspicion}$) was defined as the minimum of blood culture tests and antibiotic initiation (for at least 4 days, excluding prophylactic use) within 24 or 72 hours, depending on whether culturing or antibiotic administration occurred first, respectively.[2,128]

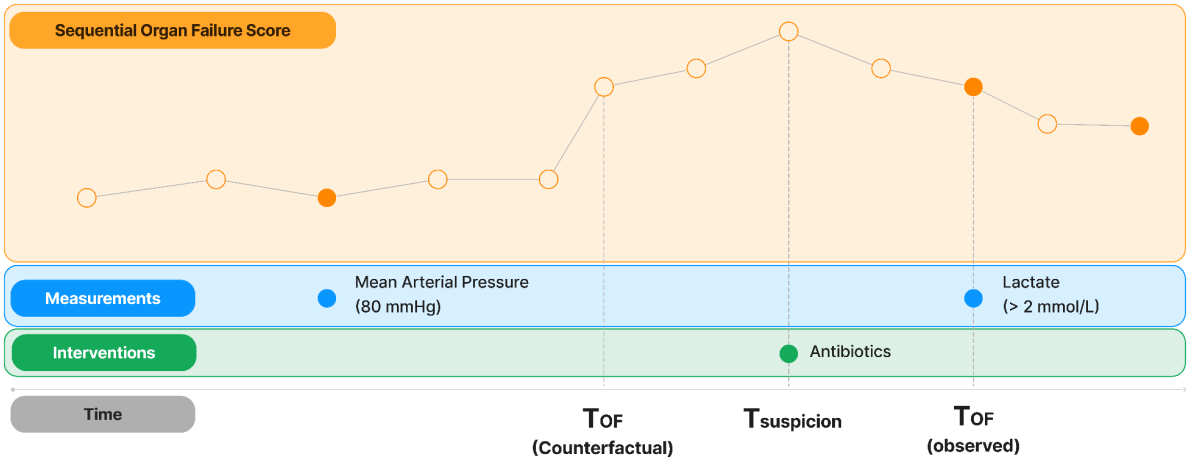


Figure 5.1: Graphical representation of CMS criteria for sepsis over time.
The top panel (orange) shows an illustrative hourly time series of the sequential organ failure score. Filled circles represent observed hourly values while unfilled circles represent latent values. The middle panel (blue) shows example measurements for assessing the presence of organ failure and the bottom panel (green) shows a sample intervention which establishes the time of suspicion ($T_{suspicion}$). In this example, the time of organ failure (T_{OF}) occurs after $T_{suspicion}$. If measurements had been taken earlier, however, the presence of organ failure would have been established prior to $T_{suspicion}$.

Patient encounters were binned into emergency department (ED), intensive care unit (ICU), or wards based on care unit location and further divided into two cohorts within each

location category depending on whether T_{OF} occurred before $T_{suspicion}$ (the 'before' cohort) or T_{OF} occurred after $T_{suspicion}$ (the 'after' cohort). Missingness of the five CMS laboratory components was determined from whether the lab was present within the 6 hours prior to $T_{suspicion}$, because SEP-1 guidelines require an element of organ failure, SIRS criteria, and suspicion of infection within a 6-hour window. A laboratory component was labeled as 'contributory' if the laboratory measurement ultimately contributed to organ dysfunction. This categorization aimed to increase the specificity in pinpointing missing data that could be potentially linked to a delayed recognition of organ failure.

Active Sensing Implementation

Figure 5.2 shows the design of the COMPOSER algorithm modified for active sensing. The AI Vigilant Information Seeking (AIVIS) system assesses the availability of data against clinical constructs (such as the SIRS criteria and the CMS SEP-1 bundle) and requests additional information to assist with the diagnostic process. The requests are handled through three distinct flowsheet entries tied to BPAs that (1) request vitals collection from the nursing team, (2) request additional lab order sets, and (3) request confirmation of a suspected or worsening infection. If the requested elements confirm a negative diagnosis of sepsis, the system is snoozed for a prespecified period. AIVIS is currently deployed in silent mode across EDs, ICUs, and Wards at UCSD Health.

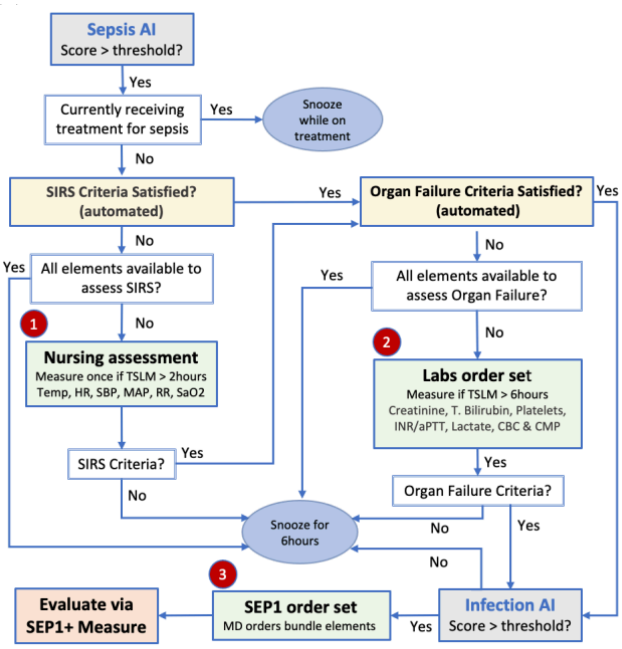


Figure 5.2: Schematic of the AIVIS system.

Statistical Analysis

P values were calculated between the 'before' and 'after' groups using the chi-square test for individual laboratory components and t-test for overall missingness between means of aggregated laboratory components. A p-value < 0.05 was considered statistically significant.

RESULTS

We identified a total of 625 patients who developed sepsis either in the ED, ICU, or wards (Supplementary Table 5.1). 71.9% of patients with sepsis had T_{OF} after $T_{suspicion}$ in the ED compared to 41.6% in the ICU and 61.5% on the wards (Table 5.1). For patients with T_{OF} after $T_{suspicion}$, the median (IQR) time difference was 1.8 (0.8-12.6) hours in the ED and 7.5 (2.4-25.4) hours in the wards. In both the ED and wards, data missingness across the 5 CMS laboratory measurements was significantly greater in patients with T_{OF} after $T_{suspicion}$. This was not observed

in the ICU. We did not observe missing values in the vitals and ventilation components of the CMS sepsis definition and excluded them from further analysis.

Table 5.1: CMS laboratory missingness (out of 5 labs) in patients with sepsis where organ failure occurred before and after $T_{suspicion}$.
The 'Before' cohort is defined as patients with T_{OF} before $T_{suspicion}$, and the 'After' cohort defined as patients with T_{OF} after $T_{suspicion}$. Number of patients presented as N (%). Missingness data presented as mean (SD).

Unit	Number of patients in the 'Before' cohort	Missing Values in the 'Before' cohort	Number of patients in the 'After' cohort	Missing Values in the 'After' cohort	p-value
ED	87 (28.1%)	2.74 (1.59)	223 (71.9%)	3.92 (1.43)	**<0.001**
ICU	90 (58.4%)	3.4 (1.54)	64 (41.6%)	3.58 (1.5)	0.47
Wards	62 (38.5%)	3.35 (1.45)	99 (61.5%)	4.08 (1.27)	**0.001**

Missingness of each CMS laboratory value was significantly greater in the 'after' cohort for all laboratory values in the ED 6 hours prior to $T_{suspicion}$ (Supplementary Table 5.2). This result was the same in the wards with the exception of PTT. When evaluating only the contributory CMS laboratory components that contributed to organ dysfunction in the 6 hours prior to T_{OF}, missingness in each laboratory feature except PTT was significantly greater in the 'after' cohort within the ED and only lactate and platelets were significantly greater in the wards (Supplementary Table 5.3).

DISCUSSION

Our study identified that there were many patients in whom the recognition of organ failure occurred after the time of clinical suspicion of sepsis (T_{OF} after $T_{suspicion}$). Most patients in the ED had T_{OF} after $T_{suspicion}$, likely reflecting the characteristics of patients who present to the ED, the delays in lab reporting, and a practice location where clinicians maintain a high suspicion for sepsis. In contrast, there were proportionally fewer patients in the ICU who had T_{OF} after $T_{suspicion}$, which might reflect frequent laboratory testing. Thus, in theory, more prompt laboratory assessment that would reveal organ dysfunction might improve the timeliness of recognition and treatment of sepsis.

Sepsis Early Warning Systems

The implications of these findings can be extended to the field of automated in-hospital warnings and alerts where notifications coming at or after clinical suspicion are not useful.[129,130] A number of machine learning-based early warning systems attempt to reduce the time to sepsis recognition by sending alerts for predicted at-risk patients before clinical suspicion.[21,35,48,79.80,] However, models often wait for physicians to order labs after clinical suspicion, leading to scenarios where such alerts are not sent preemptively for at-risk patients.[129,130] Given our findings, it is likely that some of these potential false negatives may be mitigated by including more recent lab results and that approaches for advancing the collection of critical labs may play an important role in improving early notification, especially for sepsis. Our future work will attempt to address this need through the development of an automated AI-based system that will be capable of submitting laboratory order sets on high-risk patients when it detects a suspected

sepsis diagnosis before human clinical suspicion. While the sampling frequency of EHR measurements can vary according to unit locations, patient acuity, and workflow practices, generalizable prediction tools need to possess an ability to function reliably in settings with missing data as well as complete data.[131]

Limitations

It is possible that during this retrospective analysis patients may have had a time of sepsis or time of organ failure different from the ground truth or suffer from a condition that mimics but is not sepsis. Secondly, we binned all care unit stays that were not in the ED or ICU into the wards cohort, meaning there is considerable heterogeneity in the units considered as part of this categorization. It is also unknown whether our results are consistent beyond the 10-month timeframe. Finally, there was no way to assess if patients had organ failure prior to the eventually measured lab value, but organ failure is often caused by an underlying process that occurs more than a few hours in advance.

CONCLUSION

In this study, we found significantly higher levels of data missingness in laboratory values used for the CMS sepsis definition, especially lactate, in patients who developed sepsis in the ED or wards where time of organ failure occurred after time of clinical suspicion of sepsis. These differences were minor in ICU patients. Patient outcomes in septic patients could be improved by ordering laboratory results earlier for more timely treatment.

ACKNOWLEDGEMENTS

Chapter 5, in part, has been submitted for publication of the material as it may appear in JAMIA Open as “The Impact of Laboratory Data Missingness on Sepsis Diagnosis Timeliness”. Boussina, A., Lam, J., …, & Nemati, S. The dissertation author was the primary investigator and author of this paper along with co-first author Dr. Jonathan Lam.

SUPPLEMENTARY MATERIAL

Supplementary Table 5.1: Patient characteristics

	ED cohort	ICU cohort	Wards cohort
# Encounters (%)	311	154	161
Age (in years), median [IQR]	60.1 [48.3 72.1]	60.0 [45.7 67.7]	58.1 [44.7 67.7]
Gender (Male), %	55.8%	60.3%	55.6%
Race			
White, %	48.7%	39.7%	51.6%
African American, %	10.4%	6.6%	6.4%
Asian, %	6.8%	4.4%	4.8%
Hospital Length of Stay (in hours), median [IQR]	329.5 [177.9 635.5]	580.1 [398.1 952.2]	525.7 [358.0 814.4]
CCI, median [IQR]	2 [1 4]	1 [0 4]	2 [0 5]
SOFA, median [IQR]	7 [4 9]	9 [7 11]	6 [4 9]
Mortality, %	11.0%	16.2%	5.6%
Transfer to ICU, %	71.2%	N/A	64.1%
Time from unit admission to T_{sepsis} (in hours), median [IQR]	8 [4 23]	25 [12 88]	10 [5 24]

Supplementary Table 5.2: Missingness in all CMS laboratory components in ED and Wards in the 6 hours prior to $T_{suspicion}$.
Each encounter belongs to either the 'Before' or 'After' cohort, and missingness in each encounter is determined for all 5 labs in the 6 hours prior to $T_{suspicion}$. Bolded p-values indicate labs with significant differences.

Feature	Unit	Encounters in the 'Before' Cohort Missing Feature	Encounters in the 'Before' Cohort Not Missing Feature	Encounters in the 'After' Cohort Missing Feature	Encounters in the 'After' Cohort Not Missing Feature	P-value
Bilirubin (Total)	ED	44	43	174	49	**<0.0001**
Creatinine	ED	43	44	168	55	**<0.0001**
Lactate	ED	63	24	207	16	**<0.0001**
Platelets	ED	38	39	146	77	**0.0004**
PTT	ED	62	25	195	28	**0.0007**
Bilirubin (Total)	Wards	40	22	87	12	**0.0004**
Creatinine	Wards	24	38	69	30	**0.0001**
Lactate	Wards	52	10	97	2	**0.0009**
Platelets	Wards	30	32	67	32	**0.0149**
PTT	Wards	50	12	90	9	**0.0599**

Supplementary Table 5.3: Missingness in contributory CMS laboratory components in ED and Wards that contributed to organ dysfunction in the 6 hours prior to T_{OF}.
Each encounter may have between 0 to 5 labs that were 'contributory' with respect to organ dysfunction. Bolded p-values indicate labs with significant differences.

Feature	Unit	Encounters in the 'Before' Cohort Missing Feature	Encounters in the 'Before' Cohort Not Missing Feature	Encounters in the 'After' Cohort Missing Feature	Encounters in the 'After' Cohort Not Missing Feature	P-value
Bilirubin (Total)	ED	1	9	23	2	**<0.0001**
Creatinine	ED	0	6	14	7	**0.0039**
Lactate	ED	1	20	118	2	**<0.0001**
Platelets	ED	0	8	19	3	**<0.0001**
PTT	ED	0	1	3	1	0.1709
Bilirubin (Total)	Wards	8	2	5	2	0.6818
Creatinine	Wards	2	6	3	1	0.0977
Lactate	Wards	2	6	26	1	**<0.0001**
Platelets	Wards	4	10	7	0	**0.002**
PTT	Wards	0	4	2	2	0.1025

Chapter 6 Machine learning based DNA melt curve profiling enables automated novel genotype detection

Surveillance for genetic variation of microbial pathogens, both within and among species, plays an important role in informing research, diagnostic, prevention, and treatment activities for disease control. However, large-scale systematic screening for novel genotypes remains challenging in part due to technological limitations. Towards addressing this challenge, we present an advancement in universal microbial high resolution melting (HRM) analysis that is capable of accomplishing both known genotype identification and novel genotype detection. Specifically, this novel surveillance functionality is achieved through time-series modeling of sequence-defined HRM curves, which is uniquely enabled by the large-scale melt curve datasets generated using our high-throughput digital HRM platform. Taking the detection of bacterial genotypes as a model application, we demonstrate that our algorithms accomplish an overall classification accuracy over 99.7% and perform novelty detection with a sensitivity of 0.96, specificity of 0.96 and Youden index of 0.92. Since HRM-based DNA profiling is an inexpensive and rapid technique, our results add support for the feasibility of its use in surveillance applications.

INTRODUCTION

Effective large-scale monitoring and surveillance of novel pathogens are critical components of contemporary public health strategy.[132,133] Genotypic screening of pathogens, in particular, enables researchers and clinicians to gain nuanced insights into disease transmission patterns, virulence factors, antibiotic resistance profiles, and potential outbreak sources.[134] Modern techniques for genotypic screening, however, can be cost-prohibitive, slow, and intractable to scale.[135] High Resolution Melting (HRM) analysis offers an emerging alternative

that enables rapid, effective, and economical post-amplification nucleic acid characterization for profiling DNA sequences.[136,137,138,139,140,141,142]

HRM involves the use of a DNA-binding dye added to DNA samples, which fluoresces upon intercalating into the double-stranded structure. As the sample is heated, the DNA denatures, resulting in a loss of fluorescence which is recorded to produce a melt curve. Recent advancements have positioned HRM as not just a check on amplification product homogeneity, but also as a reliable method for heterozygote detection.[136,143] Moreover, with improvements in heat transfer and reaction engineering, homozygous melt curves can now be leveraged as sequence-specific signatures.[60,61,144,145]

When applied to the problem of identifying and differentiating bacteria, universal primers have been used to target the 16S rRNA gene, which is highly conserved in bacteria but contains variable regions that are specific to different species or strains. This enables broad-based amplification of the bacterial 16S gene with universal primers, while relying on melt to genotype the hypervariable sequences specific to organism identity. The use of machine learning (ML) classification, where each unique sequence-specific curve signature represents a pathogen class, offers a principled framework to utilize HRM as a broad-based sequence profiling tool. This could be especially valuable in clinical diagnostics, where specifically identifying pathogenic bacteria is crucial for determining the appropriate treatment.[146,147,148]

The growth in the amount of HRM data available for ML training has further accelerated the practical application of these methods for pathogen profiling. The rising prominence of digital PCR has led to the evolution from traditional HRM to digital HRM (dHRM) which enables 200-fold increases in the number of melt curves.[58,59] This methodology originated from prior work in the Fraley Lab which introduced a unique dHRM platform that employs

specialized heat transfer and imaging mechanisms to simultaneously melt thousands of digital PCR (dPCR) reactions.[60,61,62] A distinct feature of this methodology is its digital design, in which each reaction is characterized by the presence or absence of a genome as its DNA template. While traditional HRM might typically operate on a 96-well plate, our dHRM technique utilizes a dPCR chip with 20,000 partitions, resulting in 20,000 HRM curves.

Existing literature has demonstrated effective application of Naïve Bayes (NB), Support Vector Machines (SVM), k-Nearest Neighbors using Dynamic Time Warping, Random Forest (RandF), and Neural Networks, for the purpose of pathogen classification using melt curve data.[149,144,145,150,151,152,153] SVM algorithms have shown notable performance even with relatively few melt curves. However, in the traditional one-versus-all application for multiclass classification, these methods are ill-suited to address melt curves that lie outside the distribution of their training set. Specifically, when presented with an out-of-distribution melt curve, these classifiers may erroneously classify the signature of an emerging pathogen as a known pathogen class. Thus, to enable broad-based pathogen surveillance, there is substantial promise in ML that can accurately identify the emergence of novel genotypes.

In this work, we leverage the massive datasets generated from our dHRM platform and broad-based 16S gene amplification strategy to evaluate the performance of common ML classifiers in the identification of novel and known bacterial genotypes. To our knowledge, this work represents the first study in which multiple broadly applicable ML classifiers have been investigated for the purpose of novelty detection with HRM. Previous work by Andini et al. utilized Naïve Bayes and a custom distance metric based on the Hilbert Transformation, but this approach was limited in efficacy since it aligned melt peaks to a single temperature.[149] We hypothesize that by incorporating a larger HRM dataset than previously reported, extracting

significant geometric features of the melt curves, selecting the optimal ML algorithm, and developing an appropriate metric of model confidence, we can more accurately automate the identification of novel genotypes. We test this hypothesis using both experimental samples and simulated melt curves generated using the uMelt tool.[154] We focus on the specific use-case of bacteremia in neonates in which a small number of bacterial organisms are typically implicated, but emerging and opportunistic infections can occur.[155]

MATERIAL AND METHODS

Bacterial Strains

The bacterial species used in this study and their corresponding melt curves are described in Table S6.1. These bacteria are the primary causative pathogens in cases of neonatal sepsis.[155,156] We obtained isolates from Dr. David Pride (University of California San Diego School of Medicine) as well as the American Tissue Culture Collection (ATCC, Old Town Manassas, VA). Bacteria were cultured in Lurie-Bertani (LB) broth or Tryptic Soy broth (TSB), as required, and incubated overnight at 37 °C.

Bacterial genomic DNA extraction and PCR

Following culturing, we performed DNA extraction using Wizard Genome DNA Purification kit (Promega Corporation, Madison, WI). We assessed the quality and concentration of the extracted DNA using spectrophotometric absorbance measurements and confirmed the identities of the species from sequencing. We prepared genomic DNA dilutions for use with dPCR and used the commercially available QuantStudio 3D Digital PCR 20 K chip v2 (Applied Biosystems, Foster City, CA) for amplification. We followed the manufacturer's recommended

process, but customized our reagents. The composition of our dPCR master mix is described in our prior work. It includes 1 µL of sample, 0.15 µM forward primer 5′-GYGGCGNACGGGTGAGTAA-3′ (Integrated DNA Technologies, Coralville, IA), 0.15 µM reverse primer 5′-AGCTGACGACANCCATGCA-3′ (Integrated DNA Technologies, Coralville, IA), 0.02 U/µL of Phusion HotStart Polymerase (Thermo Fisher Scientific, Waltham, MA), 0.2 mM dNTPs (Invitrogen, Carlsbad, CA), 1X Phusion HF Buffer containing 1.5 mM MgCl2 (Thermo Fisher Scientific, Waltham, MA), 2.5X EvaGreen (Biotium, Fremont, CA), 2X ROX (Thermo Fisher Scientific, Waltham, MA), and ultrapure PCR water (Quality Biological Inc., Gaithersburg, MD) to bring the total volume to 14.5 µL. We loaded the chip by spreading 14.5µL of the master mix per the manufacturer's recommendation. We then cycled the the dPCR on a flatbed thermocycler with the following cycle settings: an initial enzyme activation (98 °C, 30 s), followed by 70 cycles (95 °C, 30 s, 59 °C, 30 s, 72 °C, 60 s).

DNA melt curve generation and preprocessing

The architecture of our U-dHRM device has been previously described.[60,61] A copper plate hosts the microfluidic dPCR chip, separated by a thin layer of thermal grease to ensure efficient heat transfer. Temperature control is achieved through a thermoelectric module (TE Technology Inc., Traverse City, MI), PID controller (Meerstetter Engineering GmbH, Rubigen, Switzerland), Class 1/3B resistance temperature detector (RTD) (Heraeus, Hanau, Germany) embedded in the copper block, K-type thermocouple (OMEGA Engineering, Stamford, CT), and heat sink. We secure the device on-stage for optimal fluorescent imaging using a custom-made adapter. A Nikon Eclipse Ti microscope (Nikon, Tokyo, Japan) captures simultaneous fluorescent images from heat ramping the DNA-intercalating dye, EvaGreen (Ex/Em: 488

nm/561 nm) and the control dye, ROX (Ex/Em: 405 nm/488 nm). An automated image processing algorithm implemented in MATLAB is used to generate the melt curves. We perform background subtraction using the linear method described in Palais et al.[157] This horizontally aligns the tails of the melt curves with the x-axis, to ensure they are most similar to the theoretically predicted uMelt curves.

Feature engineering

The resulting preprocessed melt curves contain fluorescence loss values (−dF/dT) for 410 temperature steps in the range [51 °C, 92 °C]. We model this as a 1-dimensional time-series and apply a signature transform to each melt curve. The signature method is a non-parametric feature extractor that computes a series of integrals along a data path that fully capture its order and area.[158,159] The signature method is optionally time-shift invariant and is sensitive to the geometric shape of the path. We apply the signature transform on a rolling window with a kernel size of 20 and a stride of 8 across the time-series with time and basepoint augmentations and a signature depth of 3 to generate a set of features for the downstream classification tasks.

Machine learning model selection

We set out to compare five ML methods: Logistic Regression, Naïve Bayes (NB), Support Vector Machines (SVM), Neural Networks and Random Forest (RandF). To address correlation between the input signature features, we used L2 regularization for the logistic regression, SVM, and neural network models. At the end of this work we briefly discuss how calibrating the probabilities affects the results. We built and implemented all algorithms using the

scikit-learn package within Python programming language.[160] All data and code are available on https://github.com/aboussina/dHRM-novelty-detection.

Quantification of genotypic differences

To further assess the utility of our derived HRM signatures for the identification of distinct genotypes, we analyzed the ability of the aforementioned ML models to quantify the degree of genotypic differences between species. That is, beyond simply classifying a melt curve to a given species, we reformulated this as a regression problem where the target variable is a metric for genotypic difference. We calculated this metric by mapping our ten bacterial species on the SILVA phylogenetic tree and computing the patristic distance between the node of each species and the E. coli node.[161] We evaluated the performance of this regressor using the c-statistic as described in Austin et al.[162]

Leave-one-group-out (LOGO) experiments

We evaluated the capability of these machine learning models for the task of identifying novel melt curves, i.e. those belonging to species unrepresented in the training set, using a leave-one-group-out experimental design. For each of our ten bacterial species, we held out its melt curves and split the curves of the remaining nine species into training and test sets (80:20 ratio). Then, the held-out species' curves were added to the test set. The machine learning model was trained on the training set and then tested on its ability to recognize the curves of the held-out species as novel within the modified test set. This process was repeated for each of the ten species. The schematic of this approach is presented in Fig. 6.1. To measure our model's

efficacy in novelty detection, we used Youden's index, a metric that assesses the performance of a binary diagnostic test.

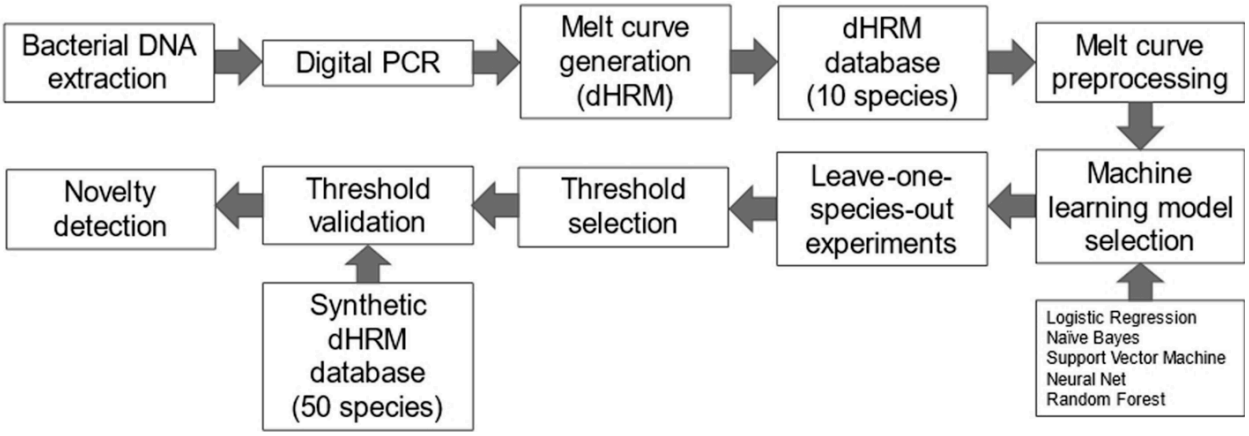


Figure 6.1: Workflow for ML novelty detection

Threshold selection

We set the threshold for novelty identification on a single LOGO experiment to the Youden Index (i.e. the point on an ROC curve at which sensitivity + specificity − 1 is maximized). However, to practically identify novel organisms, the ideal threshold should apply across all ten LOGO experiments. Thus, we calculate a 'practical' threshold by accumulating all the LOGO experiments and determining the Youden Index from their combined ROC curve. We then assess the performance of each LOGO experiment using this combined 'practical' threshold. We investigated applying sample weights to each experiment based on the number of left-out curves, but didn't observe a significant change in the selected threshold (data not shown).

Threshold validation with synthetic melt curves

We generated in-silico melt curves using uMelt for 50 clinically relevant bacterial pathogens, including category A and B biothreat agents and their surrogates from Yang et al. (Table S6.1).[163] We added real dHRM noise to these synthetic melt curves to more realistically

capture sample variation as described by Langouche et al.[152] We created 100 melt curves per species, with a unique noise residual applied to each individual melt curve.

RESULTS

Preprocessing

Figure 6.2 and Figs. S6.1–S6.2 show the results of background subtraction on the experimental melt curves. Also shown are the simulated melt curves from uMelt with added dHRM noise. Supplemental Fig. S6.3 shows an overview of the synthetically created melt curves for all 50 pathogens. As demonstrated, the background subtraction effectively aligns the curves with the x-axis. Further, the simulated curves show similar characteristics to the experimental observations.

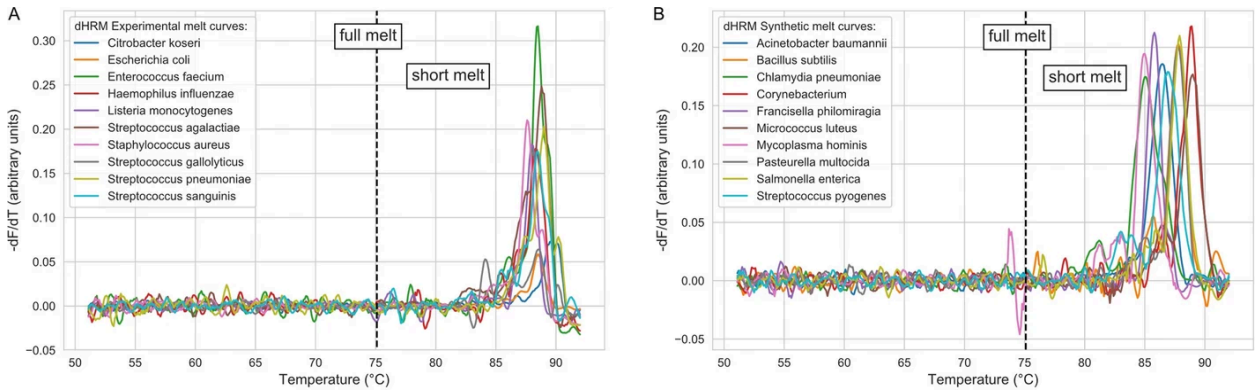


Figure 6.2: Overview of dHRM datasets.
A Experimentally obtained dHRM melt curves. **B** Ten examples of synthesized melt curves using a combination of uMelts and real dHRM melt curve noise. Full and short melt refer to using the entire length of the melt curve or a shorter window around the melt peak location.

Classification and regression performance

The ML methods show very similar classification accuracy, which is summarized in Table 6.1. No significant difference can be observed between the ‘full melt’ and ‘short melt’ classification results, which implies there might not be any additional information in the tail of

the melt curve. Figure S6.4 shows the correlation between melt curve distance (defined as the average pointwise distance from the average E. coli curve) and patristic distance from E. coli. We observe modest correlation with a couple of distinct outliers. The performance of a select ML method (RandF, n = 500) to quantify the patristic distance is listed in Table S6.2. We observe strong performance (c-statistic: 0.96) for this regression task indicating that the derived features enable quantification of the genotypic differences between organisms.

Table 6.1: ML methods overview and classification results.

	Full melt accuracy	**Short melt accuracy**
Logistic regression	0.996	0.997
Gaussian Naïve Bayes	0.963	0.975
SVM (rbf)	0.990	0.993
SVM (linear)	0.996	0.997
Neural net (identity)	0.994	0.997
Neural net (logistic)	0.997	0.998
Neural net (tanh)	0.996	0.998
Neural net (relu)	0.996	0.998
RandF (n = 100)	0.997	0.997
RandF (n = 500)	0.997	0.997

Leave-one-group-out (LOGO) experiments and threshold selection

Figure 6.3 shows the results of one LOGO experiment for one ML method (RandF, n = 500). The optimal threshold is found by plotting the ROC curve (Fig. 6.3B) and selecting its Youden index.

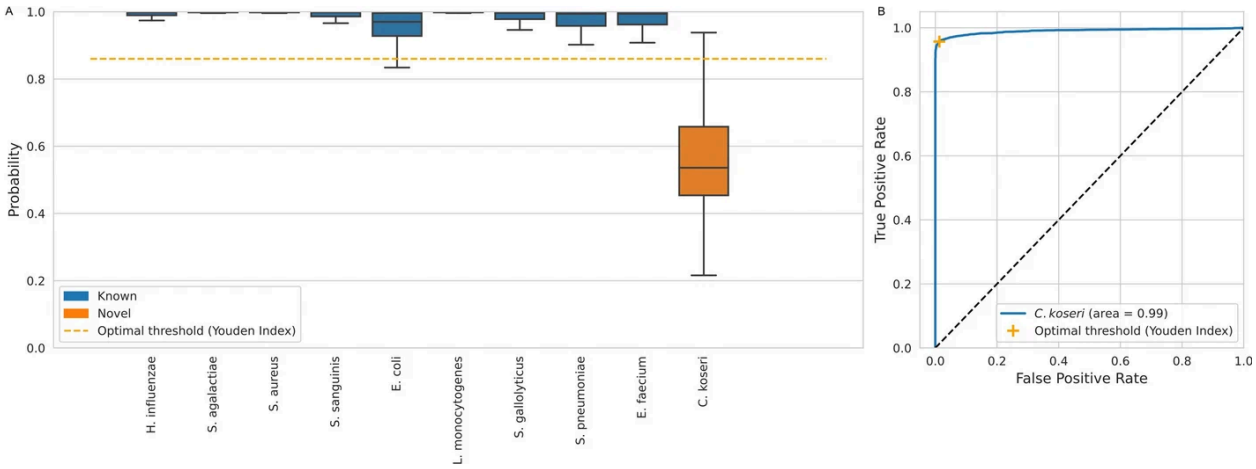


Figure 6.3: Leave-one-species-out cross validation to determine probability threshold. **A** Boxplot of the classification probabilities of each of the ten species. In this experiment C. koseri is the left-out species, which means it is left out of the training set and added to the test set. This experiment is repeated for each of the ten species, and an optimal threshold can be found for each of them. This experiment is repeated for all ML methods (method shown here is RandF (n = 500)). **B** ROC curve that is used to find the optimal threshold. Youden's Index is chosen as the optimal threshold, it is the point on the ROC curve where sensitivity + specificity − 1 is maximized.

Figures 6.4 and 6.5 summarize the results of the accumulated LOGO experiments. The ROC curves of the 10 individual LOGO experiments as well as the accumulated ROC curve are shown in Fig. 6.4. The performance, as measured by Youden's index is shown as a function of classification method in Fig. 6.5. Each bar shows the average performance across ten LOGO experiments. 'Optimal thresholds' means selecting the best threshold for each LOGO experiment individually. 'Practical threshold' means selecting the optimal threshold for the accumulated LOGO experiments and applying it to all LOGO experiments separately. Random Forest outperforms the other methods, but Neural Nets and SVMs still perform relatively well.

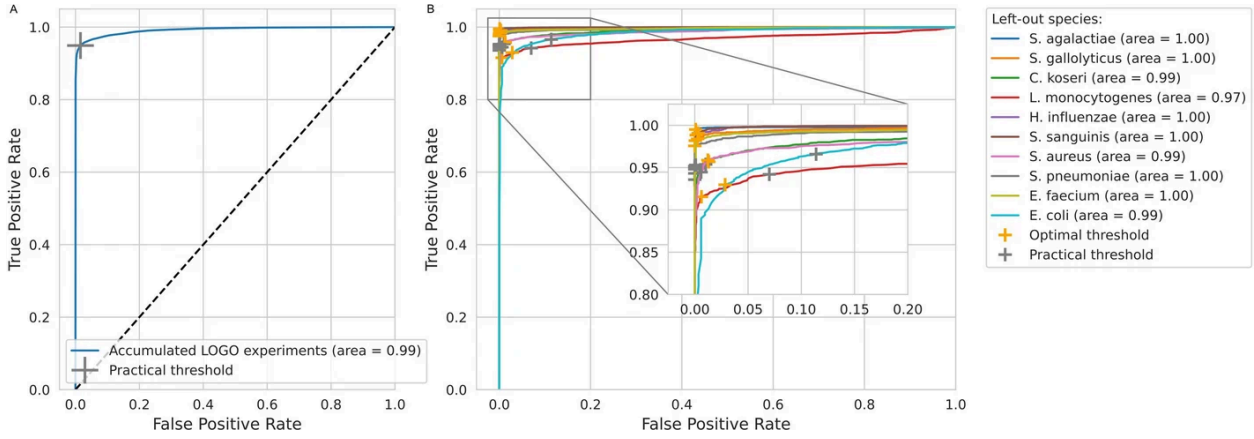


Figure 6.4: Accumulating the leave-one-group-out (LOGO) experiments results in a 'practical' threshold.
A ROC curve for all ten LOGO experiments accumulated with RandF (n = 500). The optimal threshold is again found by Youden's Index. We have named it the 'practical' threshold as one threshold has to be chosen (as opposed to a separate threshold for each LOGO experiment) when further validating the model on unseen 'novel' melt curves. It is the optimal threshold for all ten LOGO experiments combined. B Choosing a practical threshold implies that each LOGO experiment individually will be performing at a suboptimal threshold, which translates to a suboptimal operating point on the ROC curve.

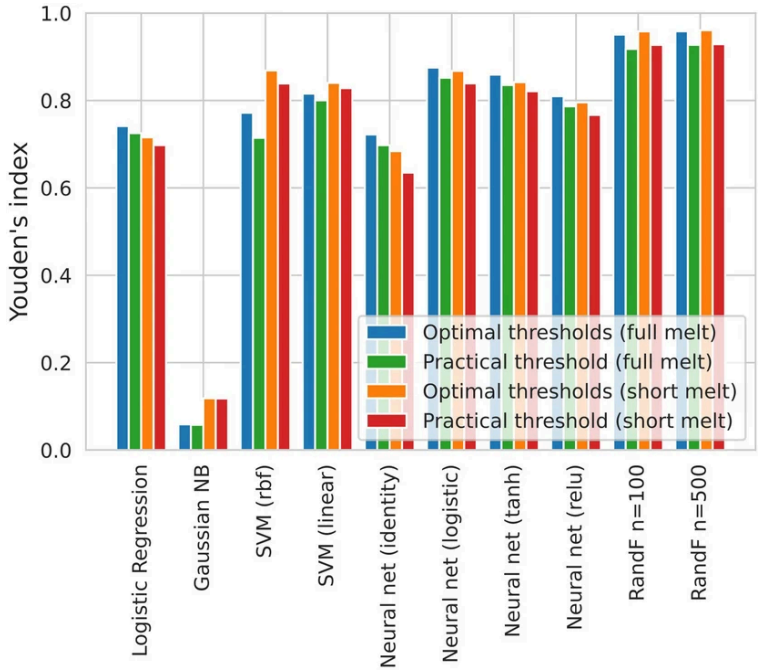


Figure 6.5: Summary of LOGO novelty detection results.
Average novelty detection performance across ten species measured by Youden's index as a function of classification method. Optimal means selecting the best threshold for each leave-one-species-out experiment. Practical means selecting one threshold and applying it to all leave-one-species-out experiments

Threshold validation and novelty detection

Figure 6.6 shows boxplots and ROC curves used to validate the previously obtained 'practical' threshold using the signature features (Fig. 6.6A, B) as well as the raw melt curves (Fig. 6.6C, D). The method shown in Fig. 6 is RandF (n = 100), which was one of the best performing methods, similar figures for all other methods are available in the supplementary data (Fig. S6.5). When the practical threshold is close to the optimal threshold, the practical operating point on the ROC curve (Fig. 6.6B) will be close to the optimal one (Youden's index). When this is the case, it confirms that our proposed method for obtaining a practical threshold for novelty detection, is indeed valid. Both feature sets (signatures and the raw melt curves) achieve strong discriminative performance of novel organisms based on the model score. The use of signature features enables an overall slight improvement but notably results in reduced novelty detection of T. pallidum. Figure 6.7A shows an overview of the results for all ML methods. The average difference between the optimal and practically attained Youden index across the ten ML methods is just 0.019 with a standard deviation of 0.025. This serves as a confirmation of our threshold selection process using the accumulated LOGO experiments. Random Forest and SVM (rbf) perform the best, with the Neural Nets a close third.

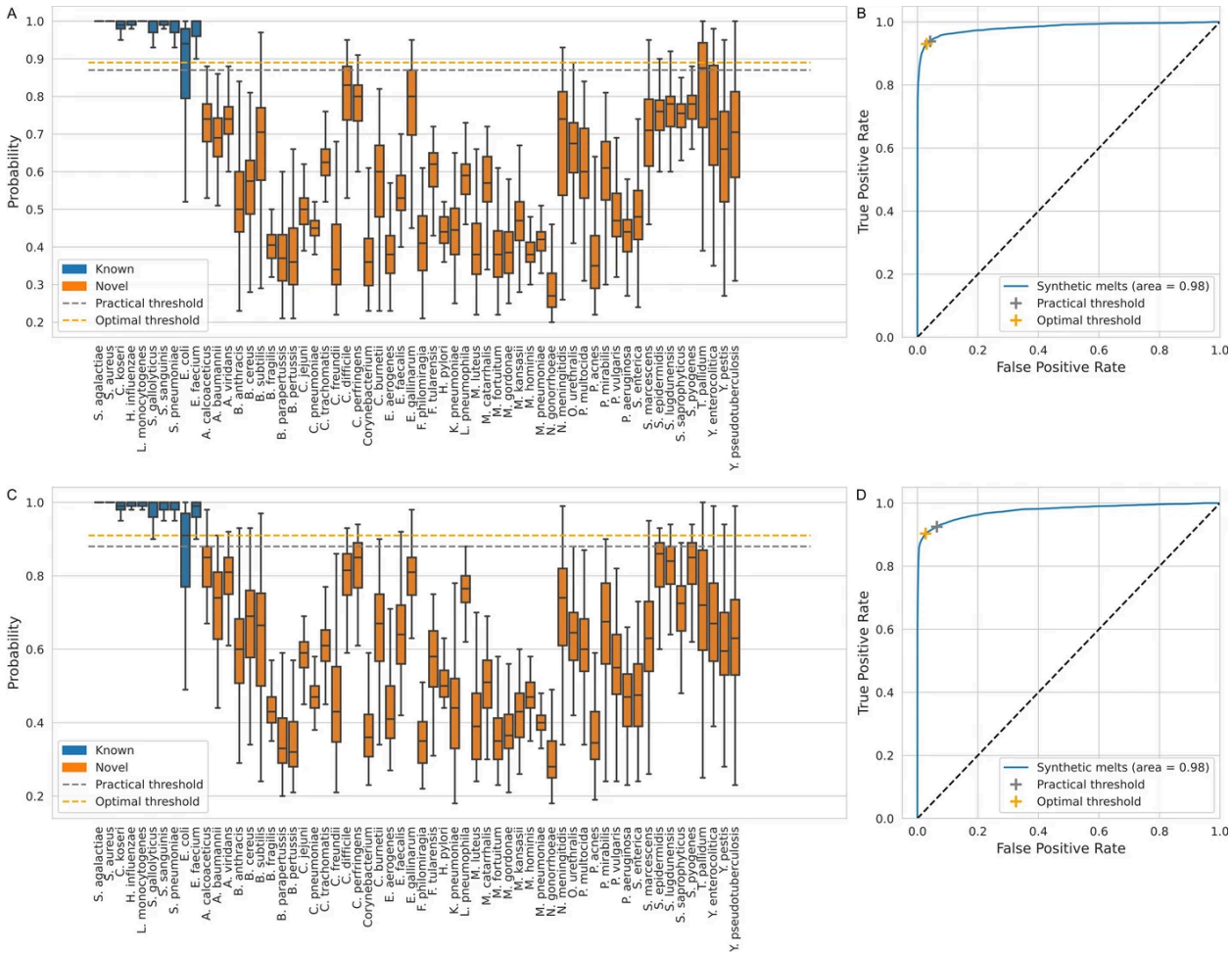


Figure 6.6: Validation of practical threshold on synthesized set of melt curves. **A** The practical threshold, selected through the LOGO experiments, was validated on the signature features from a new dataset consisting of 50 species, each with 100 synthetic melt curves with real dHRM noise. The performance of E. coli was confirmed to be an outlier using one-tailed t-tests. **B** ROC curve. When the practical threshold is close to the optimal threshold, it serves as a validation for the threshold selection process. C The practical threshold, repeated on the raw experimental and synthetic melt curves. D ROC curve using the raw melt curves. The ML method shown here is RandF (n = 100).

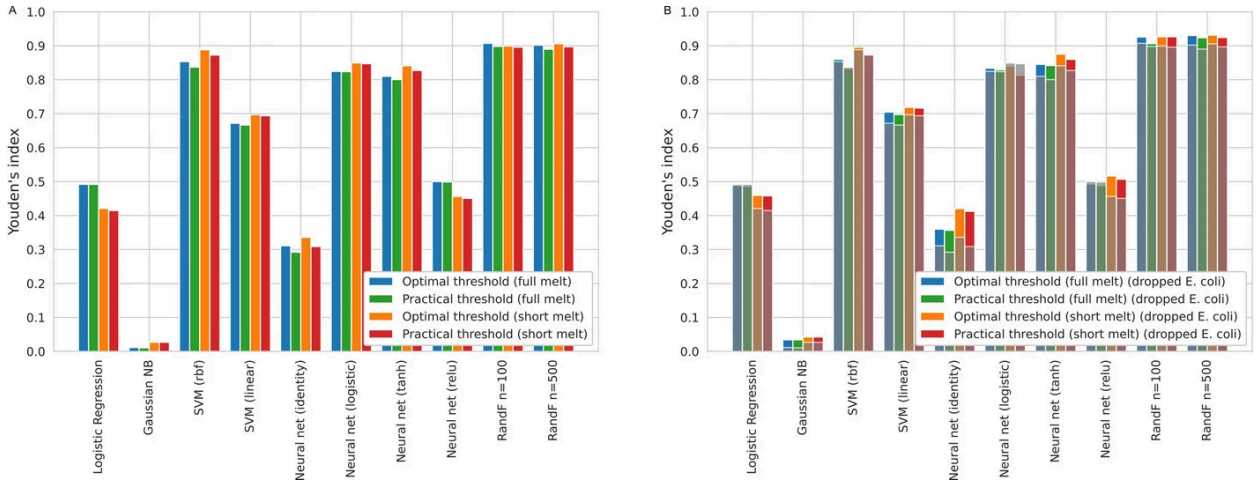


Figure 6.7: Summary of practical threshold validation results.
A Average novelty detection performance on 50 unseen species measured by Youden's index as a function of classification method. **B** Dropping E. coli, an outlier group, results in improved performance for almost all methods (results including E. coli are overlayed in gray). SVM (rbf) is more robust against this outlier behavior and sees less improvements.

Further improvements

We observed that the classification probabilities for the E. coli group of melt curves are more spread out (Fig. 6A) and might be an outlier group compared to the other species. This was apparent for multiple ML methods and was confirmed with one-tailed t-tests (e.g. for the short melt curves: $P < 0.01$ for Logistic Regression, Gaussian NB, SVM (rbf and linear), Neural Net (identity and relu) and RandF ($n = 100$ and $n = 500$)). As a result of this, we ran all steps again, but this time leaving out E. coli, to see if we could further optimize our novelty detection method. Figure 6.7B shows that results do indeed improve when leaving E. coli out. Random Forest and SVM (rbf) are the top performers, and their results are further summarized in Table 6.2. The best performance achieved is a Youden index of 0.93, corresponding to a specificity of 0.97 and sensitivity of 0.96.

Table 6.2: Overview of best results.

	Optimal specificity	**Optimal sensitivity**	**Optimal Youden**	**Practical specificity**	**Practical sensitivity**	**Practical Youden**
Full melt						
Support vector machine (rbf)	0.93	0.93	0.86	0.96	0.88	0.83
Random forest (n = 100)	0.98	0.95	0.93	0.94	0.96	0.91
Random forest (n = 500)	0.98	0.95	0.93	0.96	0.96	0.92
Short melt						
Support vector machine (rbf)	0.94	0.95	0.90	0.96	0.91	0.87
Random forest (n = 100)	0.97	0.96	0.93	0.97	0.96	0.93
Random forest (n = 500)	0.98	0.95	0.93	0.96	0.96	0.92

We also investigated whether calibrating the probabilities using scikit-learn's 'CalibratedClassifierCV' function would improve the outcome. We tested both the 'sigmoid' method, which corresponds to Platt's method (i.e., a logistic regression model) or the 'isotonic' method, which is a non-parametric approach. Results are summarized in Fig. S6.6. As expected, we see a large improvement for Naïve Bayes. We also see a significant improvement for Logistic

Regression. None of the calibrated methods outperform the best results (SVM, Neural Net, RandF) as outlined in Fig. 6.7 and Table 6.2 though.

DISCUSSION

Our work demonstrates the utility of time-series classification algorithms in resolving multiple bacterial organism melt curves, and in identifying previously unknown (novel) melt curves that are not represented in the database. The large amount of dPCR chip-generated melt curve data enabled the development of machine learning classifiers and novelty detection algorithms, which distinguishes this study from previous studies which did not assess out-of-distribution data and utilized small datasets of melt curves.[164,165]

The only other published method specifically aimed at melt curve novelty detection aligns the melt curves to one specific temperature, losing the useful melt peak location information in the process.[149] We have selected the most widely used ML methods in HRM analysis and have shown that they are all able to classify our dHRM database with very high accuracy. Interestingly, some drastically outperform others when it comes to novelty detection. We find that Neural Nets, SVMs, and Random Forest outperform the other ML methods, even after calibrating the probabilities. Random forests utilizing features extracted from a time-series have been shown to perform well on time-series classification tasks.[166] Here, we show that its well-calibrated probabilities are also particularly useful for conducting HRM novelty detection.

The performance of our approach was improved with the removal of data for the outlier species E. coli. One reason for the lower performance of E. coli compared to the other species could be that it has the lowest number of melt curves available (Table S6.1), which results in a smaller amount of training data available for the ML methods. We do not expect E. coli to

inherently have more heterogeneity in its sequence compared to other species. Melt curve shape variance might be another contributor to its outlier behavior as it has the third most variance in shape (from the ten species) as measured by dynamic time warping (DTW) in previous work.[152]

No major differences were found between using the full length and short version of the melt curves, although for most methods the short version does outperform the full length, showing that there might not be any additional information in the low-temperature tail of the curve, and including it might even confound the novelty detection performance.

There are several limitations to our work. First, while we benchmarked novelty detection across a suite of machine learning algorithms, we did not perform hyperparameter tuning. It is possible that approaches such as Bayesian optimization could improve novelty detection performance. Further, we did not utilize any distance-based metric for evaluating out-of-distribution novel organisms; opting instead to leverage the output scores from discriminative classifiers to select an optimal threshold. Our approach provides an effective way to incorporate novelty detection within a large machine learning framework, but future work is required to evaluate alternative distance-aware methodologies.[167]

In conclusion, advances in machine learning and ‘big data’ generation are opening up more opportunities for the advancement of HRM, which due to its speed, low cost, and simplicity was already attractive. The opportunity to use HRM as a discovery tool as well as profiling technology will further advance HRM technology towards its application in research and clinical diagnostics.

ACKNOWLEDGEMENTS

Chapter 6, in full, is a reprint of the material as it appears in “Machine learning based DNA melt curve profiling enables automated novel genotype detection”. Langouche, L., Obirieze, A. C., Sinha, M., Mack, H., Leineweber, W., ... & Fraley, S. I. (2024). BMC bioinformatics, 25(1), 185. The dissertation author was the primary investigator and author of this paper along with authors Dr. Lennart Langouche and Augustine C. Obirieze.

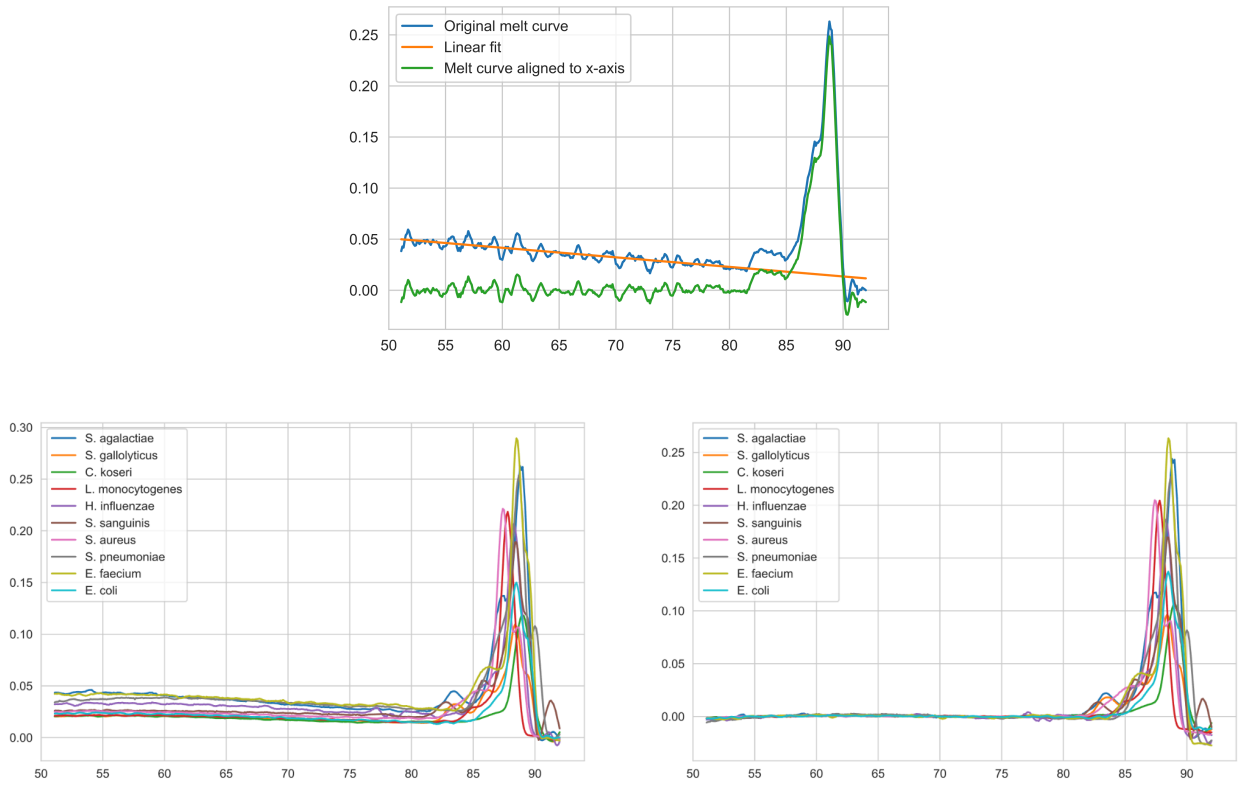


Figure S6.1. Data preparation.
Linear background is subtracted to make sure experimental melt curves resemble synthetic ones. Top: mechanism. Left: experimental mean curves. Right: experimental mean curves after background subtraction.

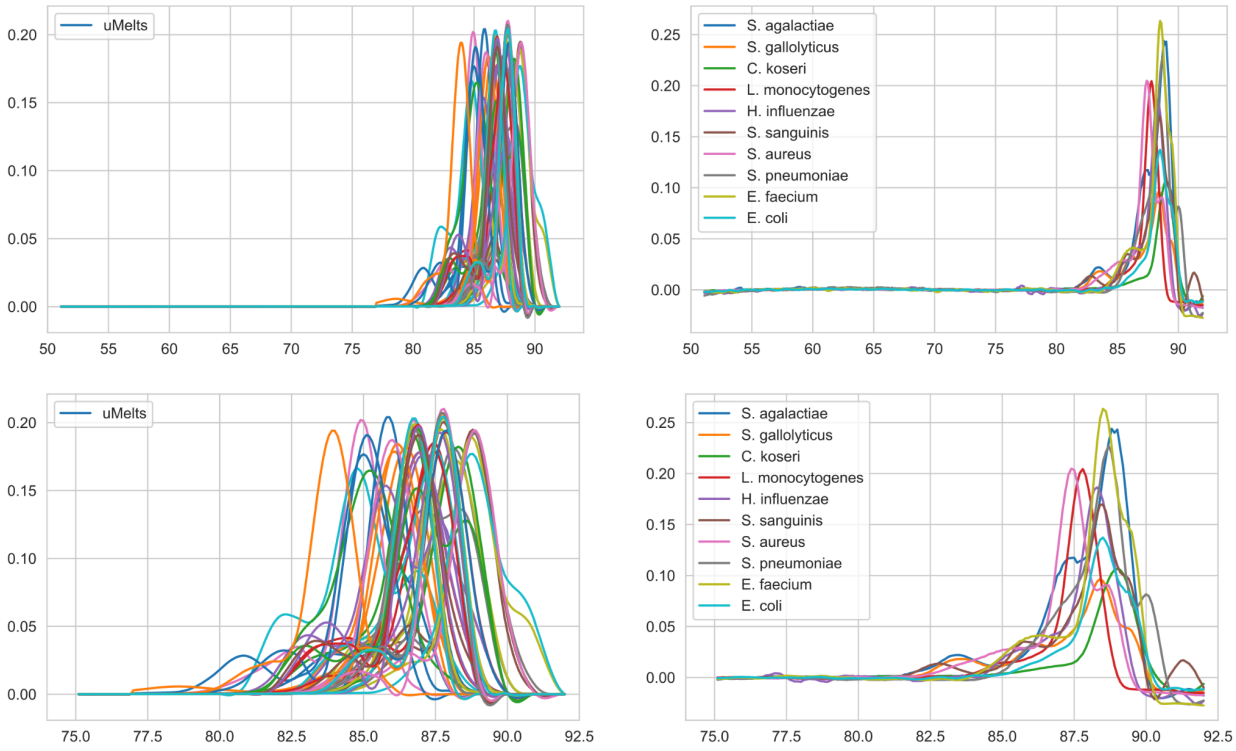


Figure S6.2. Overview of experimentally obtained and synthesized melt curves.
Top: full melts. Bottom: short versions. For the experimental curves (right), the mean curve is shown. For the uMelts (left) the original curve is shown (without applying dHRM noise). We wanted to assure these looked similar as the noise residuals to the mean (from the experimental data) are added to the uMelts.

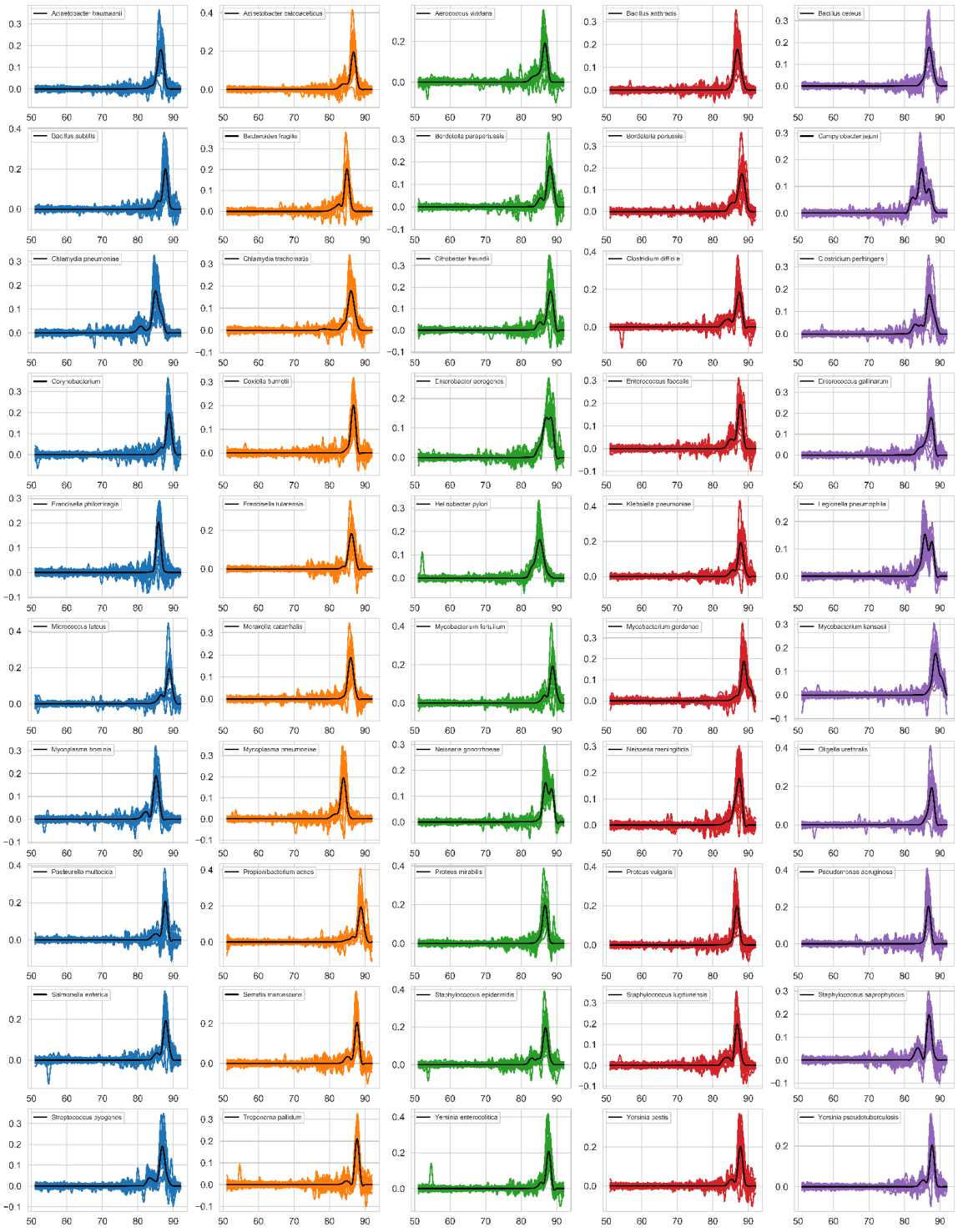


Figure S6.3. Overview of all synthesized melts. Residuals to the mean from each experimental melt curve are applied to the uMelts.
The noise is scaled to the ratio of the peak heights (real peak/uMelt peak) and shifted (position real peak – position uMelt peak) so the noise should be in a similar position relative to the peak. No residual is used twice which means all the curves shown are unique. Each uMelt uses ten residuals from each of the experimental melts, which results in 100 curves per uMelt species.

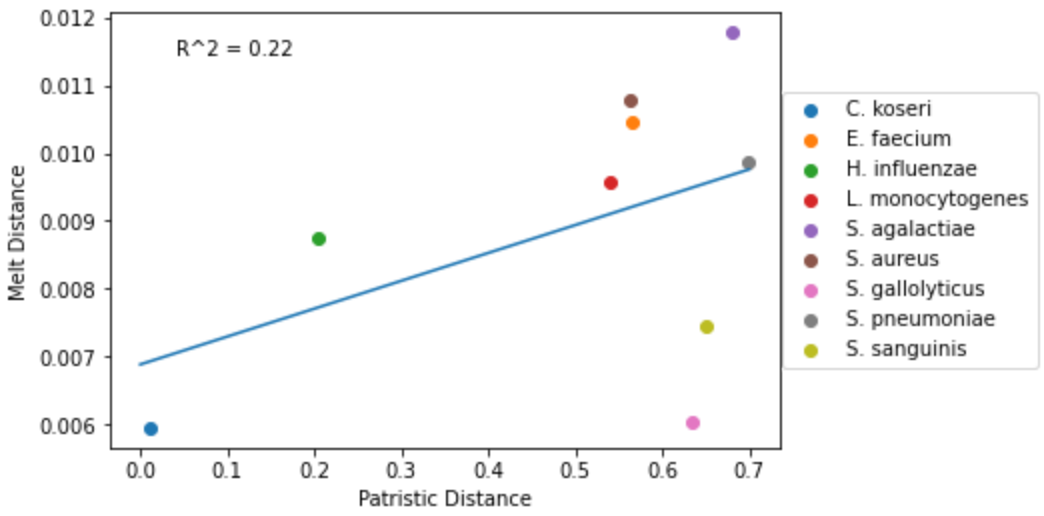


Figure S6.4: Correlation between melt distance and patristic distance.
Patristic distance was computed for each organism by utilizing the SILVA phylogenetic tree and calculating the distance from E. Coli.

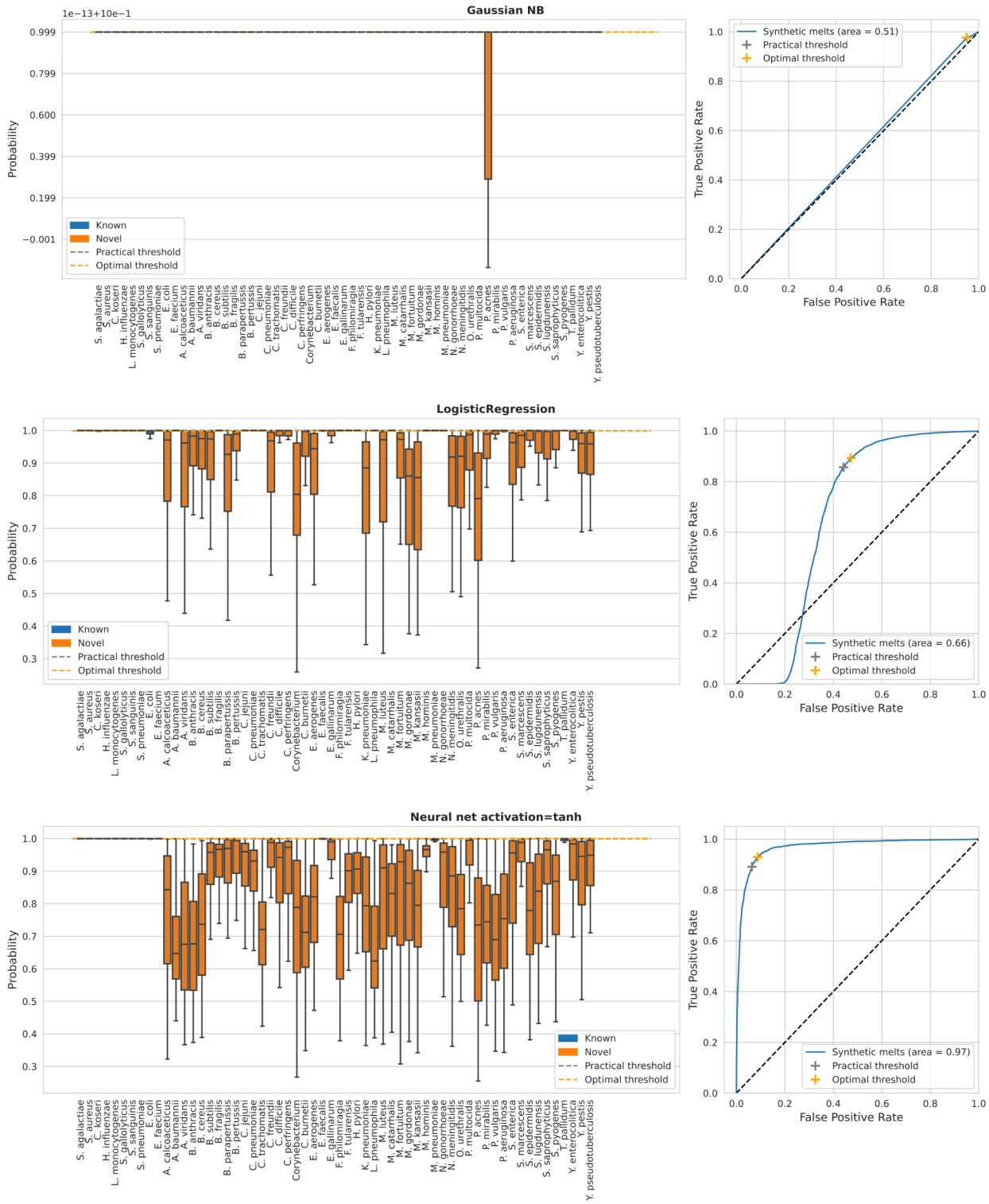


Figure S6.5. Validation of practical threshold on synthesized set of melt curves for all ML methods.

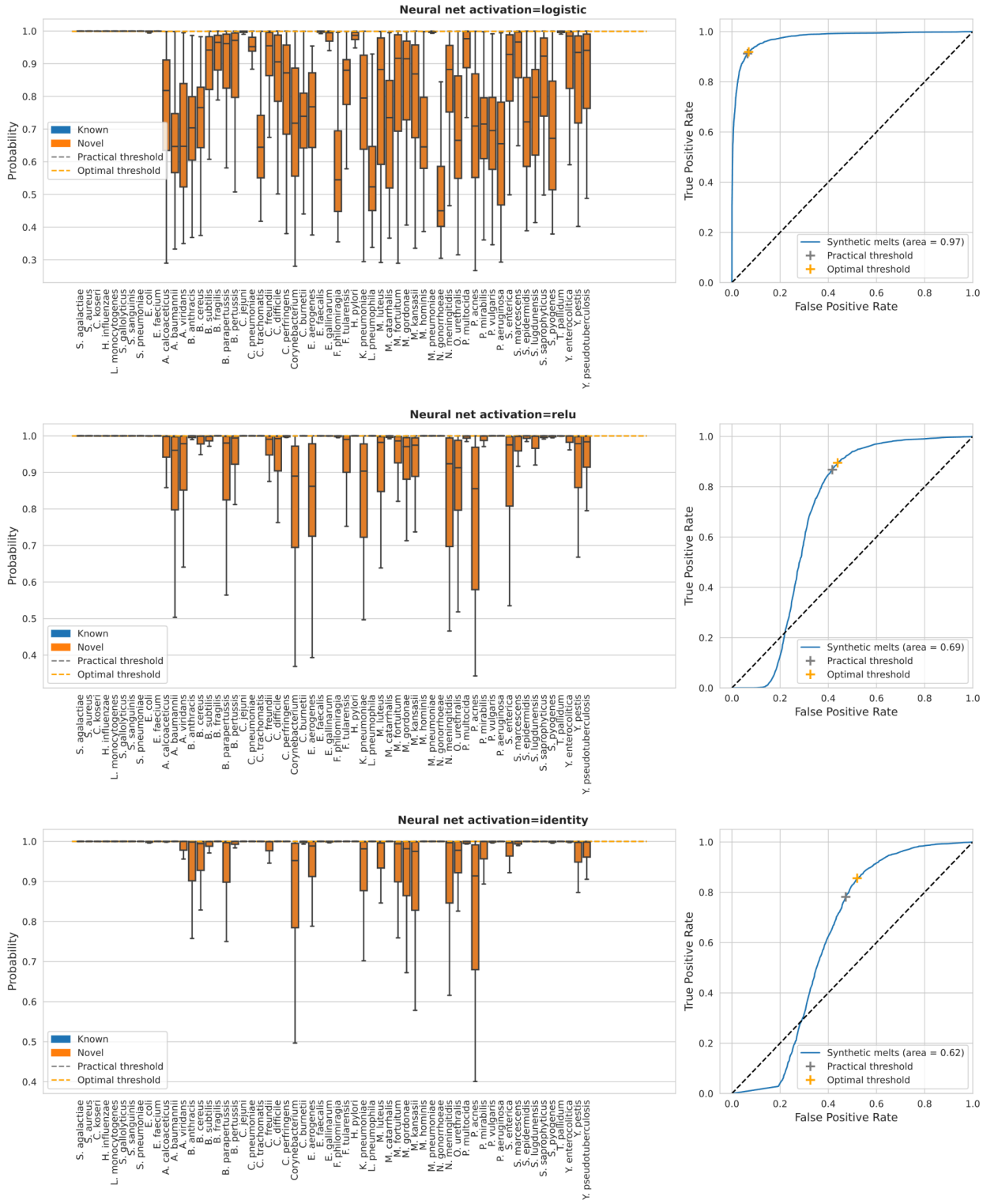


Figure S6.5. Validation of practical threshold on synthesized set of melt curves for all ML methods (Continued).

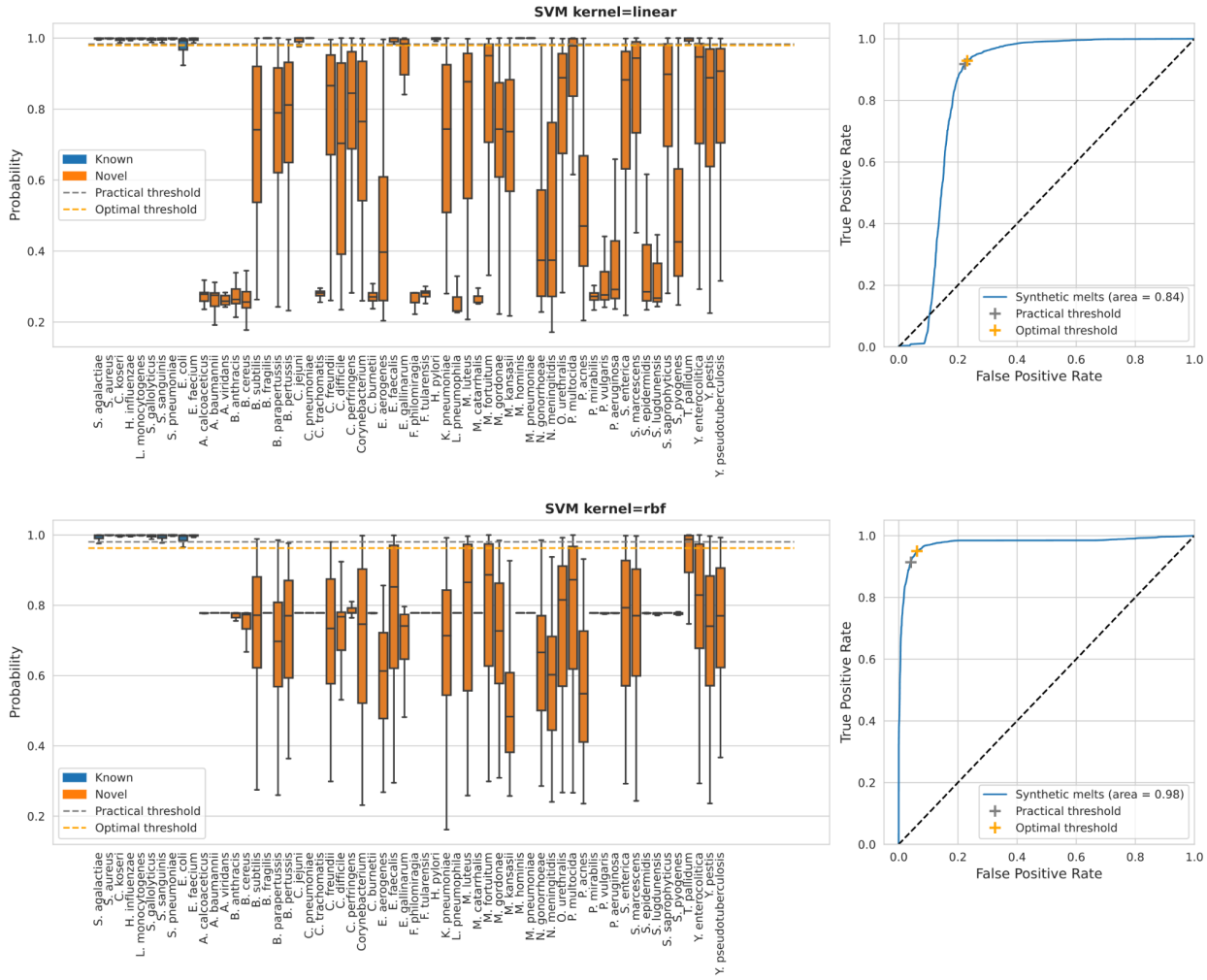


Figure S6.5. Validation of practical threshold on synthesized set of melt curves for all ML methods (Continued).

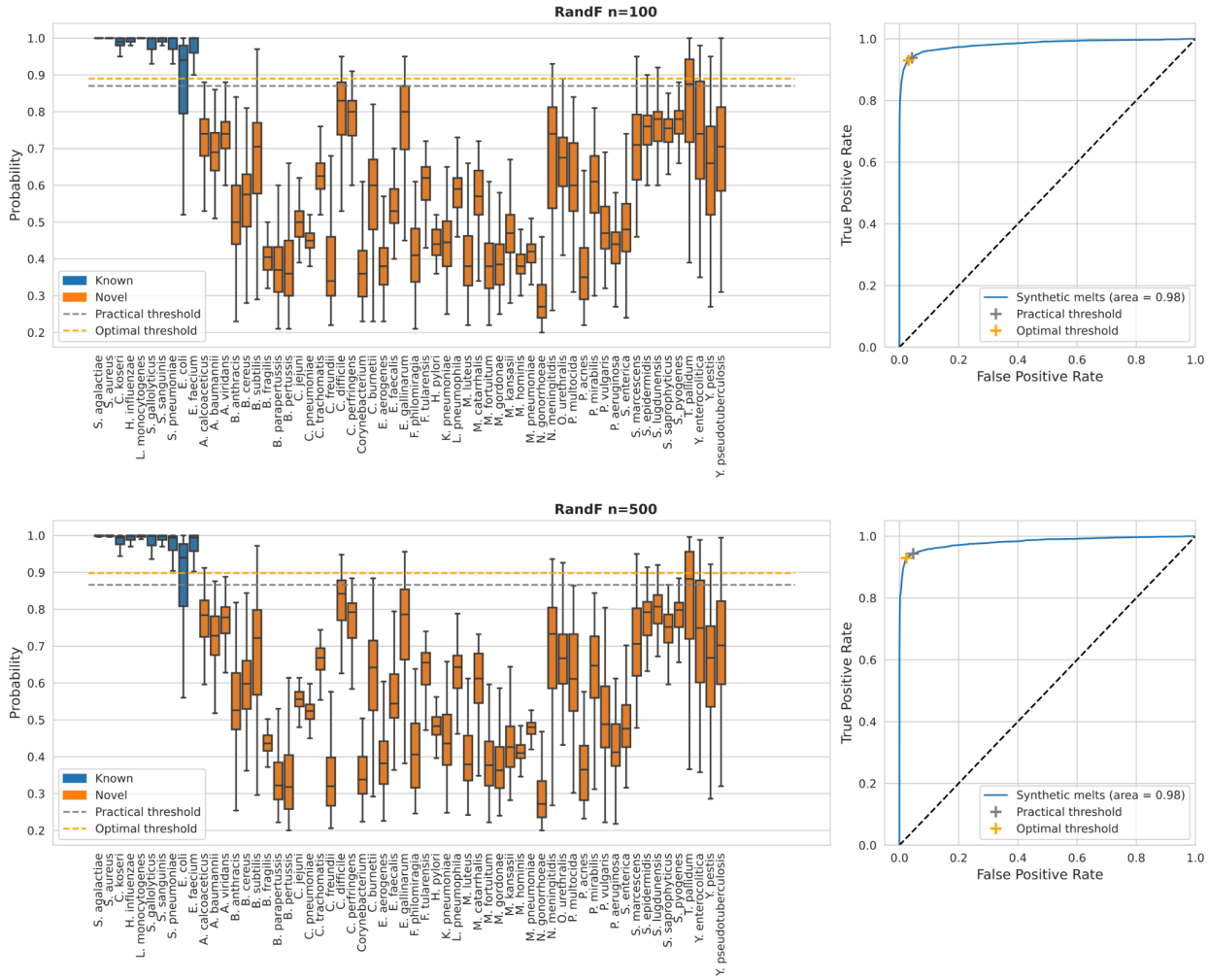


Figure S6.5. Validation of practical threshold on synthesized set of melt curves for all ML methods (Continued).

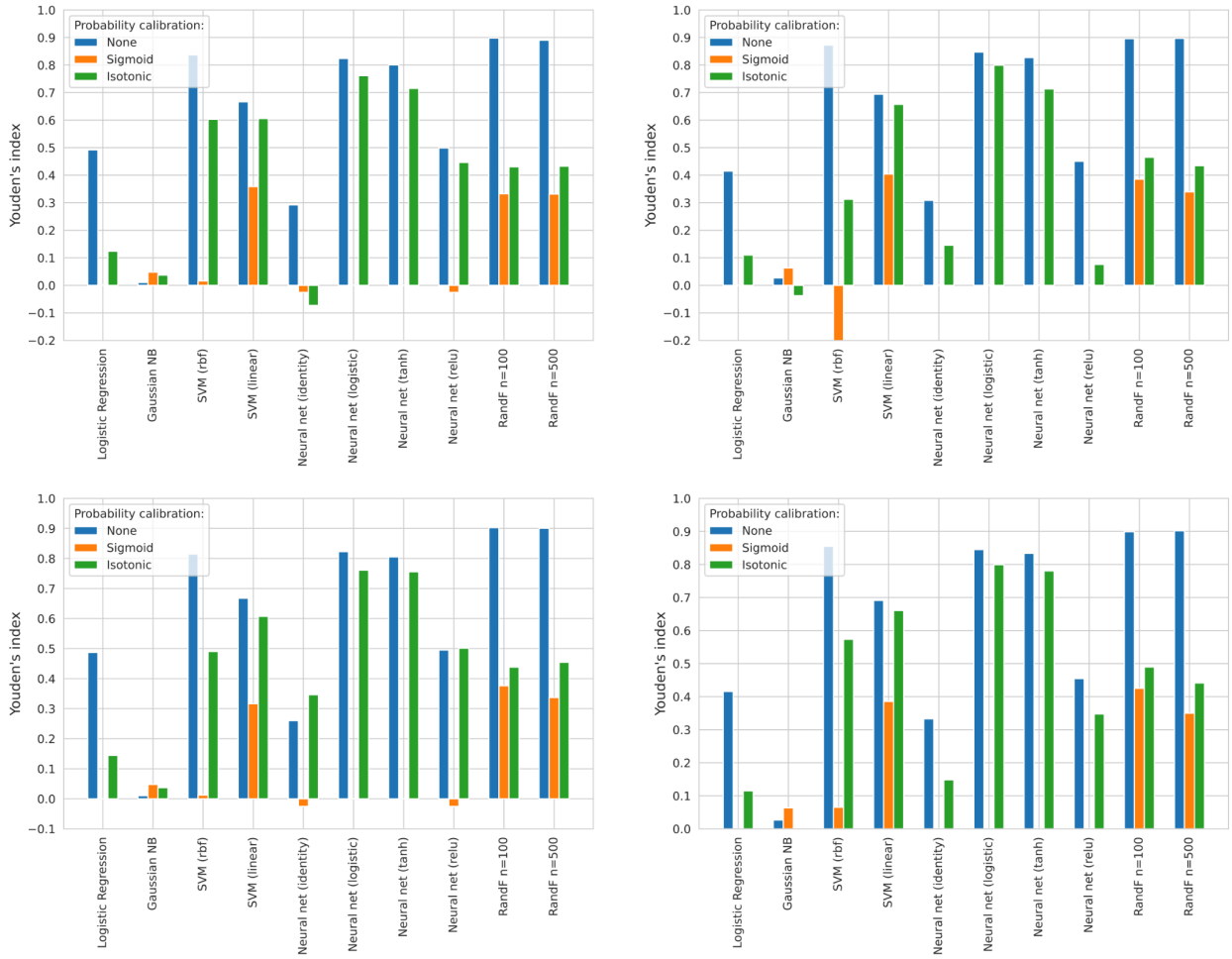


Figure S6.6: Probability calibration results.
Top: with E. coli. Bottom: dropped E. coli. Left: full melt. Right: short melt. Calibrating probabilities improves novelty detection for Naïve Bayes. None of the calibrated methods outperform the best results (SVM, Neural Net, RandF) as outlined in Figure 6.7 and Table 6.2.

Table S6.1. Overview of species

Experimentally obtained melt curves from:	**Synthesized melt curves from: (continued)**
Citrobacter koseri (taxid:545) Enterococcus faecium (taxid:1352) Escherichia coli (taxid:562) Haemophilus influenzae (taxid:727) Listeria monocytogenes (taxid:1639) Staphylococcus aureus (taxid:1280) Streptococcus agalactiae (taxid:1311) Streptococcus gallolyticus (taxid:315405) Streptococcus pneumoniae (taxid:1313) Streptococcus sanguinis (taxid:1305)	Mycobacterium fortuitum (taxid:1766) Mycobacterium gordonae (taxid:1778) Mycobacterium kansasii (taxid:1768) Mycoplasma hominis (taxid:2098) Mycoplasma pneumoniae (taxid:2104) Neisseria gonorrhoeae (taxid:485) Neisseria meningitidis (taxid:487) Oligella urethralis (taxid:90245) Pasteurella multocida (taxid:747) Propionibacterium acnes (taxid:1747) Proteus mirabilis (taxid:584) Proteus vulgaris (taxid:585) Pseudomonas aeruginosa (taxid:287) Salmonella enterica (taxid:28901) Serratia marcescens (taxid:615) Staphylococcus epidermidis (taxid:1282) Staphylococcus lugdunensis (taxid:28035) Staphylococcus saprophyticus (taxid:29385) Streptococcus pyogenes (taxid:1314) Treponema pallidum (taxid:160) Yersinia enterocolitica (taxid:630) Yersinia pestis (taxid:632) Yersinia pseudotuberculosis (taxid:633)
Synthesized melt curves from: Acinetobacter calcoaceticus (taxid:471) Acinetobacter baumannii (taxid:470) Aerococcus viridans (taxid:1377) Bacillus anthracis (taxid:1392) Bacillus cereus (taxid:1396) Bacillus subtilis (taxid:1423) Bacteroides fragilis (taxid:817) Bordetella parapertussis (taxid:519) Bordetella pertussis (taxid:520) Campylobacter jejuni (taxid:197) Chlamydia pneumoniae (taxid:83558) Chlamydia trachomatis (taxid:813) Citrobacter freundii (taxid:546) Clostridium difficile (taxid:1496) Clostridium perfringens (taxid:1502) Corynebacterium (taxid:1716) Coxiella burnetii (taxid:777) Enterobacter aerogenes (taxid:548) Enterococcus faecalis (taxid:1351) Enterococcus gallinarum (taxid:1353) Francisella philomiragia (taxid:28110) Francisella tularensis (taxid:263) Helicobacter pylori (taxid:210) Klebsiella pneumoniae (taxid:573) Legionella pneumophila (taxid:446) Micrococcus luteus (taxid:1270) Moraxella catarrhalis (taxid:480)	

Table S6.2. C-statistics for patristic distance regression using random forest with 500 trees.

	Full Melt Curve	**Short Melt Curve**
Raw Melt Curve	0.948	0.946
Signature Features	0.952	**0.956**

Chapter 7 Large Language Models Enable More Efficient Reporting of Hospital Quality Measures

Hospital quality measures are a vital component of a learning health system, yet they remain costly to report, can be statistically underpowered, and suffer from poor inter-rater reliability. Large language models (LLMs) have recently demonstrated impressive performance on healthcare-related tasks and offer a promising solution to provide accurate abstraction of the complete chart at scale. To evaluate this, we deployed an LLM-based system that ingests Fast Healthcare Interoperability Resources (FHIR) and outputs a completed Severe Sepsis and Septic Shock (SEP-1) abstraction. We tested this system on a sample of 100 manual SEP-1 abstractions which the University of California, San Diego Health reported to Centers for Medicare & Medicaid Services in 2022. The system achieved 90% agreement ($\kappa = 0.82$; 95% CI, 0.71 - 0.92) with manual abstractors on measure category assignment. Further, expert review of the 10 discordant cases identified 4 which were mistakes introduced by manual abstraction. This pilot suggests that LLMs with interoperable EHR data may perform abstractions for complex quality measures that are accurate, consistent, timely, and scalable.

INTRODUCTION

In 2022, the cost of quality reporting at a single acute care hospital was over $5 million USD and over 100,000 person-hours.[168] Within US physician practices, the annual cost of quality reporting was 785 hours per physician and over $15 billion.[169] Yet despite this massive financial and provider burden, quality measures are often assessed on a small denominator of patients, which limits their statistical validity and leads to delays in both measurement and actions to improve quality.[33,126,170,171] These limitations manifested in March 2020 when, amidst the COVID-19 pandemic, the Centers for Medicare & Medicaid Services (CMS) granted relief from

quality reporting "so the healthcare delivery system can direct its time and resources toward caring for patients."[33,172,173]

The measure for Severe Sepsis and Septic Shock (SEP-1) from CMS is a microcosm of the challenges in hospital quality reporting.[125] Previously a pay-for-reporting program, the measure will be included in the Hospital Value-Based Purchasing Program starting in 2026.[174] This has been met with opposition from various professional societies including the the Infectious Diseases Society of America (IDSA) and the American College of Emergency Physicians (ACEP) due, in part, to the measure's reporting burden and abstraction variability.[170,175] The composite "all or nothing" measure requires a complex 63-step abstraction process completed through manual chart review.[64,65] At the University of California, San Diego Health (UCSDH), abstraction involves initial determination by non-clinical analysts from an external vendor, followed by review from nurses in the Quality team and then final physician review. CMS requires monthly sampling of at least 20 patients meeting the measure inclusion criteria (e.g. inpatient, ICD-10-CM Principal or Other Diagnosis Code of Sepsis, Severe Sepsis, or Septic Shock).[125,176]

Recent work has demonstrated that large language models (LLMs) can achieve impressive performance on medical-related tasks, including human-level performance on standardized medical tests without task-specific fine-tuning.[68,177,178] Quality measurement is a complex task with a unique set of challenges based on both the requisite medical knowledge required to answer questions as well as the need to parse the temporal nature of the clinical course of diagnosis and treatment. In this work we test the hypothesis that LLMs and interoperable Electronic Health Record (EHR) data enable automated abstraction of complex quality measures. We use the SEP-1 measure as a case study due to its well-studied complexity.

METHODS

Study Design & Cohort

We developed and deployed an interoperable LLM system and tested it on a convenience sample of 100 random manual SEP-1 abstractions at UCSDH reported to CMS in 2022. This represented a 5 month reporting interval across two hospitals from three abstractors. All abstractors were from a single vendor and were non-clinical specialists trained to a standard operating procedure for SEP-1 abstraction. Our primary outcome was the measure category agreement (pass, fail, or out-of-measure) between the LLM system and current standard inclusive of human abstractors. We tested this agreement using Cohen's kappa with a two-sided test.[179] Disagreements between the LLM system and human abstractors were adjudicated by a board-certified emergency medicine and critical care medicine physician who chairs the UCSDH Sepsis Committee and are reported separately. We performed three independent trials of the LLM system to evaluate consistency. We additionally test the difference between the predicted and reported compliance rates using Cochran's Q test and report the 95% confidence interval of the predicted compliance rate using the Clopper-Pearson exact method. A *P* value less than 0.05 was interpreted as significant for all analyses. All statistical analyses were performed using Python version 3.11, the SciPy package version 1.10, and the statsmodel package version 0.13.5.[180,181] The UCSDH Institutional review board (IRB) approval was obtained with waiver of informed consent (#805726).

System Design

Our system architecture is shown in Figure 7.1. Data are retrieved as Fast Healthcare Interoperability Resources (FHIR) version R4.[182] Structured data are retrieved from the Patient,

Observation, ServiceRequest, Consent, Flag, and MedicationRequest FHIR resource types. Unstructured notes are gathered from the DocumentReference and Binary resources. Medication administration (MAR) information is not available in FHIR R4 and is retrieved from a proprietary Epic Application Programming Interface (API).

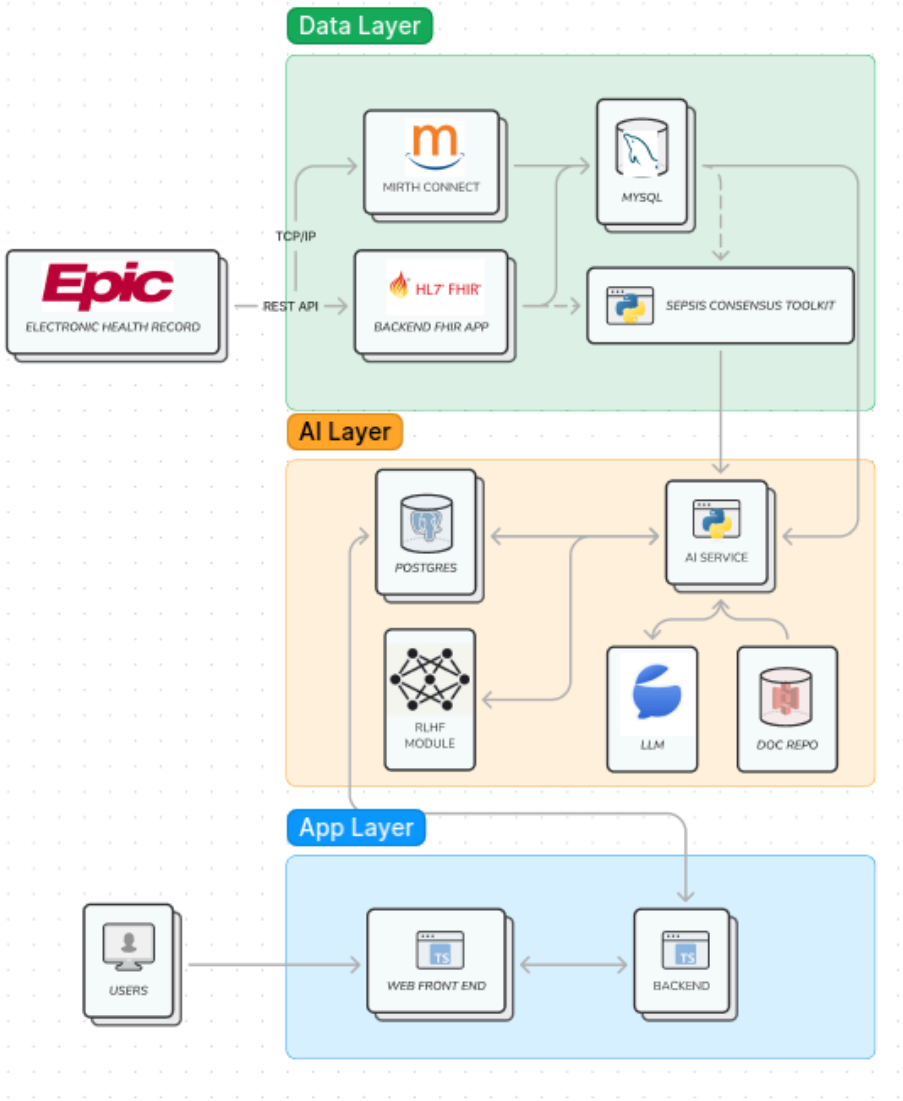


Figure 7.1: System architecture for automation of hospital quality measures.
The data layer (green) enables collection of EHR data through FHIR and computation of clinical criteria. Mirth Connect stores all encounters from Admission Discharge Transfer (ADT) messages. The Backend FHIR Application then queries encounter data and stores it in MySQL. The sepsis consensus toolkit applies standard rule-based criteria to the structured data to identify SIRS and organ failure events. The AI layer (orange) manages the LLM for abstraction. The app layer (blue) services the completed abstractions and collects human feedback.

The 63-step SEP-1 process flowchart was translated into Python and is hosted on a cloud-based virtual machine within a HIPAA-compliant Virtual Private Cloud (VPC) as described in Chapter 1.[34] The system proceeds through the measure and queries an LLM by performing Retrieval Augmented Generation (RAG) on a patient's clinical notes with CMS guidelines as prompt instructions when it reaches a step that requires information from unstructured data (see Table 7.2).[183] The system leverages a set of utilities (the sepsis consensus toolkit in Figure 7.1) to establish the presence of clinical criteria such as the Systemic Inflammatory Response Syndrome (SIRS) and the presence of organ failure from structured FHIR data. These criteria are then combined with LLM responses to select the appropriate allowable value for each SEP-1 data element.

The final output from the system is a completed SEP-1 abstraction including the measure category assignment. We service this result to users via a web application (Figure 7.2). Users can change data elements within the application's front end interface which triggers creation of a backend "human feedback" record.

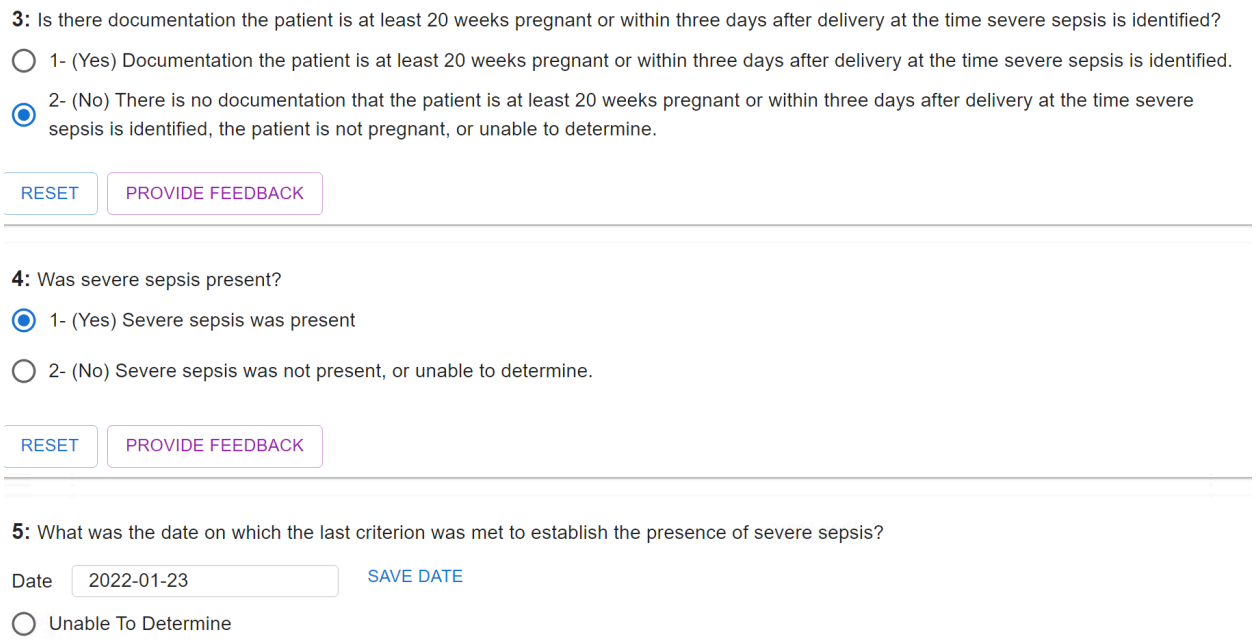


Figure 7.2. Web application front end for the system. Shown is sample output for the SEP-1 measure. Measure data elements are preloaded from a database in the AI layer. The user can change the element which creates a human feedback record.

LLM Implementation

LLM inference was performed using the open-source general-purpose SOLAR 10.7B model with 8-bit quantization and a context length of 8092 tokens.[184,185] We selected this model since it could be hosted on a single 24 GB GPU in a HIPAA-compliant environment and has relatively strong performance on standard benchmarks for its size.[186] No additional fine-tuning or prompt tuning was performed and all data remained within the VPC. We utilized chain-of-thoughts and few-shot prompting strategies with a temperature of 0.1.[69,187] This temperature is lower than the default value and was chosen to improve reproducibility of the system. LLM outputs were cast to JavaScript Object Notation (JSON) and invalid outputs were regenerated. The prompt template, all prompts, and few-shot examples are detailed in Supplementary Note 7.1 and Supplementary Note 7.2.

RAG was performed on the clinical notes by chunking the text into 1000 character segments with 50 characters of overlap, embedding the chunks and query with the Instructor model, calculating the cosine similarity between the query and the embeddings, and inserting the top 6 most similar chunks into the prompt.[188] Relevant code is available at https://github.com/aboussina/quallm.

RESULTS

Table 7.1 summarizes the patient characteristics and SEP-1 measure results based on standard reporting for the study cohort. Of the 100 cases, 62 did not qualify for the measure (B) per the SEP-1 guidelines. There were 38 patients only in the denominator (D), and 16 numerator compliant (E) for an overall compliance rate of 42.1%. Blood cultures were collected within the specified timeframe in 41/42 (97.6%) cases, timely antibiotics were administered in 45/50 (90%) cases, and in cases where initial hypotension was present, 30 mL/kg of crystalloid fluids were initiated and completely infused in 5/17 (29.4%) of cases.

Table 7.1: Study cohort demographics and SEP-1 measure results.

	SEP-1 Abstraction Cohort (N=100)
Age, median (IQR)	66.5 (53.50 - 74.25)
Sex	
Female	37 (37%)
Male	63 (63%)
Race	
White	49 (49%)
Black or African American	8 (8%)
American Indian or Alaska Native	0
Asian or Pacific Islander	10 (10%)
Other Race or Mixed Race	33 (33%)
Major Measure Elements	
Blood Culture Collection	41/42 (97.6%)
Initial Lactate Level Collection	36/39 (92.3%)
Broad Spectrum or Other Antibiotic Administration	45/50 (90%)
Repeat Lactate Level Collection	22/24 (91.7%)
Crystalloid Fluid Administration	5/17 (29.4%)
Measure Category Assignment	
B (not in measure population)	62 (62%)
D (in measure population)	22 (22%)
E (in numerator population)	16 (16%)
Compliance Rate	16/38 (42.1%)

The LLM system agreement with standard reporting inclusive of manual abstraction is detailed in Table 7.2. We observe that the LLM system generated identical measure category

assignments across all three trials and achieved 90% agreement ($\kappa = 0.82$; 95% CI, 0.71 - 0.92; $p < .001$) with manual abstractors. Physician adjudication of the 10 discordant cases is described in Table 3. For 4/10 cases, the reviewing physician determined that the LLM system was more accurate than the human abstractor. The system-computed compliance rate for the cohort 19/39 (48.7%; 95% CI, 32.4% - 65.2%) was not significantly different from the compliance rate reported to CMS (42.1%; P = 0.37). An example abstraction from the LLM system is shown in Supplementary Note 3 and example errors from the LLM are shown in Supplementary Note 7.4.

Table 7.2: LLM system agreement with manual abstraction for SEP-1.

SEP-1 Question	**N (%) of Abstractions where System Answer Resulted in Agreement with Manual Category Assignment**	**Data Element Distribution (N; %) from Manual Abstraction**
Was the patient received as a transfer from an inpatient, outpatient or emergency/observation department of an outside hospital or from an ambulatory surgery center?*	99/100 (99%)	Y (10; 10%)
		N (90; 90%)
During this hospital stay, was the patient enrolled in a clinical trial in which patients with the same condition as the measure set were being studied?*	100%	Y (0; 0%)
		N (100; 100%)
Is there documentation the patient is at least 20 weeks pregnant or within three days after delivery at the time severe sepsis is identified?*	100%	Y (0; 0%)
		N (100; 100%)
Was severe sepsis present?*	98/100 (98%)	Y (49; 53.8%)
		N (42; 46.2%)
When was the last criterion met to establish the presence of severe sepsis?*	97/100 (97%)	-
Is there documentation that the patient or authorized patient advocate refused either a blood draw, IV fluid administration, or IV antibiotic administration within the specified time frame?*	100%	Y (3; 5.4%)
		N (53; 94.6%)
Is there physician/APN/PA documentation of comfort measures only, palliative care, or another inclusion term before or within six hours after the presentation of severe sepsis?*	99/100 (99%)	Y (5; 8.9%)
		N (51; 91.1%)

Table 7.2: LLM system agreement with manual abstraction for SEP-1 (Continued).

SEP-1 Question	N (%) of Abstractions where System Answer Resulted in Agreement with Manual Category Assignment	Data Element Distribution (N; %) from Manual Abstraction
What was the patient's discharge disposition on the day of discharge?*	100%	Home (62; 62%)
		Hospice - Home (2; 2%)
		Hospice - Health Care Facility (3; 3%)
		Acute Care Facility (2; 2%)
		Other Health Care Facility (16; 16%)
		Expired (10; 10%)
		Left Against Medical Advice/AMA (5; 5%)
		Not Documented or Unable to Determine (0; 0%)
Was a broad spectrum or other antibiotic administered within the specified time frame?	100%	Y (45; 90%)
		N (5; 10%)
What was the earliest datetime on which an antibiotic was started within the specified time frame?	100%	-
Was a blood culture collected in the specified time frame?	100%	Y (41; 97.6%)
		N (1; 2.4%)
When was the blood culture collected?	100%	-
Is there documentation supporting an acceptable delay in collecting a blood culture?*	99/100 (99%)	Y (1; 33.3%)
		N (2; 66.6%)
Was an initial lactate level drawn within the specified time frame?	100%	Y (36; 92.3%)
		N (3; 7.7%)
What was the datetime on which the initial lactate level was drawn?	100%	-
What was the initial lactate level result? (<= 2 mmol/L, between 2 mmol/L and 4 mmol/L, or >= 4 mmol/L)	100%	<= 2 mmol/L (12; 33.3%)
		>2 mmol/L and < 4 mmol/L (17; 47.2%)
		>= 4 mmol/L (7; 19.4%)
Was a repeat lactate level drawn within the specified time window?	100%	Y (22; 91.7%)
		N (2; 8.3%)
What was the earliest datetime on which the repeat lactate level was drawn?	100%	-
Was initial hypotension present within the specified time frame?	100%	Y (12; 35.3%)
		N (22; 64.7%)
When was initial hypotension present six hours prior to or within six hours following Severe Sepsis Presentation Date and Time?	100%	-

Table 7.2: LLM system agreement with manual abstraction for SEP-1 (Continued).

Were crystalloid fluids initiated within the specified time frame AND completely infused based on the target ordered volume?	99/100 (99%)	Y (5; 29.4%)
		N (12; 70.6%)
What was the earliest datetime on which crystalloid fluids were initiated within the specified time frame?	100%	-
Is there documentation of the presence of septic shock?*	100%	Y (7; 26.9%)
		N (19; 73.1%)
What was the datetime on which the last criterion was met to establish the presence of septic shock?*	100%	-
Is there documentation that the patient or authorized patient advocate refused either a blood draw, IV fluid administration, or vasopressor administration before or within six hours after the Septic Shock Presentation Time?*	100%	Y (0; 0%)
		N (7; 100%)
Is there physician/APN/PA documentation of comfort measures only, palliative care, or another inclusion term before or within six hours after the presentation of septic shock?*	100%	Y (0; 0%)
		N (7; 100%)
Was persistent hypotension or new onset of hypotension present within one hour of when the target ordered volume of crystalloid fluids was completely infused?	99/100 (99%)	Y (1; 16.7%)
		N (5; 83.3%)
Was an intravenous or intraosseous vasopressor administered within the specified time frame?	100%	Y (1; 100%)
		N (0; 0%)
What was the datetime on which an intravenous or intraosseous vasopressor was administered within the specified time frame?	100%	-
Was a repeat volume status and tissue perfusion assessment performed within the specified time frame?*	100%	Y (3; 100%)
		N (0; 0%)
What datetime was a repeat volume status and tissue perfusion assessment performed?*	100%	-
Final Measure Category Assignment	**90/100 (90%)****	B (62; 62%)
		D (22; 22%)
		E (16; 16%)

* Indicates that an LLM query was utilized for the question.

** Because SEP-1 is an all-or-nothing measure, a mismatch of even a single criterion can result in a different final measure category assignment.

Table 7.3: Physician review of discrepant cases.

Root Cause of Disagreement	**Number of Cases**	**Physician Adjudication**
The patient had Chronic Kidney Disease with a creatinine baseline of (2-3). The system identified a creatinine value elevated >0.5 above baseline as evidence of organ failure. The abstractor did not identify this as a sign of organ failure.	1	LLM system is more accurate
INR[a] organ failure missed by abstractor.	1	LLM system is more accurate
Difference in documentation of infection time.	2	LLM system is more accurate
	1	Abstractor is more accurate
LLM is too sensitive to the presence of palliative care.	1	Abstractor is more accurate
LLM is too sensitive to an acceptable delay in blood cultures.	1	Abstractor is more accurate
Missing data fields (e.g. arterial line blood pressure measurements)[b]	3	Abstractor is more accurate

[a]INR: International Normalized Ratio

[b]Fields were incorporated into the system after physician adjudication.

DISCUSSION

Prior work has advocated for reducing the number of quality measures or transitioning to simpler electronic clinical quality measures (eCQM) which, historically, have had challenges in matching robust performance.[189] This work offers an alternative lens: provider relief from reporting burden may be achieved through better tools that appropriately capture case complexity and provide timely feedback. To that end, we demonstrated that LLMs with interoperable EHR data can accurately perform abstraction of the SEP-1 quality measure, and furthermore that open-source LLMs running on consumer-grade hardware are sufficiently capable. To our knowledge, this represents the first work to explore the capabilities of LLMs for hospital quality reporting. The SEP-1 measure is one of the most complex quality measures, which makes it a

suitable stress test for quality measurement writ large. The availability of previously-reported abstractions and the ability to collect user feedback from our system interface also offers opportunities for improved performance through supervised fine-tuning and reinforcement learning from human feedback.[190,191]

This approach is promising since evaluating a measure across a cohort of patients can be easily scaled beyond standard sampling for robust statistical findings. Within our study cohort, only 38 patients were included in the measure after 5 months of reporting. It is intractable to meaningfully identify quality improvement opportunities from this data. However, this LLM system affords a feasible approach that may enable SEP-1 abstraction to scale to every patient with an encounter during the reporting period.

Equally important, these findings can be generated shortly after patient discharge to shorten the time to incorporate change inclusive of process improvement within a learning health system. Timely auditing and feedback have been shown to improve measure compliance.[192,193] Unfortunately, quality measures are often prepared at either a monthly or quarterly resolution. For SEP-1 reporting at UCSDH, only cases from two months ago are prepared in a given month which precludes use of the measure to proactively target systemic issues.

Artificial intelligence for quality reporting also offers a promising avenue to reduce the variability inherent in human chart review. The National Quality Forum (NQF) takes the position that a performance measure cannot be scientifically acceptable if its data elements have poor inter-rater reliability.[194] They recommend that measure developers avoid data elements with a kappa statistic less than 0.41. Yet Rhee et al. demonstrated that the SEP-1 pass rate had a kappa of 0.39 across 3 reviewers at 3 hospitals and that abstractors agreed on time zero in only 36% of cases.[126] In this study, we observed a few examples of human error that could contribute to poor

reliability (Table 7.3). In two cases, there was clear documentation of suspected infection that was missed by reviewers resulting in a different time zero. In one case, the presence of organ failure due to an international normalized ratio (INR) value > 1.5 was missed. We also observed some clear errors and hallucinations from the LLM resulting in incorrect abstractions (Supplementary Note 7.4). In one case, the LLM inappropriately conflated palliative radiation therapy with comfort measures only. In another, the LLM inferred with insufficient evidence that an infection was being treated prior to the presentation of severe sepsis. With improved grounding and alignment, the ability to apply the same criteria, prompts, and model to a consistent set of interoperable data elements holds promise in improving intra-system and inter-system reliability.

Our study has several limitations. First, our study cohort is a small convenience sample of 100 cases across two hospitals. This represents five months of abstraction at our health system and highlights the limited scope of manual review in the current state. Second, while we used interoperable data standards wherever possible, the system is reliant on a proprietary API for MAR information and would require modification to support other EHR vendors. Third, we did not explore performance across different LLMs. However, Since we used a mid-size general-purpose LLM that was not finetuned on our data or adapted to the medical domain, we expect our results are generalizable. Finally, while we provided a web application front-end to our system, we did not explore the human-computer interaction. Future work is needed to evaluate whether human abstractors equipped with this tool can achieve greater accuracy, reliability, and efficiency. Future work will also evaluate whether automated abstraction generation can save clinician reviewer time by presenting clear evidence and enabling rapid rework.

Ultimately, the evolution of quality metrics through the adoption of interoperability standards and artificial intelligence offers a promising avenue to alleviate the clinical workload associated with manual chart reviews, thereby reallocating precious time in healthcare to enhance care quality initiatives.

ACKNOWLEDGEMENTS

Chapter 7, in full, has been submitted for publication of the material as it may appear in NEJM AI as “Large Language Models Enable More Efficient Reporting of Hospital Quality Measure”. Boussina, A., Krishnamoorthy, R., …, & Nemati, S. The dissertation author was the primary investigator and author of this paper.

SUPPLEMENTARY MATERIAL

Supplementary Note 7.1. Prompt Template.

User:
You are a CMS quality abstractor. Your task is to review patient's medical note and answer the given sepsis compliance question following the abstraction instructions.
Generate clear rationale to your answer by thinking step-by-step.

ABSTRACTION INSTRUCTIONS: {guidelines}
MEDICAL NOTE: {notes}.
QUESTION: {question}
OUTPUT FORMAT: Return the answer as a JSON object following the format, {{"rationale": str, "option": str}}.

Assistant:

Supplementary Note 7.2.

Table S7.1. Prompts for SEP-1 questions.

Measure Data Element	**Prompt**	**Options**	**Examples**
Clinical Trial	During this hospital stay, was the patient enrolled in a clinical for sepsis?	Y = Yes N = No	-
Pregnant 20 Weeks Through Day 3 Post-delivery	Is there documentation the patient is at least 20 weeks pregnant or within three days after delivery?	1 = Yes 2 = No	-
Severe Sepsis Present	Does the note clearly specify if this was prescribed to treat an infection?	Y = Yes N = No	-
	In this note, is there explicit positive qualifier for an infection? Do not infer or take a likely guess. Answer Y only if the infectious condition is clearly stated in the note. Do not use documentation of viral, fungal, or parasitic infections. The following is a list of conditions commonly associated with severe sepsis that are considered infections (this is not an all-inclusive list.) EXAMPLES: {Examples} EXAMPLES THAT AREN'T INFECTIONS: Fever Diarrhea POSITIVE QUALIFIERS: Possible Rule out (r/o) Suspected Likely Probable Differential diagnosis Suspicious for Concern for Suggestive of Presumed NEGATIVE QUALIFIERS: Impending Unlikely Doubt Ruled out Less likely Questionable	Y = Yes N = No	● Abscess ● Acute abdomen ● Acute abdominal infection ● Blood stream ● catheter infection ● Bone/joint infection ● C. difficile (C-diff) ● Chronic obstructive pulmonary disease (COPD) acute exacerbation ● Endocarditis ● Gangrene ● Implantable device infection ● Infection ● Infectious ● Meningitis ● Necrosis ● Necrotic/ischemic/infarcted bowel ● Pelvic Inflammatory Disease ● Perforated bowel ● Pneumonia ● Empyema ● Purulence/pus ● Sepsis ● Septic ● Skin/soft tissue infection ● Suspect infection, source unknown ● Urosepsis ● Urinary tract infection ● Wound infection

Table S7.1. Prompts for SEP-1 questions (Continued).

Measure Data Element	**Prompt**	**Options**	**Examples**
Severe Sepsis Present	Is there explicit documentation of a positive qualifier for severe sepsis or septic shock during this admission? POSITIVE QUALIFIERS: Possible Rule out (r/o) Suspected Likely Probable Differential diagnosis Suspicious for Concern for Suggestive of Presumed	Y = Yes N = No	Differential diagnosis includes septic shock -> Y During last admission, s/o severe sepsis -> N Low concern for severe sepsis -> N
Administrative Contraindication to Care, Severe Sepsis	Is there explicit documentation that the patient or authorized patient advocate refused either a blood draw, IV fluid administration, or IV antibiotic administration?	1 = Yes 2 = No	-
Directive for Comfort Care or Palliative Care, Severe Sepsis	Is there explicit documentation of comfort measures only, palliative care, or another inclusion term?	1 = Yes 2 = No	Left hip palliative CXR -> 2 The goal of care will be palliative -> 1 DNR/Full Care -> 2

Table S7.1. Prompts for SEP-1 questions (Continued).

Measure Data Element	Prompt	Options	Examples
Discharge Disposition	Map the discharge disposition to the closest one of the options below.	1 = Home 2 = Hospice - Home 3 = Hospice - Health Care Facility 4 = Acute Care Facility 5 = Other Health Care Facility 6 = Expired 7 = Left Against Medical Advice/AMA	HOME EXAMPLES (1): Home Healthcare IP Admit Related -> 1 Court/Law Enforcement (includes detention facilities, jails, and prison) -> 1 Home with Home Health Services -> 1 Outpatient Services including outpatient procedures at another hospital, Outpatient Chemical Dependency Programs and Partial Hospitalization -> 1 Shelter -> 1 HOSPICE - HOME EXAMPLES (2): Hospice in the home (or other 'Home' setting as above in Value 1) -> 2 HOSPICE HEALTH CARE FACILITY EXAMPLES (3): Hospice - General Inpatient and Respite -> 3 Hospice - Residential and Skilled Facilities -> 3 Hospice - Other Health Care Facilities -> 3 ACUTE CARE FACILITY EXAMPLES (4): Acute Short Term General and Critical Access Hospitals -> 4 Cancer and Childrens Hospitals -> 4 Department of Defense and Veteran's Administration Hospitals -> 4 OTHER HEALTH CARE FACILITY EXAMPLES (5): Extended or Intermediate Care Facility (ECF/ICF) -> 5 Long Term Acute Care Hospital (LTACH) -> 5 Nursing Home or Facility including Veteran's Administration Nursing Facility -> 5 Psychiatric Hospital or Psychiatric Unit of a Hospital -> 5 Rehabilitation Facility including, but not limited to: Inpatient Rehabilitation Facility/Hospital, Rehabilitation Unit of a Hospital, Chemical Dependency/Alcohol Rehabilitation Facility -> 5 Skilled Nursing Facility (SNF), Sub-Acute Care or Swing Bed -> 5 Transitional Care Unit (TCU) -> 5

			Veterans Home -> 5

Table S7.1. Prompts for SEP-1 questions (Continued).

Measure Data Element	**Prompt**	**Options**	**Examples**
Blood Culture Collection Acceptable Delay	Is there explicit documentation from the notetaker supporting an acceptable delay in collecting a blood culture?	1 = Yes 2 = No	Transfer to ICU, start antibiotics, f/o with cultures -> 1 Patient condition worsening, IV Vanco ordered stat, blood and urine cultures ordered, awaiting CXR -> 1 Patient's deteriorating condition concern for rapidly advancing infection, starting IV antibiotics now, lab on way to collect blood cultures -> 1 BC drawn? Yes. If no, explain: N/A -> 2
Septic Shock Present	Is there explicit documentation from the notetaker of a positive qualifier for septic shock during this admission? POSITIVE QUALIFIERS: Possible Rule out (r/o) Suspected Likely Probable Differential diagnosis Suspicious for Concern for Suggestive of Presumed	Y = Yes N = No	Differential diagnosis includes septic shock -> Y Prior history of septic shock and concern for sepsis -> N Previous documentation of septic shock -> N
Administrative Contraindication to Care, Septic Shock	Is there documentation that the patient or authorized patient advocate refused either a blood draw, IV fluid administration, or vasopressor administration?	1 = Yes 2 = No	-
Directive for Comfort Care or Palliative Care, Septic Shock	Is there physician/APN/PA documentation of comfort measures only, palliative care, or another inclusion term?	1 = Yes 2 = No	Left hip palliative CXR -> 2 The goal of care will be palliative -> 1 DNR/Full Care -> 2
Repeat Volume Status and Tissue Assessment Performed	Was a repeat volume status and tissue perfusion assessment performed according to the notes?	1 = Yes 2 = No	-

Supplementary Note 7.3.

Table S7.2. Example abstraction (deidentified).

Measure Data Element	**LLM Rationale**	**Answer**
Clinical Trial	To determine if the patient was enrolled in a clinical trial for sepsis during their hospital stay, we need to find clear documentation indicating this in the medical notes. While the patient's diagnosis includes sepsis and there are mentions of treatment related to sepsis, there is no specific mention or documentation about their enrollment in a clinical trial for sepsis. Therefore, based on the provided information, it cannot be confirmed that they were part of such a trial.	N
Pregnant 20 Weeks Through Day 3 Post-delivery	In the given medical note, there is no direct information about the patient's pregnancy status or delivery date. However, we can infer that she had a hysterectomy mentioned in 'Past Surgical History: Past Surgical History: Procedure Laterality Date ? HYSTERECTOMY'. This indicates that the patient does not have a uterus anymore and therefore cannot be pregnant or within three days after delivery. Additionally, there is no documentation suggesting she was pregnant at the time of severe sepsis identification. Thus, we can apply the abstraction instruction's condition 'There is documentation indicating the patient had a partial or complete hysterectomy.', which leads to selecting option 2 (No).	N
Severe Sepsis Present	To determine if the note specifies treatment for an infection, we can analyze its contents. The note mentions 'Indications: Skin / Soft Tissue Infection'. This clearly indicates that the medication is intended to treat a specific type of infection in the patient's skin or soft tissues. Therefore, based on this information, it can be concluded that the prescription is indeed for treating an infection.	1
Administrative Contraindication to Care, Severe Sepsis	To determine if there is explicit documentation regarding the patient's refusal of blood draw, IV fluid administration, or IV antibiotic administration, we need to carefully review the medical note. The note mentions that two blood pressures were documented after crystalloids completed (indicating IV fluids were administered), and initial antibiotics given. However, there is no direct mention of the patient refusing these interventions. Thus, while it can be inferred that they cooperated with these procedures, we cannot definitively conclude without explicit documentation. Therefore, based on the provided information, the answer would be:	2
Directive for Comfort Care or Palliative Care, Severe Sepsis	N/A (no notes available)	2
Discharge Disposition	To determine the discharge disposition for this patient, we first analyze the given medical note which is 'Home Healthcare IP Admit Related'. This indicates that the patient was admitted to a hospital or healthcare facility for home healthcare-related services. According to the abstraction instructions provided, Home Health Services and related admissions fall under the 'Home' category. Therefore, based on this information and the instruction mapping, we can conclude that the discharge disposition is closest to example 1 in the given examples list.	1
Blood Culture Collection Acceptable Delay	N/A	N/A
Septic Shock Present	N/A (no notes available)	2
Measure Category Assignment	N/A	**E**

Supplementary Note 7.4.

Table S7.3. Example Errors & Hallucinations.

Category	**LLM Response**	**Description of Errors**
LLM is too sensitive to the presence of palliative care.	To determine if there is explicit documentation for comfort measures only, palliative care, or another inclusion term, we need to examine the medical note based on the given abstraction instructions. The note mentions 'Left hip palliative XRT' which refers to left hip palliative radiotherapy treatment. Palliative care is one of the acceptable terms for indicating comfort measures only or palliative care. This documentation falls under an accepted context as it represents a recommendation for palliative care in relation to the left hip condition. Therefore, based on this evidence from the medical note, there is explicit documentation of palliative care which can be considered equivalent to comfort measures only in this case.	Palliative XRT can not be considered equivalent to comfort measures only.
LLM is too sensitive to an acceptable delay in blood cultures.	To determine if there is documentation supporting an acceptable delay in collecting a blood culture, we need to identify situations mentioned in the abstraction instructions. In this case, the patient presented with acute worsening of right-sided upper abdominal pain and was admitted for multiple conditions including Multiple Sclerosis, COVID pneumonia, and PE on lovenox. The notes mention that Vanc/Aztreonam/Flagyl were given due to PCN allergy before blood cultures were drawn. This indicates that the patient had an infection within 24 hours before severe sepsis was identified, and a blood culture was drawn after the antibiotic dose was started. Therefore, this falls under the 'Antibiotics Started in Hospital' situation mentioned in the instructions	1. COVID pneumonia is from a prior visit. 2. The note doesn't specify that antibiotics were given prior to cultures (hallucination). 3. The note was written after the severe sepsis presentation time, but the LLM infers that there was an infection within 24 hours prior.

CONCLUSIONS AND FUTURE WORK

SUMMARY OF CONTRIBUTIONS

In Chapter 1, we presented the design of an interoperable system to enable deployment of arbitrarily complex deep learning models onto the EHR. This system formed the foundation of the subsequent chapters and addressed entrenched challenges for the deployment of predictive analytics into clinical environments. Concretely, we described how FHIR and HL7v2 can be harnessed to ensure generalizability across healthcare systems and EHR vendors. We detailed how cloud computing could be employed to ensure reliability, security, and scalability. And we showcased how post-implementation monitoring can enforce continued performance of the system. This work goes beyond prior publications in terms of the volume of patients (the system runs on all patients at UCSDH across all care units), the diversity of data elements (the system collects laboratory results, vital-signs, demographics, problem-list items, procedures, and clinical notes), and its clinical relevance (the system hosts a deep learning model that is used in clinical practice).

We further described a playbook for safe deployment of models into clinical practice (silent mode chart reviews, refinement of inclusion/exclusion criteria, EPIS deployment framework) and, in Chapter 2, presented results from a prospective clinical evaluation of a deep learning model for sepsis prediction. We showed that the deployment of this model in two EDs was significantly associated with a 17% reduction in sepsis-related mortality, reduced organ failure, and a 10% improvement in quality of care as measured by compliance with the SEP-1 bundle. This represents the first study to demonstrate improvement in patient-centered outcomes attributable to the deployment of a deep-learning-based sepsis prediction model. We then

analyzed provider perceptions of this model in Chapter 3 and showed that nearly 80% of respondents had a positive or neutral view of the alert system.

In Chapters 4-6, we presented numerous innovations that hold promise in improving the actionability of this AI-enabled digital health system beyond early prediction. We first developed and validated novel sepsis phenotypes in Chapter 4 using unsupervised and representation learning. These phenotypes were prognostic, dynamic, and significantly associated with clinical interventions. These findings may support more personalized interventions targeted to a patient's specific sepsis phenotype.

We then explored the role of data deserts on the recognition of sepsis in Chapter 5. We presented evidence that expedited laboratory assessment may reveal organ dysfunction that could improve the timeliness of recognition and treatment of sepsis. We implemented a prototype system (AIVIS) that recommends order sets on high risk patients when there is insufficient data. We then investigated a novel bioassay for pathogen profiling and novelty detection in Chapter 6 that serves as an exemplar for potential recommendations from the AIVIS system. We show that signatures from dHRM curves can enable rapid pathogen identification and reveal emerging pathogens.

We conclude in Chapter 7 by formulating a novel method for assessing process measures at scale and in near realtime. Specifically, we leverage the platform from Chapter 1 and modern advances in LLMs to automate generation of the SEP-1 abstraction. This provides a mechanism to continuously evaluate the impact of digital health interventions on quality of care. It may further enable improved auditing and feedback to support a learning health system and reduce the clinical workload associated with manual chart reviews, thereby reallocating precious time in healthcare to enhance care quality initiatives.

FUTURE DIRECTIONS

This thesis has focused on applications for sepsis management, but many of the concepts and methods presented are agnostic to the disease area. Future work may deploy predictive models onto the analytics platform that cover a panoply of applications including inpatient deterioration, respiratory failure, acute kidney injury, and readmission, among others. In a similar vein, for various disease states the phenotypes can be characterized and the relevant quality measures can be automated and surveilled.

Further, while we performed a robust causal impact analysis to determine the salubrious impact of our sepsis AI, additional randomized controlled trials are needed to achieve the gold-standard of clinical evidence. Future research may conduct randomized trials of COMPOSER in additional care settings including inpatient wards, post-operative units, and ICUs, as well as across additional healthcare systems.

Finally, while in this work we presented evidence supporting our prototype development of AIVIS, additional studies may further characterize the utility of this system in clinical environments. For example, Chapter 6 presented a promising bioassay for pathogen profiling that may enable earlier identification of pathogens and targeted antimicrobial therapies if deployed in a prospective clinical study. The potential of AIVIS is to precisely target the patients who would most benefit from such a measurement. More broadly, there are many steps in the journey of a critical care patient where AIVIS could support diagnosis, treatment, and recovery through recommendation of novel interventions.

CONCLUSION

The research presented in this dissertation addresses significant challenges in the deployment of AI-enabled digital health interventions in critical care settings. By developing a scalable and robust predictive analytics platform, clinically validating a deep learning model for sepsis prediction, and enhancing the platform with novel bioassays and automated quality measure reporting, this work demonstrates the potential of AI to transform sepsis management and improve patient outcomes. Future research directions outlined in this dissertation provide a roadmap for further advancing AI-enabled healthcare interventions, ensuring their broader applicability and sustained impact in diverse clinical settings.

REFERENCES

1. Dantes, R. B. (2023). Sepsis program activities in acute care hospitals—National healthcare safety network, United States, 2022. MMWR. Morbidity and Mortality Weekly Report, 72.

2. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA. 2016;315(8):762–774. doi: 10.1001/jama.2016.0288

3. Singer M., Deutschman C. S., Seymour C. W., Shankar-Hari M., Annane D., Bauer M., & Angus D. C. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). Jama, 315(8), 801–810.

4. Kaukonen, K. M., Bailey, M., Suzuki, S., Pilcher, D., & Bellomo, R. (2014). Mortality related to severe sepsis and septic shock among critically ill patients in Australia and New Zealand, 2000-2012. JAMA, 311(13), 1308-1316.

5. Mayr, F. B., Yende, S., Linde-Zwirble, W. T., Peck-Palmer, O. M., Barnato, A. E., Weissfeld, L. A., & Angus, D. C. (2010). Infection rate and acute organ dysfunction risk as explanations for racial differences in severe sepsis. JAMA, 303(24), 2495–2503. https://doi.org/10.1001/jama.2010.851

6. Rhodes A., Phillips G., Beale R., Cecconi M., Chiche J. D., De Backer D., … & Levy M. (2015). The surviving sepsis campaign bundles and outcome: results from the international multicentre prevalence study on sepsis (the IMPreSS study). Intensive care medicine, 41(9), 1620–1628.

7. Sterling S. A., Miller W. R., Pryor J., Puskarich M. A., & Jones A. E. (2015). The impact of timing of antibiotics on outcomes in severe sepsis and septic shock: a systematic review and meta-analysis. Critical care medicine, 43(9), 1907.

8. Ferrer R., Martin-Loeches I., Phillips G., Osborn T. M., Townsend S., Dellinger R. P., … & Levy M. M. (2014). Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. Critical care medicine, 42(8), 1749–1755

9. Seymour CW, Gesten F, Prescott HC, et al. Time to Treatment and Mortality during Mandated Emergency Care for Sepsis. N Engl J Med. 2017;376(23):2235-2244. doi:10.1056/NEJMoa1703058

10. Baghdadi JD, Brook RH, Uslan DZ, et al. Association of a Care Bundle for Early Sepsis Management With Mortality Among Patients With Hospital-Onset or Community-Onset Sepsis. JAMA Intern Med. 2020;180(5):707-716. doi:10.1001/jamainternmed.2020.0183

11. Ferrer, R., Martin-Loeches, I., Phillips, G., Osborn, T. M., Townsend, S., Dellinger, R. P., … & Artigas, A. (2014). Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. Critical Care Medicine, 42(8), 1749-1755.

12. Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., ... & Cheang, M. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical Care Medicine, 34(6), 1589-1596.

13. Filbin, M. R., Lynch, J., Gillingham, T. D., Thorsen, J. E., Pasakarnis, C. L., Nepal, S., Matsushima, M., Rhee, C., Heldt, T., & Reisner, A. T. (2018). Presenting Symptoms Independently Predict Mortality in Septic Shock: Importance of a Previously Unmeasured Confounder. Critical care medicine, 46(10), 1592–1599. https://doi.org/10.1097/CCM.0000000000003260

14. Pines, J. M., Hilton, J. A., Weber, E. J., Alkemade, A. J., Al Shabanah, H., Anderson, P. D., Bernhard, M., Bertini, A., Gries, A., Ferrandiz, S., Kumar, V. A., Harjola, V. P., Hogan, B., Madsen, B., Mason, S., Ohlén, G., Rainer, T., Rathlev, N., Revue, E., Richardson, D., … Schull, M. J. (2011). International perspectives on emergency department crowding. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine, 18(12), 1358–1370. https://doi.org/10.1111/j.1553-2712.2011.01235.x

15. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, Campbell E, Bates DW. Grand challenges in clinical decision support. J Biomed Inform. 2008 Apr;41(2):387-92. doi: 10.1016/j.jbi.2007.09.003. Epub 2007 Sep 21. PMID: 18029232; PMCID: PMC2660274.

16. Kaukonen, K. M., Bailey, M., Pilcher, D., Cooper, D. J., & Bellomo, R. (2015). Systemic inflammatory response syndrome criteria in defining severe sepsis. The New England journal of medicine, 372(17), 1629–1638. https://doi.org/10.1056/NEJMoa1415236

17. Griffiths, P., Maruotti, A., Recio Saucedo, A., Redfern, O. C., Ball, J. E., Briggs, J., Dall'Ora, C., Schmidt, P. E., Smith, G. B., & Missed Care Study Group (2019). Nurse staffing, nursing assistants and hospital mortality: retrospective longitudinal cohort study. BMJ quality & safety, 28(8), 609–617. https://doi.org/10.1136/bmjqs-2018-008043

18. Pines, J. M., Hollander, J. E., Localio, A. R., & Metlay, J. P. (2006). The association between emergency department crowding and hospital performance on antibiotic timing for pneumonia and percutaneous intervention for myocardial infarction. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine, 13(8), 873–878. https://doi.org/10.1197/j.aem.2006.03.568

19. Rajkomar, A., Dean, J., & Kohane, I. (2018). Machine Learning in Medicine. New England Journal of Medicine, 378, 1347-1358.

20. Shashikumar, S.P., Wardi, G., Malhotra, A. et al. Artificial intelligence sepsis prediction algorithm learns to say “I don’t know”. npj Digit. Med. 4, 134 (2021). https://doi.org/10.1038/s41746-021-00504-6

21. Adams, R., Henry, K.E., Sridharan, A. et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 28, 1455–1460 (2022). https://doi.org/10.1038/s41591-022-01894-0

22. Shashikumar SP, Wardi G, Paul P, Carlile M, Brenner LN, Hibbert KA, North CM, Mukerji SS, Robbins GK, Shao YP, Westover MB, Nemati S, Malhotra A. Development and Prospective Validation of a Deep Learning Algorithm for Predicting Need for Mechanical Ventilation. Chest. 2021 Jun;159(6):2264-2273. doi: 10.1016/j.chest.2020.12.009. Epub 2020 Dec 17. PMID: 33345948; PMCID: PMC8027289.

23. Shashikumar, S. P., Josef, C. S., Sharma, A., & Nemati, S. (2021). DeepAISE–an interpretable and recurrent neural survival model for early prediction of sepsis. Artificial intelligence in medicine, 113, 102036.

24. Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. *International journal of medical informatics*, *114*, 57–65. https://doi.org/10.1016/j.ijmedinf.2018.03.013

25. Sudat S. E., Robinson S. C., Mudiganti S., Mani A., & Pressman A. R. (2021). Mind the clinical-analytic gap: Electronic health records and COVID-19 pandemic response. Journal of Biomedical Informatics, 116, 103715.

26. Norrie J., "The challenge of implementing AI models in the ICU," The Lancet Respiratory Medicine, vol. 6, no. 12, pp. 886–888, 2018.

27. Wiens, J., Shenoy, E. S. (2019). Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. Clinical Infectious Diseases, 66(1), 149-153.

28. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science (New York, N.Y.), 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

29. Yealy DM, Mohr NM, Shapiro NI, Venkatesh A, Jones AE, Self WH. Early care of adults with suspected sepsis in the emergency department and out-of-hospital environment: a consensus-based task force report. Ann Emerg Med. 2021;78(1):1–19. doi: 10.1016/j.annemergmed.2021.02.006.

30. Rhee C, Chiotos K, Cosgrove SE, Heil EL, Kadri SS, Kalil AC, Gilbert DN, Masur H, Septimus EJ, Sweeney DA, Strich JR, Winslow DL, Klompas M. Infectious diseases society of America position paper: recommended revisions to the national severe sepsis and septic shock early management bundle (SEP-1) sepsis quality measure. Clin Infect Dis. 2021;72(4):541–552. doi: 10.1093/cid/ciaa059.

31. Seymour CW, Kennedy JN, Wang S, Chang CCH, Elliott CF, Xu Z, Berry S, Clermont G, Cooper G, Gomez H, Huang DT, Kellum JA, Mi Q, Opal SM, Talisa V, van der Poll T, Visweswaran S, Vodovotz Y, Weiss JC, Yealy DM, Yende S, Angus DC. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA. 2019;321(20):2003–2017. doi: 10.1001/jama.2019.5791.

32. Moskowitz A, McSparron J, Stone DJ, Celi LA. Preparing a New Generation of Clinicians for the Era of Big Data. Harv Med Stud Rev. 2015;2(1):24-27.

33. Austin JM, Kachalia A. The State of Health Care Quality Measurement in the Era of COVID-19: The Importance of Doing Better. JAMA. 2020;324(4):333–334. doi:10.1001/jama.2020.11461

34. Boussina, A., Shashikumar, S., Amrollahi, F., Pour, H., Hogarth, M., & Nemati, S. (2023, July). Development & Deployment of a Real-time Healthcare Predictive Analytics Platform. In 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 1-4). IEEE.

35. Boussina, A., Shashikumar, S. P., Malhotra, A., Owens, R. L., El-Kareh, R., Longhurst, C. A., ... & Wardi, G. (2024). Impact of a deep learning sepsis prediction model on quality of care and survival. npj Digital Medicine, 7(1), 14.

36. Boussina A, Wardi G, Shashikumar SP, Malhotra A, Zheng K, Nemati S. Representation Learning and Spectral Clustering for the Development and External Validation of Dynamic Sepsis Phenotypes: Observational Cohort Study. J Med Internet Res. 2023 Jun 23;25:e45614. doi: 10.2196/45614. PMID: 37351927; PMCID: PMC10337434.

37. Boussina, A., Langouche, L., Obirieze, A. C., Sinha, M., Mack, H., Leineweber, W., ... & Fraley, S. I. (2024). Machine learning based DNA melt curve profiling enables automated novel genotype detection. BMC bioinformatics, 25(1), 185.

38. van de Sande D, van Genderen ME, Huiskens J, et al. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med 2021;47:750–760. 10.1007/s00134-021-06446-7

39. Wilkinson J, Arnold KF, Murray EJ, et al. Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Health 2020;2:e677–80. 10.1016/S2589-7500(20)30200-4

40. Chen, J. H., Asch, S. M. (2020). Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. New England Journal of Medicine, 376, 2507-2509.

41. Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of clinical epidemiology, 110, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

42. Shah NH, Milstein A, Bagley, PhD SC. Making Machine Learning Models Clinically Useful. JAMA. 2019;322(14):1351–1352. doi:10.1001/jama.2019.10306

43. Shorr, Andrew F. MD, MPH; Micek, Scott T. PharmD; Jackson, William L. Jr MD; Kollef, Marin H. MD. Economic implications of an evidence-based sepsis protocol: Can we improve outcomes and lower costs?*. Critical Care Medicine 35(5):p 1257-1262, May 2007. | DOI: 10.1097/01.CCM.0000261886.65063.CC

44. ektaş M, Tuynman JB, Costa Pereira J, Burchell GL, van der Peet DL. Machine Learning Algorithms for Predicting Surgical Outcomes after Colorectal Surgery: A Systematic Review. World J Surg. 2022

Dec;46(12):3100-3110. doi: 10.1007/s00268-022-06728-1. Epub 2022 Sep 15. PMID: 36109367; PMCID: PMC9636121.

45. Narayanan N, Gross AK, Pintens M, Fee C, MacDougall C. Effect of an electronic medical record alert for severe sepsis among ED patients. Am J Emerg Med. 2016 Feb;34(2):185-8. doi: 10.1016/j.ajem.2015.10.005. Epub 2015 Oct 14. PMID: 26573784.

46. Berger T, Birnbaum A, Bijur P, Kuperman G, Gennis P. A Computerized Alert Screening for Severe Sepsis in Emergency Department Patients Increases Lactate Testing but does not Improve Inpatient Mortality. Appl Clin Inform. 2010 Nov 17;1(4):394-407. doi: 10.4338/ACI-2010-09-RA-0054. PMID: 23616849; PMCID: PMC3633314.

47. Makam AN, Nguyen OK, Auerbach AD. Diagnostic accuracy and effectiveness of automated electronic sepsis alert systems: a systematic review. Journal of hospital medicine. 2015 Jun;10(6):396-402.

48. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ open respiratory research. 2017 Nov 1;4(1):e000234.

49. Nemati, S., Shashikumar, S. P., Holder, A. L., Wardi, G., & Owens, R. L. (2022). Randomized Clinical Trials or Convenient Controls: TREWS or FALSE?. medRxiv, 2022-08.

50. Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med. 2021;181(8):1065–1070. doi:10.1001/jamainternmed.2021.2626.

51. Henry, K.E., Adams, R., Parent, C. et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nat Med 28, 1447–1454 (2022). https://doi.org/10.1038/s41591-022-01895-z

52. Köpcke F, Prokosch HU. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res. 2014 Jul 1;16(7):e161. doi: 10.2196/jmir.3446. PMID: 24985568; PMCID: PMC4128959.

53. Russ, A. L., Zillich, A. J., McManus, M. S., Doebbeling, B. N., & Saleem, J. J. (2012). Prescribers' interactions with medication alerts at the point of prescribing: A multi-method, in situ investigation of the human-computer interaction. International journal of medical informatics, 81(4), 232–243. https://doi.org/10.1016/j.ijmedinf.2012.01.002

54. Zimek A, Schubert E, Kriegel HP. A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analy Data Mining 2012;5(5):363-387

55. Radovanovic M, Nanopoulos A, Ivanovic I. Hubs in space: popular nearest neighbors in high-dimensional data. J Mach Learn Res 2010;11(sept):2487-2531

56. Aggarwal C, Hinneburg A, Keim D. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In: Van den Bussche J, Vianu V, editors. Database Theory–ICDT 2001. Berlin, Heidelberg: Springer; 2001.

57. O'Grady, Naomi P. MD; Barie, Philip S. MD, MBA, FCCM; Bartlett, John G. MD; Bleck, Thomas MD, FCCM; Carroll, Karen RN; Kalil, Andre C. MD; Linden, Peter MD; Maki, Dennis G. MD; Nierman, David MD, FCCM; Pasculle, William MD; Masur, Henry MD, FCCM. Guidelines for evaluation of new fever in critically ill adult patients: 2008 update from the American College of Critical Care Medicine and the Infectious Diseases Society of America. Critical Care Medicine 36(4):p 1330-1349, April 2008. | DOI: 10.1097/CCM.0b013e318169eda9

58. Athamanolap P, Hsieh K, O'Keefe CM, Zhang Y, Yang S, Wang T-H. Nanoarray Digital Polymerase Chain Reaction with High-Resolution Melt for Enabling Broad Bacteria Identification and Pheno-Molecular Antimicrobial Susceptibility Test. Anal Chem. 2019 Oct 15;91(20):12784–12792.

59. Rolando JC, Jue E, Barlow JT, Ismagilov RF. Real-time kinetics and high-resolution melt curves in single-molecule digital LAMP to differentiate and study specific and non-specific amplification. Nucleic Acids Res. 2020 Apr17;48(7):e42.

60. Velez DO, Mack H, Jupe J, Hawker S, Kulkarni N, Hedayatnia B, et al. Massively parallel digital high resolution melt for rapid and absolutely quantitative sequence profiling. Sci Rep. 2017 Feb 8;7:42326.

61. Sinha M, Mack H, Coleman TP, Fraley SI. A High-Resolution Digital DNA Melting Platform for Robust Sequence Profiling and Enhanced Genotype Discrimination. SLAS Technol. 2018 Apr 13;23(6):580–591

62. Aralar A, Yuan Y, Chen K, Geng Y, Ortiz Velez D, Sinha M, et al. Improving Quantitative Power in Digital PCR through Digital High-Resolution Melting. J Clin Microbiol. 2020 May 26;58(6).

63. Baghdadi JD, Brook RH, Uslan DZ, et al. Association of a Care Bundle for Early Sepsis Management With Mortality Among Patients With Hospital-Onset or Community-Onset Sepsis. JAMA Intern Med. 2020;180(5):707-716. doi:10.1001/jamainternmed.2020.0183

64. Aaronson, E. L., Filbin, M. R., Brown, D. F., Tobin, K., & Mort, E. A. (2017). New mandated Centers for Medicare and Medicaid Services requirements for sepsis reporting: caution from the field. The Journal of Emergency Medicine, 52(1), 109-116.

65. Gesten, F., & Evans, L. (2021). SEP-1—taking the measure of a measure. JAMA Network Open, 4(12), e2138823-e2138823.

66. Rodos, A., Aaronson, E., Rothenberg, C., Goyal, P., Sharma, D., Slesinger, T., ... & Venkatesh, A. (2022). Improving Sepsis Management Through the Emergency Quality Network Sepsis Initiative. The Joint Commission Journal on Quality and Patient Safety, 48(11), 572-580.

67. Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023). Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277.

68. Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.

69. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

70. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601.

71. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... & Christiano, P. F. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33, 3008-3021.

72. Kanbar, L. J., Wissel, B., Ni, Y., Pajor, N., Glauser, T., Pestian, J., & Dexheimer, J. W. (2022). Implementation of machine learning pipelines for clinical practice: development and validation study. JMIR Medical Informatics, 10(12), e37833.

73. Henry, J. R., Lynch, D., Mals, J., Shashikumar, S. P., Holder, A., Sharma, A., & Nemati, S. (2018, July). A FHIR-enabled streaming sepsis prediction system for ICUs. In 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC) (pp. 4093-4096). IEEE.

74. Amrollahi, F., Shashikumar, S. P., Kathiravelu, P., Sharma, A., & Nemati, S. (2020, July). AIDEx-an open-source platform for real-time forecasting sepsis and a case study on taking ML algorithms to production. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 5610-5614). IEEE.

75. Benson, T., & Grieve, G. (2016). Hl7 version 2. In Principles of health interoperability (pp. 223-242). Springer, Cham.

76. Davis, S. E., Greevy Jr, R. A., Lasko, T. A., Walsh, C. G., & Matheny, M. E. (2020). Detection of calibration drift in clinical prediction models to inform model updating. Journal of biomedical informatics, 112, 103611.

77. Davis, S. E., Greevy Jr, R. A., Fonnesbeck, C., Lasko, T. A., Walsh, C. G., & Matheny, M. E. (2019). A nonparametric updating method to correct clinical prediction model drift. Journal of the American Medical Informatics Association, 26(12), 1448-1457.

78. Liu, A., Lu, J., & Zhang, G. (2020). Concept drift detection via equal intensity k-means space partitioning. IEEE transactions on cybernetics, 51(6), 3198-3211.

79. Giannini, H. M., Ginestra, J. C., Chivers, C., Draugelis, M., Hanish, A., Schweickert, W. D., ... & Umscheid, C. A. (2019). A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. Critical care medicine, 47(11), 1485-1492.

80. McCoy, A., & Das, R. (2017). Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. BMJ open quality, 6(2), e000158.

81. Lyons, P. G., Hofford, M. R., Sean, C. Y., Michelson, A. P., Payne, P. R., Hough, C. L., & Singh, K. (2023). Factors associated with variability in the performance of a proprietary sepsis prediction model across 9 networked hospitals in the US. JAMA internal medicine, 183(6), 611-612.

82. Wardi, G., Owens, R., Josef, C., Malhotra, A., Longhurst, C., & Nemati, S. (2023). Bringing the promise of artificial intelligence to critical care: what the experience with sepsis analytics can teach us. Critical care medicine, 51(8), 985-991.

83. Classen, D. C., Longhurst, C., & Thomas, E. J. (2023). Bending the patient safety curve: how much can AI help?. NPJ Digital Medicine, 6(1), 2.

84. Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., Futoma, J., Gao, M., ... & O'Brien, C. (2020). Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. JMIR medical informatics, 8(7), e15182.

85. Carlile, M., Hurt, B., Hsiao, A., Hogarth, M., Longhurst, C. A., & Dameff, C. (2020). Deployment of artificial intelligence for radiographic diagnosis of COVID‐19 pneumonia in the emergency department. Journal of the American College of Emergency Physicians Open, 1(6), 1459-1464.

86. Amrollahi, F., Shashikumar, S. P., Razmi, F., & Nemati, S. (2020). Contextual embeddings from clinical notes improves prediction of sepsis. In AMIA annual symposium proceedings (Vol. 2020, p. 197). American Medical Informatics Association.

87. Goh, K. H., Wang, L., Yeow, A. Y. K., Poh, H., Li, K., Yeow, J. J. L., & Tan, G. Y. H. (2021). Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nature communications, 12(1), 711.

88. Amrollahi, F., Shashikumar, S. P., Boussina, A., Yhdego, H., Nayebnazar, A., Yung, N., ... & Nemati, S. (2023, July). Predicting Hospital Readmission among Patients with Sepsis Using Clinical and Wearable Data. In 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 1-4). IEEE.

89. Wardi, G., Carlile, M., Holder, A., Shashikumar, S., Hayden, S. R., & Nemati, S. (2021). Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. Annals of emergency medicine, 77(4), 395-406.

90. Holder, A. L., Shashikumar, S. P., Wardi, G., Buchman, T. G., & Nemati, S. (2021). A locally optimized data-driven tool to predict sepsis-associated vasopressor use in the ICU. Critical care medicine, 49(12), e1196-e1205.

91. Amrollahi, F., Shashikumar, S. P., Holder, A. L., & Nemati, S. (2022). Leveraging clinical data across healthcare institutions for continual learning of predictive risk models. Scientific Reports, 12(1), 8380.

92. Li, R. C., Smith, M., Lu, J., Avati, A., Wang, S., Teuteberg, W. G., ... & Shah, N. H. (2022). Using AI to empower collaborative team workflows: two implementations for advance care planning and care escalation. NEJM Catalyst Innovations in Care Delivery, 3(4), CAT-21.

93. Gibbs, K. D., Shi, Y., Sanders, N., Bodnar, A., Brown, T., Shah, M. D., & Hess, L. M. (2021). Evaluation of a sepsis alert in the pediatric acute care setting. Applied Clinical Informatics, 12(03), 469-478.

94. Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The lancet, 370(9596), 1453-1457.

95. Lauritsen, S. M. et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat. Commun. 11, 1–11 (2020).

96. Moullin, J.C., Dickson, K.S., Stadnick, N.A. et al. Systematic review of the Exploration, Preparation, Implementation, Sustainment (EPIS) framework. Implementation Sci 14, 1 (2019). https://doi.org/10.1186/s13012-018-0842-6

97. R Core Team, R. (2013). R: A language and environment for statistical computing.

98. Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. The Annals of Applied Statistics, 247-274.

99. Takyi, P. O., & Bentum-Ennin, I. (2021). The impact of COVID-19 on stock market performance in Africa: A Bayesian structural time series approach. Journal of Economics and Business, 115, 105968.

100. Jalan, A., Matkovskyy, R., & Urquhart, A. (2021). What effect did the introduction of Bitcoin futures have on the Bitcoin spot market?. The European Journal of Finance, 27(13), 1251-1281.

101. Martin, W., Sarro, F., & Harman, M. (2016, November). Causal impact analysis for app releases in google play. In Proceedings of the 2016 24th ACM SIGSOFT international symposium on foundations of software engineering (pp. 435-446).

102. Danai, P. A., Sinha, S., Moss, M., Haber, M. J., & Martin, G. S. (2007). Seasonal variation in the epidemiology of sepsis. Critical care medicine, 35(2), 410-415.

103. Woodworth, L. (2020). Swamped: Emergency department crowding and patient mortality. Journal of health economics, 70, 102279.

104. Mlodzinski, E., Wardi, G., Viglione, C., Nemati, S., Crotty Alexander, L., & Malhotra, A. (2023). Assessing Barriers to Implementation of Machine Learning and Artificial Intelligence–Based Tools in Critical Care: Web-Based Survey Study. JMIR Perioperative Medicine, 6, e41056.

105. Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., ... & Terry, M. (2019, May). Human-centered tools for coping with imperfect algorithms during medical decision-making. In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1-14).

106. Kesselheim, A. S., Cresswell, K., Phansalkar, S., Bates, D. W., & Sheikh, A. (2011). Clinical decision support systems could be modified to reduce 'alert fatigue'while still minimizing the risk of litigation. Health affairs, 30(12), 2310-2317.

107. McCoy, A. B., Thomas, E. J., Krousel-Wood, M., & Sittig, D. F. (2014). Clinical decision support alert appropriateness: a review and proposal for improvement. Ochsner Journal, 14(2), 195-202.

108. Ginestra JC, Giannini HM, Schweickert WD, et al. Clinician Perception of a Machine Learning-Based Early Warning System Designed to Predict Severe Sepsis and Septic Shock. Crit Care Med. 2019;47(11):1477-1484. doi:10.1097/CCM.0000000000003803

109. Eysenbach, G. (2004). Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). Journal of medical Internet research, 6(3), e34.

110. York, T., Jenney, H., & Jones, G. (2020). Clinician and computer: a study on patient perceptions of artificial intelligence in skeletal radiography. BMJ Health & Care Informatics, 27(3).

111. Rogers, P., Boussina, A. E., Shashikumar, S. P., Wardi, G., Longhurst, C. A., & Nemati, S. (2023). Optimizing the implementation of clinical predictive models to minimize national costs: Sepsis case study. Journal of Medical Internet Research, 25, e43486.

112. Kostick-Quenet, K. M., & Gerke, S. (2022). AI in the hands of imperfect users. npj Digital Medicine, 5(1), 197.

113. DeMerle, K. M., Angus, D. C., Baillie, J. K., Brant, E., Calfee, C. S., Carcillo, J., ... & Seymour, C. W. (2021). Sepsis subclasses: a framework for development and interpretation. Critical care medicine, 49(5), 748-759.

114. Elbattah, M., & Molloy, O. (2017, January). Data-Driven patient segmentation using K-Means clustering: The case of hip fracture care in Ireland. In Proceedings of the Australasian Computer Science Week Multiconference (pp. 1-8).

115. Nicolet, A., Assouline, D., Le Pogam, M. A., Perraudin, C., Bagnoud, C., Wagner, J., ... & Peytremann-Bridevaux, I. (2022). Exploring patient multimorbidity and complexity using health insurance claims data: a cluster analysis approach. JMIR Medical Informatics, 10(4), e34274.

116. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1798-1828.

117. Landi, I., Glicksberg, B. S., Lee, H. C., Cherng, S., Landi, G., Danieletto, M., ... & Miotto, R. (2020). Deep representation learning of electronic health records to unlock patient stratification at scale. NPJ digital medicine, 3(1), 96.

118. Ng A, Jordan M, Weiss Y. On spectral clustering: analysis and an algorithm. Adv neural Inf Process Syst. 2002:849–856.

119. Wardi, G., Joel, I., Villar, J., Lava, M., Gross, E., Tolia, V., ... & Beitler, J. R. (2020). Equipoise in appropriate initial volume resuscitation for patients in septic shock with heart failure: results of a multicenter clinician survey. Journal of intensive care medicine, 35(11), 1338-1345.

120. Douglas, I. S., Alapat, P. M., Corl, K. A., Exline, M. C., Forni, L. G., Holder, A. L., ... & Hansell, D. M. (2020). Fluid response evaluation in sepsis hypotension and shock: a randomized clinical trial. Chest, 158(4), 1431-1445.

121. Villar, J., Clement, J. P., Stotts, J., Linnen, D., Rubin, D. J., Thompson, D., ... & Fee, C. (2014). Many emergency department patients with severe sepsis and septic shock do not meet diagnostic criteria within 3 hours of arrival. Annals of emergency medicine, 64(1), 48-54.

122. Gauer, R., Forbes, D., & Boyer, N. (2020). Sepsis: diagnosis and management. American family physician, 101(7), 409-418.

123. Neilson, H. K., Fortier, J. H., Finestone, P. J., Ogilby, C. M., Liu, R., Bridges, E. J., & Garber, G. E. (2023). Diagnostic delays in sepsis: lessons learned from a retrospective study of Canadian medico-legal claims. Critical Care Explorations, 5(2), e0841.

124. Liu, V. X., Fielding-Singh, V., Greene, J. D., Baker, J. M., Iwashyna, T. J., Bhattacharya, J., & Escobar, G. J. (2017). The timing of early antibiotics and hospital mortality in sepsis. American journal of respiratory and critical care medicine, 196(7), 856-863.

125. Centers for Medicare and Medicaid Services. QualityNet - Inpatient Hospitals Specifications Manual. https://www.qualitynet.org. Accessed April 16th, 2024.

126. Rhee, C., Brown, S. R., Jones, T. M., O'Brien, C., Pande, A., Hamad, Y., ... & CDC Prevention Epicenters Program. (2018). Variability in determining sepsis time zero and bundle compliance rates for the centers for medicare and medicaid services SEP-1 measure. Infection Control & Hospital Epidemiology, 39(8), 994-996.

127. Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., ... & Sibbald, W. J. (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Chest, 101(6), 1644-1655.

128. Rhee, C., Zhang, Z., Kadri, S. S., Murphy, D. J., Martin, G. S., Overton, E., ... & CDC Prevention Epicenters Program. (2019). Sepsis surveillance using adult sepsis events simplified eSOFA criteria versus sepsis-3 sequential organ failure assessment criteria. Critical care medicine, 47(3), 307-314.

129. Kamran, F., Tjandra, D., Heiler, A., Virzi, J., Singh, K., King, J. E., ... & Wiens, J. (2024). Evaluation of Sepsis Prediction Models before Onset of Treatment. NEJM AI, AIoa2300032.

130. Beaulieu-Jones, B. K., Yuan, W., Brat, G. A., Beam, A. L., Weber, G., Ruffin, M., & Kohane, I. S. (2021). Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians?. NPJ digital medicine, 4(1), 62.

131. Agniel, D., Kohane, I. S., & Weber, G. M. (2018). Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. Bmj, 361.

132. Tacconelli, E. (2017). Global Priority List of Antibiotic-Resistant Bacteria to Guide Research, Discovery, and Development.

133. Heymann, D. L., & Shindo, N. (2020). COVID-19: what is next for public health?. The lancet, 395(10224), 542-545.

134. Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E., & Crook, D. W. (2012). Transforming clinical microbiology with bacterial genome sequencing. Nature Reviews Genetics, 13(9), 601-612.

135. Kwong, J. C., McCallum, N., Sintchenko, V., & Howden, B. P. (2015). Whole genome sequencing in clinical and public health microbiology. Pathology, 47(3), 199-210.

136. Reed, G. H., & Wittwer, C. T. (2004). Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. Clinical chemistry, 50(10), 1748-1754.

137. Liew, M., Pryor, R., Palais, R., Meadows, C., Erali, M., Lyon, E., & Wittwer, C. (2004). Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons. Clinical chemistry, 50(7), 1156-1164.

138. Bidet, P., Liguori, S., Plainvert, C., Bonacorsi, S., Courroux, C., d'Humières, C., ... & Bingen, E. (2012). Identification of group A streptococcal emm types commonly associated with invasive infections and antimicrobial resistance by the use of multiplex PCR and high-resolution melting analysis. European journal of clinical microbiology & infectious diseases, 31, 2817-2826.

139. Roth, A. L., & Hanson, N. D. (2013). Rapid detection and statistical differentiation of KPC gene variants in Gram-negative pathogens by use of high-resolution melting and ScreenClust analyses. Journal of clinical microbiology, 51(1), 61-65.

140. Zianni, M. R., Nikbakhtzadeh, M. R., Jackson, B. T., Panescu, J., & Foster, W. A. (2013). Rapid discrimination between Anopheles gambiae ss and Anopheles arabiensis by High-Resolution Melt (HRM) analysis. Journal of Biomolecular Techniques: JBT, 24(1), 1.

141. Pritt, B. S., Mead, P. S., Johnson, D. K. H., Neitzel, D. F., Respicio-Kingry, L. B., Davis, J. P., ... & Petersen, J. M. (2016). Identification of a novel pathogenic Borrelia species causing Lyme borreliosis with unusually high spirochaetaemia: a descriptive study. The Lancet Infectious Diseases, 16(5), 556-564.

142. Langouche, L. (2021). Advancing Rapid Infectious Disease Screening Using a Combined Experimental/Computational Approach. University of California, San Diego.

143. Cheng, J. C., Huang, C. L., Lin, C. C., Chen, C. C., Chang, Y. C., Chang, S. S., & Tseng, C. P. (2006). Rapid detection and identification of clinically important bacteria by high-resolution melting analysis after broad-range ribosomal RNA real-time PCR. Clinical chemistry, 52(11), 1997-2004.

144. Athamanolap, P., Parekh, V., Fraley, S. I., Agarwal, V., Shin, D. J., Jacobs, M. A., ... & Yang, S. (2014). Trainable high resolution melt curve machine learning classifier for large-scale reliable genotyping of sequence variants. PloS one, 9(10), e109094.

145. Fraley, S. I., Athamanolap, P., Masek, B. J., Hardick, J., Carroll, K. C., Hsieh, Y. H., ... & Yang, S. (2016). Nested machine learning facilitates increased sequence content for large-scale automated high resolution melt genotyping. Scientific reports, 6(1), 19218.

146. Aralar, A., Goshia, T., Ramchandar, N., Lawrence, S. M., Karmakar, A., Sharma, A., ... & Fraley, S. I. (2024). Universal digital high-resolution melt analysis for the diagnosis of bacteremia. The Journal of Molecular Diagnostics, 26(5), 349-363.

147. Goshia, T., Aralar, A., Wiederhold, N., Jenks, J. D., Mehta, S. R., Karmakar, A., ... & Fraley, S. I. (2024). Universal digital high-resolution melting for the detection of pulmonary mold infections. Journal of Clinical Microbiology, e01476-23.

148. Sinha, M., Jupe, J., Mack, H., Coleman, T. P., Lawrence, S. M., & Fraley, S. I. (2018). Emerging technologies for molecular diagnosis of sepsis. Clinical microbiology reviews, 31(2), 10-1128.

149. Andini, N., Wang, B., Athamanolap, P., Hardick, J., Masek, B. J., Thair, S., ... & Yang, S. (2017). Microbial typing by machine learned DNA melt signatures. Scientific reports, 7(1), 42097.

150. Lu, S., Mirchevska, G., Phatak, S. S., Li, D., Luka, J., Calderone, R. A., & Fonzi, W. A. (2017). Dynamic time warping assessment of high-resolution melt curves provides a robust metric for fungal identification. PLoS One, 12(3), e0173320.

151. Bowman, S., McNevin, D., Venables, S. J., Roffey, P., Richardson, A., & Gahan, M. E. (2019). Species identification using high resolution melting (HRM) analysis with random forest classification. Australian Journal of Forensic Sciences, 51(1), 57-72.

152. Langouche, L., Aralar, A., Sinha, M., Lawrence, S. M., Fraley, S. I., & Coleman, T. P. (2020). Data-driven noise modeling of digital DNA melting analysis enables prediction of sequence discriminating power. Bioinformatics, 36(22-23), 5337-5343.

153. Adelman, J. D., McKay, W. R., Lillis, J., & Lawson, K. (2018). U.S. Patent No. 9,864,834. Washington, DC: U.S. Patent and Trademark Office.

154. Dwight, Z., Palais, R., & Wittwer, C. T. (2011). uMELT: prediction of high-resolution melting curves and dynamic melting profiles of PCR products in a rich web application. Bioinformatics, 27(7), 1019-1020.

155. Klinger, G., Levy, I., Sirota, L., Boyko, V., Reichman, B., Lerner-Geva, L., & Israel Neonatal Network. (2009). Epidemiology and risk factors for early onset sepsis among very-low-birthweight infants. American journal of obstetrics and gynecology, 201(1), 38-e1.

156. Stoll, B. J., Hansen, N. I., Sánchez, P. J., Faix, R. G., Poindexter, B. B., Van Meurs, K. P., ... & Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. (2011). Early onset neonatal sepsis: the burden of group B Streptococcal and E. coli disease continues. Pediatrics, 127(5), 817-826.

157. Palais, R., & Wittwer, C. T. (2009). Mathematical algorithms for high-resolution DNA melting analysis. Methods in enzymology, 454, 323-343.

158. Chevyrev, I., & Kormilitzin, A. (2016). A primer on the signature method in machine learning. arXiv preprint arXiv:1603.03788.

159. Morrill J, Fermanian A, Kidger P, Lyons T. A generalised signature method for multivariate time series feature extraction; 2020. arXiv preprint arXiv:2006.00873.

160. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

161. Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., ... & Glöckner, F. O. (2014). The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. Nucleic acids research, 42(D1), D643-D648.

162. Austin, P. C., & Steyerberg, E. W. (2012). Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC medical research methodology, 12, 1-8.

163. Yang, S., Ramachandran, P., Rothman, R., Hsieh, Y. H., Hardick, A., Won, H., ... & Gaydos, C. (2009). Rapid identification of biothreat and other clinically relevant bacterial species by use of universal PCR coupled with high-resolution melting analysis. Journal of clinical microbiology, 47(7), 2252-2255.

164. Athamanolap, P., Hsieh, K., & Wang, T. H. (2018, July). Integrated bacterial identification and antimicrobial susceptibility testing for Polymicrobial infections using digital PCR and digital high-resolution melt in a microfluidic Array platform. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 5346-5349). IEEE.

165. O'Keefe, C. M., & Wang, T. H. (2018, July). Digital High-Resolution Melt Platform for Rapid and Parallelized Molecule-by-Molecule Genetic Profiling. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 5342-5345). IEEE.

166. Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data mining and knowledge discovery, 31, 606-660.

167. Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. IEEE transactions on pattern analysis and machine intelligence, 20(3), 226-239.

168. Saraswathula, A., Merck, S. J., Bai, G., Weston, C. M., Skinner, E. A., Taylor, A., ... & Berry, S. A. (2023). The volume and cost of quality metric reporting. Jama, 329(21), 1840-1847.

169. Casalino, L. P., Gans, D., Weber, R., Cea, M., Tuchovsky, A., Bishop, T. F., ... & Evenson, T. B. (2016). US physician practices spend more than $15.4 billion annually to report quality measures. Health Affairs, 35(3), 401-406.

170. Ash, A. S., Fienberg, S. F., Louis, T. A., Normand, S. L. T., Stukel, T. A., & Utts, J. (2012). Statistical issues in assessing hospital performance.

171. Bauer, S. R., Gonet, J. A., Rosario, R. F., Griffiths, L. A., Kingery, T., & Reddy, A. J. (2019). 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, 45(2), 108-111.

172. Centers for Medicare & Medicaid Services. (2020). CMS announces relief for clinicians, providers, hospitals and facilities participating in quality reporting programs in response to COVID-19.

173. Rosenbaum, L. (2022). Reassessing quality assessment—the flawed system for fixing a flawed system. New England Journal of Medicine, 386(17), 1663-1667.

174. Centers for Medicare & Medicaid Services. (2023). Medicare program: proposed hospital inpatient prospective payment systems for acute care hospitals and the long-term care hospital prospective payment system and policy changes and fiscal year 2024 rates. Fed Regist, 88, 27193.

175. Rhee, C., Strich, J. R., Chiotos, K., Classen, D. C., Cosgrove, S. E., Greeno, R., ... & Klompas, M. (2024). Improving sepsis outcomes in the era of pay-for-performance and electronic quality measures: a joint IDSA/ACEP/PIDS/SHEA/SHM/SIDP position paper. Clinical Infectious Diseases, 78(3), 505-513.

176. Centers for Medicare and Medicaid Services. QualityNet - Hospital Quality Reporting Important Dates and Deadlines. https://www.qualitynet.org. Accessed April 23rd, 2024.

177. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180.

178. Jiang, L. Y., Liu, X. C., Nejatian, N. P., Nasir-Moin, M., Wang, D., Abidin, A., ... & Oermann, E. K. (2023). Health system-scale language models are all-purpose prediction engines. Nature, 619(7969), 357-362.

179. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica, 22(3), 276-282.

180. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods, 17(3), 261-272.

181. Seabold, S., & Perktold, J. (2010). Statsmodels: econometric and statistical modeling with python. SciPy, 7, 1.

182. HL7 FHIR Release 4. http://hl7.org/fhir/R4/overview.html. Accessed April 17th, 2024.

183. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

184. Kim, D., Park, C., Kim, S., Lee, W., Song, W., Kim, Y., ... & Kim, S. (2023). Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. arXiv preprint arXiv:2312.15166.

185. Dettmers, T., & Zettlemoyer, L. (2023, July). The case for 4-bit precision: k-bit inference scaling laws. In International Conference on Machine Learning (pp. 7750-7774). PMLR.

186. Hugging Face. Open LLM Leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard. Accessed April 18th, 2024.

187. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

188. Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., ... & Yu, T. (2022). One embedder, any task: Instruction-finetuned text embeddings. arXiv preprint arXiv:2212.09741.

189. Ahmad, F. S., Rasmussen, L. V., Persell, S. D., Richardson, J. E., Liss, D. T., Kenly, P., ... & Kho, A. N. (2019). Challenges to electronic clinical quality measurement using third-party platforms in primary care practices: the healthy hearts in the heartland experience. JAMIA open, 2(4), 423-428.

190. Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.

191. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.

192. Borgert, M., Binnekade, J., Paulus, F., Goossens, A., Vroom, M., & Dongelmans, D. (2016). Timely individual audit and feedback significantly improves transfusion bundle compliance—a comparative study. International Journal for Quality in Health Care, 28(5), 601-607.

193. Clark, R. C., Carter, K. F., Jackson, J., & Hodges, D. (2018). Audit and feedback: a quality improvement study to increase pneumococcal vaccination rates. Journal of Nursing Care Quality, 33(3), 291-296.

194. Glance, L. G., Maddox, K. J., Johnson, K., Nerenz, D., Cella, D., Borah, B., ... & Lin, Z. (2020). National Quality Forum guidelines for evaluating the scientific acceptability of risk-adjusted clinical outcome measures: a report from the National Quality Forum Scientific Methods Panel. Annals of surgery, 271(6), 1048-1055.