Predictive Approaches for Acute Adverse Events in Electronic Health Records

Dae Hyun Lee

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Meliha Yetisgen, Co-Chair

Eric Horvitz, Co-Chair

Lucy Vanderwende

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington

Abstract

Predictive Approaches for Acute Adverse Events in Electronic Health Records

Dae Hyun Lee

Chair of the Supervisory Committee:
Meliha Yetisgen and Eric Horvitz
Biomedical and Health Informatics

Medical errors have been cited as the third leading cause of death in the United States in 2013. Failure to rescue (FTR) is a subtype of medical errors and refers to the loss of an opportunity to save a patient's life after the development of one or more preventable and treatable complications. Focusing on detecting early signs of deterioration may therefore provide opportunities to prevent and/or treat an illness in a timely manner, which may in turn reduce the number of FTR cases. When implementing a data-driven model to predict the risk of potential FTR onsets in a supervised setting, gold standard information for the target FTR onset is often not directly retrievable in electronic health records (EHR) so that it requires to manually annotate clinical observations with corresponding labels. This method results in a bottleneck to scalability and the full utilization of

the clinical observations available in EHRs for model training. In this dissertation, I propose a machine learning framework that can be used to derive a risk prediction model using proxy events of the disease of interest, the administration of relevant clinical interventions, as a noisy label via a distant supervision approach. Moreover, this study evaluated the effects of considering the temporal progression of FTR risk estimates calculated using myopic evidence. Lastly, a case study is presented to demonstrate that the proposed prediction models can be deployed to quantify the adverse effects of clinical interventions with regard to the target disease of interest. This dissertation demonstrates 1) the feasibility of using proxy events of the target disease as a label for supervised model training, 2) the performance improvement when temporal progression is considered in the risk prediction model design, and 3) the applicability of the proposed risk prediction model to quantify the adverse effects of clinical interventions regarding the target disease. Suggestions are also provided on how the proposed model could be further improved by integrating experts' knowledge with the proposed framework.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

From taking my first steps into Seattle–Tacoma International Airport to the time of writing these acknowledgments, I have encountered numerous people whom I would like to thank for their kind support. Without their help, I don't think I would have been able to complete this long journey.

First and foremost, I cannot thank my family back in South Korea enough for their financial and emotional support, which allowed me to focus solely on my PhD studies. To my wife, Chloe, who quit her promising career as an equity analyst and successfully started a new career as an accountant, thank you for being a partner I can rely on in every moment of my life.

I would also like to thank my committee chairs, Eric Horvitz and Meliha Yetisgen, for their unconditional support from both academic and career perspectives. For Eric, all of the dissertation work was started from his words: "Why don't you focus on saving lives with the skill you have?" and "There is a field of study called machine learning. Why don't you take a look?" Without Eric's guidance, I also do not believe that I could have safely established a career in industry. Without Meliha's help, I sincerely doubt whether I would have come this far. Her guidance on writing and critics on my results during weekly meetings pushed me to go the extra mile. I would also like to thank Lucy Vanderwende in the reading committee for her patience and advices as I slowly made progress with my writing skills. I further thank Karol Bombsztyk,

the graduate school representative, for accepting this position at short notice and Ira Kalet, my first mentor and teammate, who helped me decide to come to Seattle.

Finally, I would like to extend my gratitude to all the faculties and my colleagues in the Biomedical and Health Informatics at the University of Washington for sharing their insights and resources, as well as the Hackers for allowing me to occasionally detach from my work with a stick and a puck.

Chapter 1. Introduction

1.1. Significance of the problem

The aftermath of medical errors resulted in $17.1 billion in unnecessary expenditures in 2008 [1]. To mitigate the adverse effects of medical errors, many government agencies have implemented countermeasures, such as funding medical research to better understand the extent of medical error[1] and penalizing clinical malpractice by denying reimbursement[2]. However, the mitigation of medical errors is still not fully addressed or understood, leaving medical error the third leading cause of death in the US in 2014 [2].

The medical errors grabbed the attention of the medical community when they were first highlighted by the Institute of Medicine's seminar paper "To Err is Human" in the late 1990s [3]. Upon being published, the report categorized different types of medical errors (Table 1.1). Of these different types of medical errors, patient harm caused by intra-management activities, such as error in the dose or method of using a drug and in the performance of an operation, procedure, or test, have been decreased because of safety guidelines [4] and policy measures [5]; the prevalence of catheter-induced infections was decreased significantly by implementing additional verification procedures before and after the interventions while the rate of errors around drug administration (e.g. administering the wrong dose or to the wrong patient) decreased significantly as barcode-based patient verification processes were implemented as a standard practice [6], [7]. Since the current electronic health records (EHRs) tend to log more information, not only about patients but also clinical practice around them, it would be less challenging to track how these types of errors were committed, thereby allowing practitioners to develop counteractive measures to systematically avoid such incidents [8].

Diagnostic
- Error or delay in diagnosis
- Failure to employ indicated tests
- Use of outmoded tests or therapy
- Failure to act on results of monitoring or testing
Treatment
- Error in the performance of an operation, procedure, or test
- Error in administering the treatment
- Error in the dose or method of using a drug
- Avoidable delay in treatment or in responding to an abnormal test
- Inappropriate (not indicated) care
Preventive
- Failure to provide prophylactic treatment
- Inadequate monitoring or follow-up of treatment
Other
- Failure of communication
- Equipment failure
- Other system failure

Table 1.1. Types of Medical Errors [3]

[1]https://www.ahrq.gov/news/newsroom/press-releases/health-affairs-patient-safety-research.html

[2]https://www.cms.gov/newsroom/fact-sheets/eliminating-serious-preventable-and-costly-medical-errors-never-events

In contrast, medical errors caused by omission—such as errors due to avoidable delays in treatment or in responding to an abnormal test, failure to act on results of monitoring or testing, and failure to provide prophylactic treatment—are still a challenge to be addressed in current practice [9], [10] because these types of errors are hard to foresee, thereby making them challenging to be defined and be extracted systematically from the existing EHR for further analyses. Of such cases, Silber et al. first identified groups of patients with preventable complications after cardiac surgeries [11]; they then introduced the concept of failure to rescue (FTR) [12]. The FTR refers to the failure to save patients who presented early signs of deterioration but on whom the current clinical workflow failed to apply proactive managements, even though such measures are currently available. This concept is now adopted as a standard quality metric from the Agency of Healthcare Research and Quality(AHRQ)'s patient safety indicators [13] because the proper management of FTR incidents not only improves the quality of care [14] but also would decrease hospital operation costs for handling unnecessary patient deterioration [15], [16]. However, although retrospective analyses could be done on such cases based on patients' discharge diagnoses and evidence from the EHRs [17], [18], implementing a framework to prevent such cases in practice is still challenging as it is hard to pin-point the time where such incidents are prevalent compared to the patient harms committed through intra-management activities. Therefore, if such an alerting system can be implemented into the workflow, thereby providing a quantitative estimates of the risk of future FTR incidents, it would allow caregivers to either verify biomarkers for the potential FTR incidents or handle such risk factors proactively in order to reduce the likelihood of these FTR event onsets.

Within the clinical domains, experts came up with the early warning scores (EWSs), which aim to quantify patients' risks of the adverse events or specific physiological states, as a tool to triage patients under management. The scores, such as the Glasgow Coma Score (GCS) [19] and the Acute Physiology and Chronic Health Evaluation (APACHE) [20] score, were designed to abstract multiple physiological measurements into fewer quantities so that caregivers could quickly assess the patient's status quo, thereby allowing them to reevaluate or to intervene if necessary. As they were also designed to be calculated in time-critical care settings, they use the latest physiological measurements and simple scoring criteria to derive scores to minimize the chance of calculation errors. Moreover, the risks estimated by the framework are straightforward to caregivers due to their simplicity. The scoring systems, however, are not able to consider a patient's physiological trajectory due to their static nature. Moreover, as most of EWSs are designed by groups of domain experts, they require significant effort to derive a standard scoring system that could be applied in various clinical settings.

In the clinical informatics domain, many prior studies strove to build a counterpart to such EWS with more clinical variables so that they can estimate a likelihood of the target diseases more accurately, and they were able to show a reasonable accuracy when predicting the events at the time of discharge [21]. For clinical outcomes related to FTR incidents, however, there are still some challenges that need to be addressed. First, most machine-learning models predicting clinical outcomes are trained in a supervised setting, and there is no clear way to extract the exact time of FTR incidents systematically in most of the EHRs that are currently deployed. The terminal outcomes of patients are available from the EHRs as discharge diagnoses. However, since gold-standard information regarding the exact time of FTR event onsets are mostly unavailable, prior studies have been annotating the event onsets by relying on expert-driven criteria for the event onset or manually annotate instances based on richer but noisy resources

such as clinical notes [22]. Such approaches are labor-intensive and expensive, and the cost would be even higher in labeling clinical outcomes compared to generic tasks (e.g. such as annotating general images or texts) as it requires annotators with domain expertise. Therefore, the current practice using criteria-based or manual annotations has some limitations regarding the scalability of data preprocessing. As the complexity of models tends to be increased in the current machine learning practice, and as they require more training data for the optimization, the scarcity of available training instances would result in a bottleneck when deriving models for estimating FTR risks.

To address this limitation, other studies have focused on using the timing of clinical interventions as the potential indicator of these event onsets. For example, Henry et al. annotated the onset of septic shock when hypotensive patients received volume resuscitations, which are frequently administered for patients with the event, and the trained model showed a reasonable accuracy for predicting the likelihood of the event [23]. Moreover, Suresh et al. showed that the timing of the clinical interventions dedicated to resuscitating a patient from severe deterioration could also be predicted with reasonable accuracy [24]. Although clinical indications of each intervention vary, the timing of clinical interventions is already documented in the EHRs and has the potential to be used as a proxy indicator for the FTR incidents. If clinical interventions frequently administered to the specific FTR events could be identified systematically, and the likelihood of receiving such interventions and the risk of developing the FTR incidents show a positive correlation, this approach would have the potential to generate a risk prediction model with less human labor compared to the current practice.

1.2. Specific Aims

The dissertation study focused on providing a framework that can derive risk estimation model on FTR incidents by using the timing of clinical intervention as a proxy label for the event onset. The framework first selects a list of clinical interventions that are frequently observed in patients discharged with the target FTR event, which could be verified through discharge diagnoses documented in the EHR. To improve the correlation between the timing of the clinical intervention and the event onset, the framework only considered the intervention administration from patients discharged with the target events as a proxy event for the event onset. The study specifically focused on acute organ failures (AOFs) developed in Intensive Care Units (ICUs) as FTR events because patients' prognoses are comparably more volatile than patients in wards, and early intervention on patients with high risk of AOF are known to improve their prognoses.

Similar to the EWS design currently used in the clinical settings, the study first evaluated whether such proxy events could be predicted by only using patients' physiologies measured up to 24 hours prior to the time of prediction, and it evaluated whether the likelihood of proxy event onsets could be used as a risk estimator for the AOF onsets. Then, with risk estimates from the model trained above, not only temporal progression presented in the risk trajectory of the target event was considered but also the trajectories of potentially relevant FTR events as to whether they would improve the prediction performance was evaluated, which is not utilized in the most of EWS designs. Lastly, a case study was conducted to examine how these risk estimates could be used in the current practice as a clinical decision support. The specific aims for the dissertation are covered in detail below.

- Aim 1: Learn about the risk of target AOF onsets based on clinical intervention and discharge diagnosis using distant supervision

Aim 1 focused on implementing a framework that can derive a risk prediction model for four target AOFs—Acute Heart Failure (AHF), Acute Lung Injury (ALI), Acute Kidney Injury (AKI), and Acute Liver Failure (ALF)—solely based on the data available from EHRs without experts' annotations. The study was conducted with the hypothesis that clinical interventions related to the target AOFs would be observed more frequently from patients discharged with the diseases, and it could serve as proxy indicator of the occurrence of the target AOF. The trained model, the Acute Organ Failure Intervention (AOFI) model, was aimed at predicting the likelihood of receiving relevant intervention to the target AOF within the next 24 hours from the time of prediction and being discharged with the target AOF based on previous 24 hours of physiologic observations and demographics. After the training, clinical validation was conducted to examine the quality of the likelihood as risk estimates of the target AOF.

- Aim 2: Improve clinical risk estimation by integrating risk trajectory holistically

In Aim 2, I focused on evaluating whether integrating the estimated risk trajectories of both the target AOF and the other AOFs could improve the prediction accuracy. As mentioned above, many EWSs have been proposed to quantify the risk of target diseases, and the AOFI models also aimed to predict the risk of each AOF onset. For scores from EWSs and probabilities from AOFI models, they only utilize the most recent physiologic measurements (e.g. using summary statics of the measurements within previous 24 hours or the most recent measurement) in order to derive a simpler model (for EWSs) and to facilitate clinical validation purposes (for AOFI models). Therefore, risks estimated in the earlier time periods were not utilized during the prediction in both cases. Aim 2 focused on evaluating the effects of considering temporal dynamics in risk trajectories on both the target AOF and the other AOFs where each quantity was estimated based on myopic evidence. I hypothesized that these trajectories would improve the prediction performance by providing additional information, which was either unavailable or underestimated in the original risk estimate. To do so, I first evaluated how the prediction performance of the expert-driven EWS, Multiple Organ Dysfunction Score (MODS) [25], on four different organ system (heart, respiratory, kidney, and liver) could be improved when screening high-risk AOF patients by consolidating EWS trajectories on four organ systems into the probabilistic framework using the hidden Markov model (HMM). The second experiment was conducted by considering the risk estimates from the AOFI models as an EWS for each AOF, and it measured how the performance was changed after the integration. Furthermore, the trained hidden Markov models in both experiments were clinically validated to evaluate how each model described patients' prognoses.

- Aim 3: Quantify intervention-induced risks using counterfactual analyses

Aim 3 was conducted as a case study showing how the risk prediction model trained from earlier aims could be used in the clinical workflow as a clinical decision support. In the critical care setting, the prevalence of clinical interventions that rapidly revert a patient's adverse physiology is higher than in wards, and their intensity comes with a higher risk of side effects. To quantitatively compare the magnitude of toxicity due to such clinical interventions, the objective of aim 3 was to decompose the adverse influence of target clinical interventions from the risks estimated by observed patients' physiologies. To do so, two submodels were trained jointly: one

model estimating a patient's baseline risk trajectory with the assumption that no clinical intervention was administered and another model estimating the magnitude of risk induced by the administered clinical interventions. Then, the estimated risk increments from the latter model were used to conduct a clinical validation using a counterfactual analysis: comparing the risk predicted with and without the clinical intervention. The first part of the experiment aimed to quantify the nephrotoxicity of antibiotics administered to ICU patients, and the model estimated the baseline kidney function trajectory and the nephrotoxicity of the administered antibiotics based on patient's serum creatinine level trajectories. The second part aimed to quantify the organ toxicity of clinical interventions frequently administered in ICUs based on the risk of AOF estimated by AOFI models.

1.3. Outline

In Chapter 2, I discuss how Aim 1 was conducted.

In Chapter 3, I discuss how the trajectory consideration changed the performance of both expert-driven risk estimates and the estimates from Aim 1 when predicting the target AOF onsets.

In Chapter 4, I discuss how the risk increment of individual interventions could be quantified and analyzed in the clinical sense.

In Chapter 5, I summarize my dissertation's findings. I discuss modeling recommendations for future studies that could potentially improve the prediction performance of modeling approaches presented in this dissertation. Furthermore, I outline how the proposed modeling approach could benefit the current clinical workflow as a clinical decision support.

Chapter 2. Predicting Severe Clinical Events by Learning about Life-Saving Actions and Outcomes using Distant Supervision

The methods described in this chapter are adopted from the following manuscript:

Lee D, Yetisgen M, Vanderwende L, Horvitz E. Predicting Severe Clinical Events by Learning about Life-Saving Actions and Outcomes using Distant Supervision. Journal of biomedical informatics (Under review)

Introduction

Medical errors have been cited as the third leading cause of death in the United States. A recent study estimated that more than 250,000 deaths were caused by medical error across the United States in 2013 [2]. Failure to Rescue (FTR) is a subtype of medical error, referring to the loss of opportunity to save a patient's life after the development of one or more preventable and treatable complications [26]. Thus, focusing on detecting early signs of deterioration may provide opportunities for preventing and/or treating illness in a timely manner, which promises to reduce the number of FTR cases [27].

To decrease the occurrence of FTR, several Early Warning Scores (EWSs) have been designed to detect and guide actions in time-critical care settings. The aim of EWSs is to give healthcare providers easily computable measures that provide insight into a patient's physiological status. Many EWSs, including the Acute Physiologic Assessment and Chronic Health Evaluation (APACHE) score [20], the Simplified Acute Physiology Score (SAPS) [28], and the Sequential (Sepsis-related) Organ Failure Assessment (SOFA) score [29], employ static scoring tables, predefined by domain experts, that map clinical measurements to discretized scores. Figure 2.1 presents the scoring table for the SOFA score as an example. Sepsis is a complex, time-critical condition, and prior studies have demonstrated reduced patient mortality with early detection and treatment of severe sepsis or septic shock [30], [31]. Therefore, if deployed carefully in existing clinical workflows, EWSs have the potential to improve patient outcomes, especially for those with acute onset diseases.

While manually curated EWSs can provide useful alerts in clinical settings [32], they make use of only a small portion of the information available in the Electronic Health Record (EHR). Today's EHRs include content ranging from hospital-operation-related information—such as patient locations and hospital charges—to care-related information—such as charted clinical observations, clinical notes, and demographic information—for each patient. Among such content, charted clinical observations with timestamps, including vital signs and lab test results, serve as valuable data sources when inferring a patient's prognosis. Therefore, instead of using only variables manually curated by specialists, implementing clinical risk prediction models that use all available information from the EHR could improve the accuracy of predicting the onset of clinically adverse events when compared to existing EWSs [23], [33], [34].

Table 1. Sequential [Sepsis-Related] Organ Failure Assessment Score[a]

System	Score 0	1	2	3	4
Respiration					
Pao_2/Fio_2, mm Hg (kPa)	≥400 (53.3)	<400 (53.3)	<300 (40)	<200 (26.7) with respiratory support	<100 (13.3) with respiratory support
Coagulation					
Platelets, $\times 10^3/\mu L$	≥150	<150	<100	<50	<20
Liver					
Bilirubin, mg/dL (μmol/L)	<1.2 (20)	1.2-1.9 (20-32)	2.0-5.9 (33-101)	6.0-11.9 (102-204)	>12.0 (204)
Cardiovascular	MAP ≥70 mm Hg	MAP <70 mm Hg	Dopamine <5 or dobutamine (any dose)[b]	Dopamine 5.1-15 or epinephrine ≤0.1 or norepinephrine ≤0.1[b]	Dopamine >15 or epinephrine >0.1 or norepinephrine >0.1[b]
Central nervous system					
Glasgow Coma Scale score[c]	15	13-14	10-12	6-9	<6
Renal					
Creatinine, mg/dL (μmol/L)	<1.2 (110)	1.2-1.9 (110-170)	2.0-3.4 (171-299)	3.5-4.9 (300-440)	>5.0 (440)
Urine output, mL/d				<500	<200

Abbreviations: Fio_2, fraction of inspired oxygen; MAP, mean arterial pressure; Pao_2, partial pressure of oxygen.

[a] Adapted from Vincent et al.[27]

[b] Catecholamine doses are given as μg/kg/min for at least 1 hour.

[c] Glasgow Coma Scale scores range from 3-15; higher score indicates better neurological function.

Figure 2.1. Scoring table for the SOFA score [35]

Although data-driven counterparts to EWSs have shown a higher accuracy in predicting clinical events in general, the medical community continues to focus on evaluating EWSs as clinical decision support tools for in-patient care settings [36], [37]. This reluctance to study data-driven models might suggest that their practicality is questioned [38], and physicians insist that the structure and representations provided by EWSs are more straightforward and actionable in practice. One aspect of EWSs that may make them preferable is that existing EWSs often quantify the patient's physiological status by simplifying multiple clinical measurements into a few clinically actionable representations, such as organ-level severities. For example, calculating a SOFA score entails first determining organ-level subscores using the predefined scoring table, then calculating the final severity score from those subscores. Similarly, the APACHE score uses selected representative physiologic variables for each organ system to calculate the final score from the predefined scoring table. As these two examples illustrate, the organ-level abstraction of a patient's physiological status is common in EWSs since physicians see it as a convenient decision support tool within existing clinical workflows. To be seamlessly merged with existing clinical workflows, data-driven models should also provide some form of risk estimates that physicians can interpret and use to plan corresponding action proactively.

In this paper, we present a machine learning approach for predicting the risk of acute onset diseases in an Intensive Care Unit (ICU) setting, which has the potential to allow a timely detection of patient deterioration and thereby prevention of possible FTR incidents. Compared to EWSs, the proposed approach considers all available physiological variables observed during a patient's ICU stay when training risk prediction models, and systematically selects variables that can maximize model accuracy in predicting the risk of developing the target acute-onset diseases in the near future.

Supervised machine learning approaches require annotated data to train models. In the study setting, however, human annotation is expensive and does not scale to large datasets. Instead of relying on human annotation, we use distant supervision [39]–[41] by defining the administration of clinical interventions dedicated to managing the target acute-onset disease as *proxy events* for the onset of the disease. Then, we predicted the proxy event onset based on previous physiological observations from the EHR. In an attempt to provide actionability of our approach comparable to that of EWSs, we examined four types of acute organ failures (AOFs)—acute heart failure (AHF), acute lung injury (ALI), acute kidney injury (AKI), and acute liver failure (ALF)—to answer the following research questions: (1) *Can we systematically identify clinical interventions, which will serve as distant supervision, that are frequently administrated for each type of acute organ failure?* And (2) *Can we use these interventions as supervisory signals to build models that predict the onset of acute organ failure?*

2.1. Related Work

In the clinical informatics domain, risk prediction models have been studied to predict the onset of diseases within a specified timeframe using supervised learning methods [42]–[46]. When implementing such supervised models, researchers often try to balance the tradeoff between prediction accuracy and model interpretability, though the definition of interpretability varies based on each model's objectives [47], [48]. When representing the influence of the features used in the model is important, simple logistic regression and linear regression are preferred because they can provide both the magnitude and direction of each feature's influence [49], [50]. Alternatively, if a proposed model is intended for use in clinical decision-making—such as medical imaging classification models—it can focus more on improving prediction accuracy while providing the evidence used to derive the predictions; in this case, there is less emphasis on why such portion of the instance were identified as evidence because a model's user is assumed to have background knowledge of potential relationships between predictions and evidence provided [51]. These relaxed constraints on interpretability allow models to use more complex conditions to provide more accurate predictions than those of logistic regression and linear regression models.

Based on the results of many comparative analyses [52], [53], tree-based ensemble modeling approaches tend to show moderately better performance over other modeling approaches by offering reasonable training time and computational resource requirements, while showing the extent to which each feature contributes to the model. This higher performance of tree-based ensemble approaches can be explained by the following: first, such ensemble approaches train submodels using subsampling. Therefore, although the objective of each submodel is to maximize accuracy of the given subsamples (making it prone to overfit), the prediction from an ensemble model can be more robust compared to other modeling approaches because it considers the predictions of all submodels trained with different parts of the training dataset. Moreover, the main criteria used by each tree-based submodel are logical conditions. They are therefore less influenced by how inputs are preprocessed compared to other approaches [54]. As the dataset for our study was expected to have heterogenous characteristics among different types of clinical observations, we chose to use a tree-based ensemble approach for modeling. Since the prevalence of the clinical problem we aimed to predict is known to be small, we selected a gradient boosting tree modeling approach over a random forest modeling approach because training error and generalization error can be bounded [55].

A major challenge of using supervised learning in a clinical informatics setting is creating a dataset annotated with the target event onsets for model training and testing. Such annotated datasets are often not readily extractable from EHRs, so manual annotation of disease onset is necessary. For example, Bejan et al. [22] identified the onset of ventilator-associated pneumonia among patients in the ICU using manual annotation according to criteria predefined by domain experts. Next, the labeled dataset and identified features from clinical notes were used to train a support vector machine to predict pneumonia onset within the upcoming 24 hours. However, such a manual annotation approach is labor-intensive, and it does not scale well to larger datasets. It is also prone to annotator bias. Therefore, if the annotation process can rely solely on information available in the EHR without any human intervention, it will render the process more scalable and more systematic, thereby reducing potential biases during the annotation process.

As an alternative to manual annotation, some studies have used distant supervision—leveraging patterns frequently observed from target events as noisy labels—to train supervised models. This approach is more widely used in natural language understanding, including sentiment analysis [41] and relation extraction [39], [40]. For example, Go et al. [41] trained a sentiment classification model from Twitter feeds, where tweets containing ":)" were labeled as positive sentiments and tweets containing ":(" were labeled as negative sentiments. The authors achieved accuracy above 80% for the task, and they showed that a reliance on distant supervision can yield performances similar to that of models trained by manually annotated labels [56].

Distant supervision has also been used in clinical informatics research by leveraging the administration timing of certain clinical interventions documented in EHRs as evidence of the time of disease onset. First-line interventions are actions taken by clinicians in response to a rapid deterioration in a patient's physiological status, and these measures aim to stabilize patients and improve their outcomes over a short period of time. Such interventions can provide strong signals that can be used when systematically annotating target events, thereby having the potential to be used as labels when training a supervised model in larger datasets. For example, Henry et al. [23] considered using the time of the initiation of fluid resuscitation with hypotension to identify the time at which septic shock onset was likely. Then, they fit a Cox proportional hazards model, which predicted the onset of septic shock with high accuracy. Although fluid resuscitation is frequently employed during septic shock and hypovolemia is often comorbid with septic shock, some patients develop hypovolemic shock from non-infectious causes. Distinguishing between shock with infectious and non-infectious causes is important when deploying this model into practice, but an analysis of these cases was not reported in the study. As this example illustrates, if the annotation process relies only on proxy events when labeling a potential onset of target clinical events, it is expected to yield labels with high uncertainty. However, since patients' discharge diagnoses are available in EHRs, using the time of proxy events as a marker of the potential event onset and verifying that annotations using discharge diagnoses might reduce the level of uncertainty in labels so that it can improve the performance of trained models.

In another example, Suresh et al. [24] trained an unsupervised switching state autoregressive model and then used the learned states, along with clinical variables, to predict future first-line intervention administration behavior in the ICU. The study showed that a subset of clinical intervention administrations could be predicted with high accuracy. Since these types of interventions tend to be administered in reaction to a specific physiological status, the

administration behavior could be considered as a marker of the physiological status that required the intervention to sustain the patient's life. Although the study did not further analyze the relationship between the predicted probabilities of intervention administration and the prognosis of potentially relevant diseases, it may exist between such factors, such as the probability of initiating mechanical ventilation and the risk of a patient developing hypoxemia or hypercarbia. Our study aimed to compare the probability of receiving interventions relevant to the target diseases with the likelihood of the target disease using the confirmatory measures available in the EHR.

Compared to the related works mentioned above, our study used a combination of intervention administrations and discharge diagnoses to annotate the potential onset of target acute organ failure. We then aimed to predict the identified potential onsets in the near future using physiological observations from larger-scale EHR datasets. We also conducted an additional analysis to examine whether the model—predicting the probability of the administration of first-line intervention in proximity and observing target acute organ failure discharge diagnoses—could be used as a risk estimator for the onset of target acute organ failure.

2.2. Methods

Through our work, we aimed to derive risk prediction models for four different acute organ failures (AHF, ALI, AKI, and ALF) as identified in patients' discharge diagnoses. Our approach included four steps: (1) representation of patient physiology from clinical data, (2) selection of relevant interventions for the onset of each acute organ failure, (3) annotation of the potential onset with selected interventions and discharge diagnoses, and (4) training models that estimate acute organ failure risks by predicting the likelihood of receiving one of the relevant interventions in the near future and being discharged with the target acute organ failure according to discharge diagnoses from the EHR. In the following sections, we will explain each step of our approach in detail.

2.2.1. Representation of Patient Physiology from Clinical Data

We represented each patient with multiple instances, in which each instance captured demographic information and clinical observations within a specific time window. In our experiments, we only considered clinical variables that were measured for at least 50% of patient-day observations in the EHR as clinical observations—these included vital signs, patient assessments documented in clinical charts (e.g., physical exam results), and lab test results. For demographic information, we used age, gender, admission source, ethnicity, insurance, language, religion, and marital status.

In our representation, the feature vector $x_{pts,[t_{i-1},t_i)}$ captured clinical observations from t_{i-1} to t_i for the patient *pts* in addition to the demographic information, and it was possible for there to be several clinical observations associated with a feature for the given time interval. We experimented with several summary functions—$\phi_j(obs_{T_0}, \dots, obs_{T_l})$, where $\phi_j: R^k \rightarrow R, T_s \in [t_{i-1}, t_i)$, and $s = 0,1, \dots, l$—and used average, min, max, and standard deviation to populate the all non-demographic and numeric features in the vector. To quantify the sustainment of each clinical observation from t_{i-1} to t_i (e.g., sustained high heart rate due to acutely developed tachycardia during the day), the sustainment quantifiers proposed in our previous work [57] were used as an additional summary function. For each type of numerical clinical observation,

sustainment quantifiers were derived by conducting *t*-tests comparing observations during the time window $[t_{i-1}, t_i)$ for the patient to population observations. The resulting *p*-value was used as a measure of sustainment for numerical observations during the given time period. Two sustainment quantifiers were added for each numerical clinical observation—one from a two-tailed *t*-test and another from a one-tailed *t*-test. Categorical variables, such as ethnicity and admission type, were transformed using one-hot encoding.

Observational studies have suggested that the signs of clinical deterioration can be seen up to 24 hours before a serious clinical event requiring intensive interventions [58]. We, therefore, set the length of the time interval for feature representation, $|t_i - t_{i-1}|$, to 24 hours. When the feature vector could not be populated because of missing observations, we imputed such features with mean values from the training set.

2.2.2. Selection of Relevant Interventions for Each Acute Organ Failure

We used statistical testing to identify clinical interventions that were potentially relevant to each acute organ failure. First, we identified subtypes for each acute organ failure discharge diagnosis in the ICD-9 (International Classification of Diseases, 9th revision). ICD-9 diagnosis codes with "acute" and "acute on chronic" for each acute organ failure were selected as subtypes because both included patients who acutely developed the target acute organ failure during their hospital stay, with or without being predisposed to the condition [59]. Table 2.1 presents the list of selected ICD-9 diagnosis codes. For each acute organ failure, we aggregated the selected codes into a single binary outcome variable, *discharge status*, in order to increase the statistical power of the findings.

Target Acute Organ Failure	ICD-9 Diagnosis Code	
AHF	428.21	Acute systolic heart failure
	428.23	Acute on chronic systolic heart failure
	428.31	Acute diastolic heart failure
	428.33	Acute on chronic diastolic heart failure
	428.41	Acute combined systolic and diastolic heart failure
	428.43	Acute on chronic combined systolic and diastolic heart failure
ALI	518.81	Acute respiratory failure
	518.51	Acute respiratory failure following trauma and surgery
	518.84	Acute on chronic respiratory failure
	518.53	Acute on chronic respiratory failure following trauma and surgery
AKI	584.9	Acute kidney failure, unspecified
	584.6	Acute kidney failure with lesion of cortical necrosis
	584.7	Acute kidney failure with lesion of medullary necrosis
	584.5	Acute kidney failure with lesion of tubular necrosis
	584.8	Acute kidney failure with specified pathology NEC
ALF	570	Acute and subacute necrosis of liver

Table 2.1. List of ICD-9 diagnosis codes for each target acute organ failure

We hypothesized that the probability of observing interventions which are relevant to managing the target acute organ failure would be higher for patients discharged with the corresponding

diagnoses than for the overall population of patients. To identify interventions relevant to each acute organ failure, we conducted a binomial test. To conduct the test, we first considered medication administrations and interventions charted in the EHR with Current Procedural Terminology (CPT), the Healthcare Common Procedure Coding System (HCPCS), or ICD-9 procedural codes with timestamps as clinical interventions. The binomial test then compared the probability of observing a clinical intervention in patients discharged with the target acute organ failure with the same probability for all patients in the training set. The *p*-value was used as a quantifier representing the strength of the relationship between the intervention and each acute organ failure, as presented in the dataset. For each type of acute organ failure, all clinical interventions under consideration from training hospital admissions were ranked by *p*-value in ascending order. Then, first *k* interventions (k=5, 10, 20, 50) were assumed to be relevant to the target acute organ failure and were subsequently used for the annotation.

2.2.3. Annotation with Interventions and Discharge Diagnoses

In our annotation approach of each target acute organ failure (AHF, ALI, AKI, and ALF), we labeled an instance $x_{pts,[t_{i-1},t_i)}$ as positive if the patient received at least one of the identified interventions between t_i and $t_i + threshold$ (Figure 2.2) and was discharged with one of the discharge diagnoses selected for the target organ failure as shown in Table 2.1.

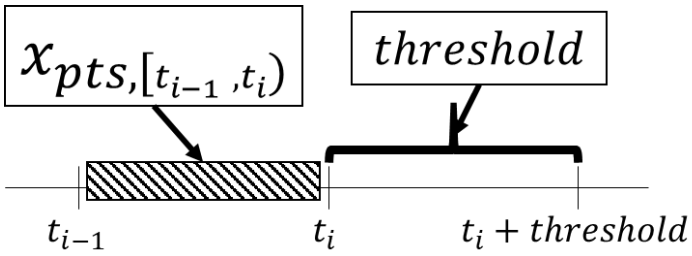

Figure 2.2. Illustration of clinical vector representation

On the other hand, we labeled an instance $x_{pts,[t_{i-1},t_i)}$ as negative based on the three following conditions: (1) if the patient did not receive any of selected interventions and was not discharged with the target acute organ failure, (2) if the patient did not receive any of the selected interventions but was discharged with the target acute organ failure, or (3) if the patient received one of the selected interventions between t_i and $t_i + threshold$ but was not discharged with the target organ failure.

For the second condition, instances were labeled negative because our primary focus was to learn about physiological patterns that preceded the potential onset of the target acute organ failure. Therefore, if there was no indicator of the onset of the patient's acute organ failure on which we could rely, we treated those instances as negative.

For the third condition, we hypothesized that the administration of the selected interventions in patients who were not discharged with the target acute organ failure could only indicate the development of comorbidity of the target acute organ failure, rather than the onset of target acute organ failure. Since the developed comorbidities did not result in the target acute organ failure for the patient according to discharge diagnoses, we did not consider the physiology that necessitated the selected interventions as a preceding physiological pattern of the target acute organ failure onset. For example, heparin sodium is commonly used as an anticoagulant, and is

administered to patients who have a higher risk of thromboembolism. However, even though thrombotic complication is common in heart failure [60], [61], the indication of heparin sodium injection varies from "prophylaxis and treatment of venous thromboembolism and pulmonary embolism" to "prevention of clotting in arterial and cardiac surgery" [62]. Therefore, even if heparin sodium was identified as one of the relevant interventions for acute heart failure, we were not confident in considering all of the physiological characteristics that necessitate heparin sodium injection as a preceding sign of the acute heart failure onset, unless the patient was specifically discharged with this disease. As explained in Section 2.2.2, we aimed to select interventions that were frequently observed in patients discharged with the target acute organ failure but comparably infrequent from the patient population in the dataset. Therefore, when the selected intervention is known to manage certain comorbidities, and those are prevalent from patients discharged with the target acute organ failure, we expected that the number of potential positive instances that were labeled as negative by this condition would be marginal and that their effect on the trained model's quality would be insignificant.

2.2.4. Training Models to Predict the Proxy Events of the Onsets of Acute Organ Failure

We employed the annotation criteria described in the previous section to train the acute organ failure intervention (AOFI) models and used the Gradient Boosted Tree (spark.ml library version 2.2.0 [63]) as our model building algorithm. The trained models predicted whether the patient would receive selected interventions within the next 24 hours ($=threshold$) and be discharged with the target acute organ failure based on available demographic information and the previous 24 hours ($= |t_i - t_{i-1}|$) of clinical observation: $p^{AOF}_{\mathrm{pts,[t_i,t_i+24hrs)}} = f_{AOF}\left(x_{pts,[t_{i-1},t_i)}\right)$, where f_{AOF} is the trained AOFI model for the target acute organ failure *AOF*. Before training the AOFI models, we excluded patients who were discharged with the target discharge diagnoses but did not receive any of the relevant interventions during their entire stay in the ICU from the training set. However, we kept such patients in the test set as discharge diagnoses were not known during the testing time. For this study, we did not make predictions on the date when the patient was admitted to the ICU, as there is no available evidence that AOFI models can use to make predictions.

2.2.5. Evaluation

To evaluate the performance of the AOFI models, we calculated the *aggregated probability* $p^{AOF}_{pts,agg}$ for each acute organ failure *AOF* and for each patient *pts* with the following formula:

$$p^{AOF}_{pts,agg} \triangleq 1 - \prod_i \left[1 - p^{AOF}_{pts,[t_i,t_i+24hrs)}\right]$$

The formula allowed us to represent the probability of patients receiving selected interventions at least once during their ICU stay and being discharged with the target acute organ failure. Then, the aggregated probability for each acute organ failure was compared with the discharge status presented in Table 2.1.

Optimal hyperparameters, including the number of relevant interventions for the annotation process and other hyperparameters for the Gradient Boosting Tree (such as the number of trees, the learning rate, and max depth) were selected based on the highest area under the precision-

recall curve (AUPRC) with the acute organ failure discharge status and aggregated probabilities from the validation set. We used the AUPRC as the performance metric for selecting the best model because it is a more sensitive metric for evaluating supervised machine learning models with unbalanced datasets (i.e., datasets in which the proportions of different types of labels are significantly different) for their discriminative power [64].

After we selected the highest performing AOFI model for each acute organ failure using the validation set, each model was then evaluated using observations from the test set. Evaluation was performed at two different levels: (1) predictions at instance-level (24-hours, $p^{AOF}_{pts,[t_i,t_i+24hrs)}$) and (2) predictions at patient-level (entire ICU stays, $p^{AOF}_{pts,agg}$). We measured the prediction performance of trained AOFI models on the test set using the following standard evaluation metrics: precision, recall, and F1-score.

2.3. Results

2.3.1. Dataset

We trained and evaluated the AOFI models with two datasets. The first dataset was the MIMIC-3 dataset [65], which is composed of ICU stays at the Beth-Israel Hospital in Boston, MA. This dataset contained two different EHR systems documenting patients admitted between 2001 and 2013. The first EHR system, CareVue (Phillips), covers admissions from 2001 to 2008; the second, MetaVision (iMD Soft), covers admissions from 2008 to 2013. For this study, we only considered clinical data from MetaVision because it contained more detailed information on clinical interventions. The second dataset was extracted from the University of Washington Clinical Data Repository (UW-CDR dataset). This dataset contained information about patients admitted to ICUs at the University of Washington Medical Center and Harborview Medical Center between 2014 and 2016.

For both the MIMIC-3 and UW-CDR datasets, patients under the age of 18 were excluded because the normal range of physiological variables differs between children and adults. We also censored clinical observations after patients' code status was changed from full code (a care preference indicating that patients desire all necessary clinical measures be taken to prolong their lives) to *Do Not Intubate*, *Do Not Resuscitate*, or *Comfort Measure Only* because care preferences other than full code significantly alters the pattern of standard clinical intervention administrations. Table 2.2 shows a brief summary of patient demographics in the final datasets.

Descriptor	MIMIC-3	UW-CDR
Number of hospital admissions	22,020	14,506
Number of ICU admissions	23,593	16,612
Number of patient-days	98,529	65,875
Age	64.39±17.06*	57.63±17.19
Admission type – Number	Elective: 3,006 Emergency: 18,744 Urgent: 243	Elective: 3,861 Emergency: 7,819 Urgent: 1,631 Trauma: 281 Unknown: 6
Number of in-hospital deaths	2,294 (10.42%)	2,275 (16.77%)

Table 2.2. Patient demographics. *981 hospital admissions were for patients over the age of 90, and their ages were hidden by the data provider for de-identification purposes.

To train and test the AOFI models, we divided hospital admissions in each MIMIC-3 and UW-CDR dataset into training, validation, and test sets, with a 7:2:1 ratio, respectively. After preprocessing, as described in Section 3.1, we generated 1,177 features for the MIMIC-3 dataset and 287 features for the UW-CDR dataset for each feature vector $x_{pts,[t_{i-1},t_i)}$.

2.3.2. Selection of Relevant Interventions and Dataset Annotation

For each acute organ failure, we performed experiments on different numbers of interventions (k=5, 10, 20, 50) to select the best *k* value that maximized the AUPRC on aggregated probabilities, in addition to other hyperparameters (see Section 2.2.4), using the validation set. To
accomplish this, aggregated probabilities for each patient were calculated using the predicted probabilities from each AOFI model. We then compared those aggregated probabilities against the patients' discharge statuses on each acute organ failure. As shown in Table 2.3, for the MIMIC-3 dataset, 10 interventions were selected to annotate AHF and AKI; 5 interventions were selected for ALI; and 20 interventions were selected for ALF. In the UW-CDR dataset, 5 interventions were selected for AHF and ALI; 20 interventions were selected for AKI; and 10 interventions were selected for ALF. The selected interventions are presented in Table 2.4.

# Intervention (k)	MIMIC-3				UW-CDR			
	AHF	**ALI**	**AKI**	**ALF**	**AHF**	**ALI**	**AKI**	**ALF**
5	0.3451	**0.6754**	0.6209	0.1667	**0.4132**	**0.5745**	0.5252	0.0802
10	**0.3863**	0.6754	**0.6570**	0.1464	0.3599	0.5732	0.5986	**0.1662**
20	0.3833	0.6481	0.6397	**0.2215**	0.3912	0.5688	**0.6131**	0.1495
50	0.3644	0.6010	0.6214	0.2201	0.3977	0.5501	0.5993	0.1451

Table 2.3. AUPRC of aggregated probability and discharge diagnoses from the validation set

In our analysis of the selected interventions presented in Table 2.4, we observed that those from the MIMIC-3 dataset were mostly first-line interventions for each acute organ failure onset (e.g., heparin sodium—AHF, invasive ventilation—ALI, continuous renal replacement therapy—AKI, and fresh frozen plasma transfusion—ALF). For the UW-CDR dataset, confirmatory test orders for the target acute organ failure were also selected along with first-line interventions in the cases of AHF, AKI, and ALF (e.g., echocardiography—AHF, creatinine and urine tests—AKI, and hepatitis infection tests—ALF). In the case of ALI, first-line interventions were mainly selected, similar to the MIMIC-3 dataset.

Target Acute Organ Failure	AHF	ALI	AKI	ALF
MIMIC-3	• Heparin Sodium • Insulin – Humalog • Non-invasive Ventilation • Dopamine • Milrinone • Furosemide (Lasix) 500/100 • Furosemide (Lasix) • Coumadin (Warfarin) • Cardiac Cath • Nitroglycerin	• Invasive Ventilation • Midazolam (Versed) • Fentanyl • Fentanyl (Concentrate) • Ventilation assist and management, subsequent day [C]	• ACD-A Citrate (1000ml) • Calcium Gluconate (CRRT) • Insulin – Humalog • Vasopressin • Sodium Bicarbonate 8.4% • KCl (CRRT) • Albumin 25% • Dialysis – CRRT • Citrate • Norepinephrine	• Intubation • Platelets • Citrate • KCl (CRRT) • Albumin 25% • Ventilation assist and management, subsequent day [C] • Calcium Gluconate (CRRT) • Pantoprazole (Protonix) • Cryoprecipitate • Midazolam (Versed) • ACD-A Citrate (1000ml) • Sodium Bicarbonate 8.4% • Cisatracurium • Vasopressin • Fresh Frozen Plasma • Calcium Gluconate • Dialysis – CRRT • Fentanyl (Concentrate) • Norepinephrine • Fentanyl
UW-CDR	• Dobutamine • Assay of Natriuretic Peptide [C] • Echocardiography [C] • Blood gases, O_2 saturation only [C] • Dobutamine, Hydro-Chloride Injection [H]	• Emergency endotracheal intubation [C] • Ventilation assist and management, initial day [C] • Radiologic exam, abdomen [C] • Blood culture, aerobic bacteria [C] • Ventilation assist and management, subsequent day [C]	• Infusion, Albumin (human), 25% [H] • Gases, blood, O_2 saturation; direct measurement [C] • Creatinine; other source[C] • Hepatitis B surface antigen detection [C] • Norepinephrine • Lactate dehydrogenase [C] • Creatinine measurement, other source [C] • Urine bacterial culture [C] • Urine sodium measurement [C] • Hepatitis B surface antibody (HBsAb) • Blood bacterial culture, aerobic [C] • Urine Chloride measurement [C] • Vancomycin HCL injection, 250 MG [H] • Automated complete blood count and diff. WBC count [C] • Hepatitis C antibody [C] • Vancomycin drug assay [C] • Urine potassium measurement [C] • Vancomycin, HCL injection, 500 MG [H] • Total Hepatitis B core antibody (HBcAb) [C] • Albumin, human	• Phytonadione • Lactulose • Infectious agent antigen detection with immunoassay, hepatitis B surface antigen (HBsAg) [C] • Ammonia [C] • Duplex scan of arterial inflow and venous outflow of abdominal, pelvic, scrotal contents, and/or retroperitoneal organs; complete study [C] • Hepatitis B core antibody (HBcAb); total [C] • Hepatitis C antibody [C] • Abdominal Ultrasound [C] • Hepatitis B surface antibody (HBsAb) [C] • 25% Albumin infusion [H]

Table 2.4. Selected interventions used for annotating the onset of each acute organ failure; [C]: interventions documented with CPT code; [H]: interventions documented with HCPCS code.

Table 2.5 shows the number of positively labeled instances from training, validation, and the test set for each dataset, according to our annotation criteria with the selected interventions and discharge statuses.

	MIMIC-3			UW-CDR		
	Training	**Validation**	**Test**	**Training**	**Validation**	**Test**
# of hospital admissions	15,855	3,963	2,202	10,445	2,611	1,450
# of instances (patient-day)	70,452	18,207	9,870	47,174	12,426	6,275
AHF	6,063	1,602	657	1,621	723	215
ALI	14,962	4,136	2,222	9,494	2,823	1,184
AKI	8,862	2,773	1,258	7,952	2,458	924
ALF	2,263	906	311	1,180	349	113

Table 2.5. Number of positive training, validation, and test instances for the MIMIC-3 and UW-CDR datasets.

2.3.3. Prediction Performance

2.3.3.1. Instance-Level Prediction Performance

For instance-level evaluation, predicted probabilities $p^{AOF}_{pts,[t_i,t_i+24hrs)}$ were compared with the labels assigned according to annotation criteria (i.e., when the patient received one of the selected interventions for the target acute organ failure within the next 24 hours and was discharged with the target acute organ failure). Table 2.6 shows the contingency tables for instance-level predictions in the test set from the MIMIC-3 and UW-CDR datasets.

In both datasets, the ALI and AKI AOFI models showed comparably higher F1 scores (0.6758 and 0.4858 for ALI; 0.4569 and 0.4583 for AKI) than the AHF and ALF AOFI models (0.2518 and 0.2891 for AHF; 0.2202 and 0.0906 for ALF). As shown in Table 2.5, the proportion of positive instances for training AHF and ALF AOFI models was lower than the proportion of positive instances for training the ALI and AKI AOFI models in both datasets. Consequently, we suspect that the lower performance on AHF and ALF AOFI models is mainly due to an insufficient number of positive instances, which might not yield enough distinctive patterns for training AHF- and ALF-positive cases. The difference in performance between datasets on the same acute organ failure might be due to the difference in the strength of relationship between selected interventions and the discharge status presented in each dataset.

(a) MIMIC-3

AHF		Label	
		0	1
Prediction	0	8447	452
	1	766	205
	Precision	**Recall**	**F1**
	0.2111	0.3120	0.2518

ALI		Label	
		0	1
Prediction	0	6637	572
	1	1011	1650
	Precision	**Recall**	**F1**
	0.6201	0.7426	0.6758

AKI		Label	
		0	1
Prediction	0	7705	617
	1	907	647
	Precision	**Recall**	**F1**
	0.4141	0.5095	0.4569

ALF		Label	
		0	1
Prediction	0	9296	240
	1	263	71
	Precision	**Recall**	**F1**
	0.2126	0.2283	0.2202

(b) UW-CDR

AHF		Label	
		0	1
Prediction	0	5914	154
	1	146	61
	Precision	**Recall**	**F1**
	0.2947	0.2837	0.2891

ALI		Label	
		0	1
Prediction	0	4361	570
	1	730	614
	Precision	**Recall**	**F1**
	0.4568	0.5186	0.4858

AKI		Label	
		0	1
Prediction	0	4657	443
	1	694	481
	Precision	**Recall**	**F1**
	0.4094	0.5206	0.4583

ALF		Label	
		0	1
Prediction	0	5980	99
	1	182	14
	Precision	**Recall**	**F1**
	0.0714	0.1239	0.0906

Table 2.6. Contingency table and performance metrics on instance-level prediction from (a) the MIMIC-3 dataset, and (b) the UW-CDR dataset.

To measure the clinical validity of the probabilities predicted by the AOFI models as risk estimates for the target AOF, we first conducted a literature review to identify lab tests that were suggested as biomarkers for each acute organ failure by the clinical community. From the review, we identified brain natriuretic peptide for AHF [66], arterial CO_2 pressure for ALI [67], urea nitrogen for AKI [68], and total bilirubin for ALF [69]. Then, the predicted probabilities from each AOFI model on test instances, $p^{AOF}_{pts,[t_i,t_i+24hrs)}$, were compared to selected lab test measurements available during the prediction window, $[t_i, t_i + 24hrs)$, by calculating Pearson's correlation coefficient (see Figure 2.3). The correlation analysis results in Table 2.7 indicate a high correlation between predicted probabilities from the AKI and ALF AOFI models and the confirmatory lab test results in both datasets (min: 0.2165 from the ALF AOFI model in the UW-CDR dataset; max: 0.3492 from the AKI AOFI model in the MIMIC-3) while the AHF and ALI AOFI models still showed the expected direction of correlation (min: 0.0526 from the AHF AOFI model in the MIMIC-3 dataset; max: 0.1786 from the ALI AOFI model in the UW-CDR dataset). This indicates that when lab tests do not provide immediate results that can be used to evaluate a patient's risk of acute organ failure onset, the AOFI model's capability to provide information on the probability of receiving relevant interventions and being discharged with the target acute organ failure might offer critical information in a more timely manner.

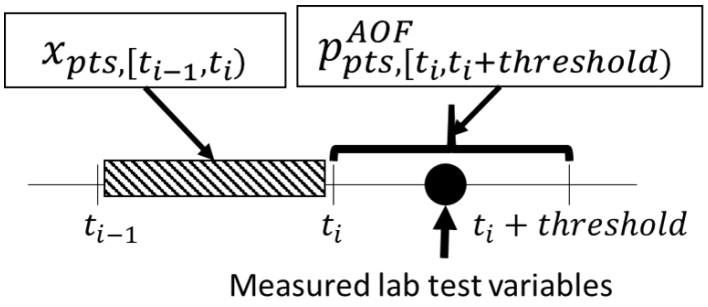

Figure 2.3. Study design of the correlation analysis between lab test measurements and predicted probabilities.

AOFI Model	**Lab Test**	**MIMIC-3**	**UW-CDR**	Reference
AHF	Brain Natriuretic Peptide	0.0526 (58)	0.0693 (98)	[66]
ALI	pCO_2	0.0613 (9427)	0.1786 (650)	[67]
AKI	Urea Nitrogen	0.3492 (10596)	0.3492 (6917)	[68]
ALF	Total Bilirubin	0.3086 (2563)	0.2165 (1481)	[69]

Table 2.7. Pearson's correlation coefficient between predicted probabilities and lab test measurements (number of available lab test measurements from test patients).

2.3.3.2. Patient-Level Prediction Performance

For patient-level evaluation, aggregated probabilities from each patient, $p^{AOF}_{pts,agg}$, were compared with discharge statuses. The contingency tables of the aggregated probabilities and discharge statuses with corresponding precision, recall, and F1 scores are shown in Table 2.8.

(a) MIMIC-3

AHF		Label	
		0	1
Prediction	0	1891	169
	1	79	48
	Precision	**Recall**	**F1**
	0.3780	0.2211	0.2791

ALI		Label	
		0	1
Prediction	0	1692	147
	1	120	228
	Precision	**Recall**	**F1**
	0.6552	0.6080	0.6307

AKI		Label	
		0	1
Prediction	0	1614	321
	1	68	184
	Precision	**Recall**	**F1**
	0.7302	0.3644	0.4861

ALF		Label	
		0	1
Prediction	0	2105	31
	1	37	14
	Precision	**Recall**	**F1**
	0.2745	0.3111	0.2917

(b) UW-CDR

AHF		Label	
		0	1
Prediction	0	1278	42
	1	15	10
	Precision	**Recall**	**F1**
	0.4000	0.1923	0.2597

ALI		Label	
		0	1
Prediction	0	1096	76
	1	79	94
	Precision	**Recall**	**F1**
	0.5434	0.5529	0.5481

AKI		Label	
		0	1
Prediction	0	1105	80
	1	72	88
	Precision	**Recall**	**F1**
	0.5500	0.5238	0.5366

ALF		Label	
		0	1
Prediction	0	1279	21
	1	35	10
	Precision	**Recall**	**F1**
	0.2222	0.3226	0.2632

Table 2.8. Contingency table and performance metric on patient-level prediction performance from (a) the MIMIC-3 dataset, and (b) the UW-CDR dataset.

Similar to the instance-level prediction performance presented in Table 2.6, the ALI and AKI AOFI models showed higher F1 scores on both datasets (0.6307 and 0.5481 for ALI; 0.4861 and 0.5366 for AKI) compared to the AHF and ALF AOFI models (0.2741 and 0.2597 for AHF; 0.2917 and 0.2632 for ALF). As discussed in the instance-level prediction performance in Section 2.3.3.1, the performance difference between the AHF and ALF AOFI models and the ALI and AKI AOFI models in both datasets could be a result of an insufficient number of positive instances for training the AHF and ALF AOFI models. For the ALF AOFI model in the UW- CDR dataset, we suspect that the labeling criteria might focus too highly on specific ALF subtypes. Although it performed the worst in instance-level predictions, the ALF AOFI model in the UW-CDR dataset nevertheless showed a correlation with total bilirubin levels and had comparable results in patient-level predictions to the ALF AOFI model from the MIMIC-3 dataset.

There is medical consensus that the risk of one organ failure depends on the risk of other organ failures [70]. As the scatter matrix plot in Figure 2.4 illustrates, we observed a strong positive correlation in all pairs of aggregated probabilities from the AOFI models in both datasets. Moreover, the aggregated probabilities of ALF tend to be lower than the aggregated probabilities of ALI and AKI from most patients in both datasets, and these trends can be written as the inequality with a conditional probability: $p(AOF_1) > p(AOF_2) \leftrightarrow p(AOF_1|AOF_2) > p(AOF_2|AOF_1)$. Accordingly, two probabilistic inequalities shared by both datasets can be derived, $p(ALI|ALF) > p(ALF|ALI)$ and $p(AKI|ALF) > p(ALF|AKI)$, which agree with the observation from discharge diagnoses in both datasets (Table 2.9). Through literature review, we found supporting evidence of these findings—that subtypes of acute kidney injury are frequently observed in patients with liver dysfunction [71], and pulmonary infection is prevalent in patients with acute liver failure [72], [73]; this might explain the higher aggregated probabilities from ALI and AKI AOFI models than the probabilities from ALF AOFI models in both datasets.

Data Sources	$\frac{\#(ALI \cap ALF)}{\#(ALF)}$	$\frac{\#(ALI \cap ALF)}{\#(ALI)}$	$\frac{\#(AKI \cap ALF)}{\#(AKI)}$	$\frac{\#(AKI \cap ALF)}{\#(ALF)}$
MIMIC-3	**0.4222**	0.0507	**0.7333**	0.0661
UW-CDR	**0.7667**	0.1353	**0.8667**	0.1576

Table 2.9. Conditional probabilities of observing discharge diagnoses of ALI, AKI, and ALF; $\#(A \cap B)$ represents the number of patients discharged with both AOF A and B.

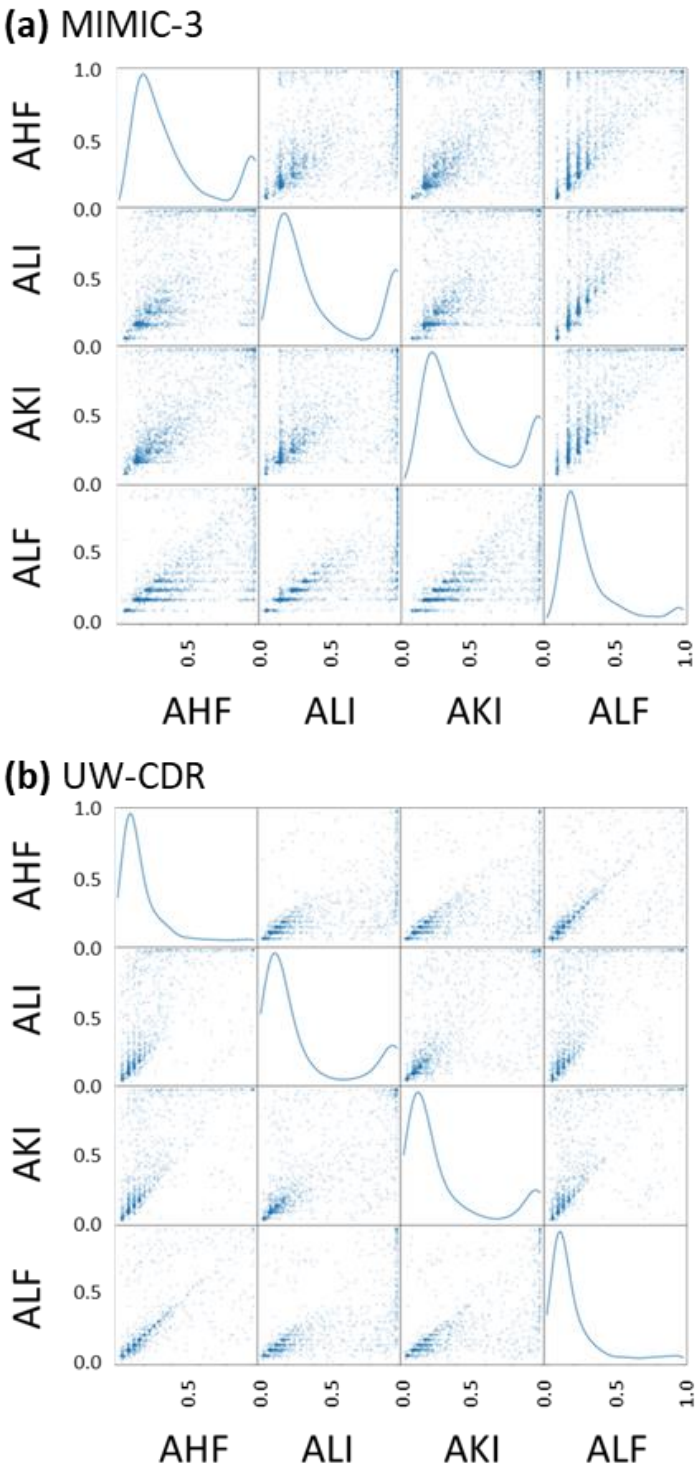

Figure 2.4. Scatter matrix plot for aggregated probabilities from (a) the MIMIC-3 dataset, and (b) the UW-CDR dataset. Diagonal entries represent the probability density of aggregated probabilities for each AOFI model.

2.3.4. Error Analysis

2.3.4.1. Instance-Level Analysis

We conducted error analyses at both the instance and patient levels. An analysis of instance-level predictions showed that an average of 18.83% false-positive instances were generated from patients discharged with the target acute organ failure—12.14% (93 instances) from the AHF AOFI model, 13.85% (140 instances) from the ALI AOFI model, 38.15% (346 instances) from the AKI AOFI model, and 3.80% (10 instances) from the ALF AOFI model in the MIMIC-3 dataset; 26.71% (39 instances) from the AHF AOFI model, 21.78% (159 instances) from the ALI AOFI model, 26.51% (184 instances) from the AKI AOFI model, and 7.69% (14 instances) from the ALF AOFI model in the UW-CDR dataset. Of those false-positive predictions, an average of 17.74% occurred one day before the target date (i.e., when patients who discharged with the target acute organ failure received selected interventions; Figure 2.5a)—11.83% from the AHF AOFI model, 17.14% from the ALI AOFI model, 16.73% from the AKI AOFI model, and 20% from the ALF AOFI model in the MIMIC-3 dataset; 30.77% from the AHF AOFI model, 26.42% from the ALI AOFI model, 19.02% from the AKI AOFI model, and none from the ALF AOFI model in the UW-CDR dataset. Although classified as false-positives according to our annotation criteria, these false-positive predictions could be useful as early warning signals when trained AOFI models are deployed.

Similarly, of the all false-negative predictions generated from patients discharged with the target acute organ failure according to our annotation criteria, we found that an average of 63.12% were made one day after the target date (Figure 2.5b)—58.85% from the AHF AOFI model, 62.59% from the ALI AOFI model, 53.97% from the AKI AOFI model, and 77.08% from the ALF AOFI model in the MIMIC-3 dataset; and 55.84% from the AHF AOFI model, 68.95% from the ALI AOFI model, 61.85% from the AKI AOFI model, and 57.57% from the ALF AOFI model in the UW-CDR dataset. We suspect these false-negative predictions occurred because the selected interventions had been administered as late as the day before the target date to manage patients' risks of the target acute organ failure, and the interventions might control abnormal physiological statuses. Therefore, the predictions, which are based on clinical observations after the interventions, might yield lower risks. Consequently, we could consider those predictions as true-negative predictions regarding the risk of the target acute organ failure onset, even though they were classified as false-negatives according to the annotation criteria.

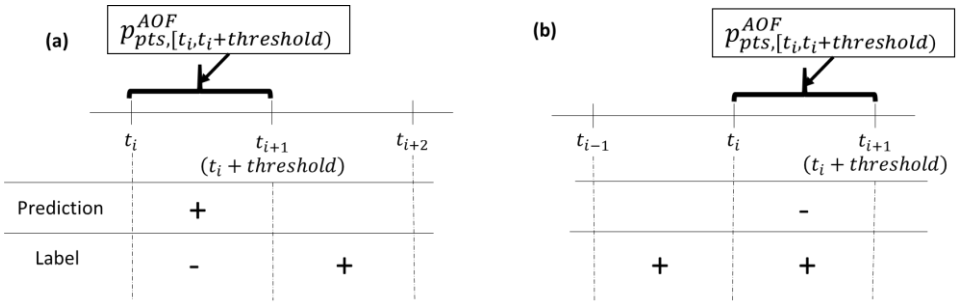

Figure 2.5. Illustration of (a) false-positive occurred one day before the target date and (b) false-negative instance-level predictions one day after the target date.

By assuming false-negative predictions the day after the target date and false-positive predictions the day before the target date as clinically valid predictions, we reevaluated the models by changing the evaluation criteria. First, we evaluated the performance after accepting false-positives occurred one day before the target date as true-positive predictions (Table 2.10, "Accepting FP" column). This yielded F1 scores that were an average of 5.24% higher (min: 0% in the ALF AOFI model from the UW-CDR dataset; max: 16.36% in the AHF AOFI model from the UW-CDR dataset). We also evaluated the performance change by accepting false-negatives occurring one day after the target date as true-negative predictions (Table 2.10, "Accepting FN" column). This yielded F1 scores that were an average of 20.34% higher (min: 7.92% in the ALI AOFI model from the MIMIC-3 dataset; max: 40.19% in the ALF AOFI from the MIMIC-3 dataset). When combined (Table 2.10, "Accepting Both" column), we observed F1 scores that were an average of 26.39% higher compared to the F1 scores evaluated by the annotation criteria (min: 8.90% in the ALI AOFI model from the MIMIC-3 dataset; max: 45.10% in the AHF AOFI model from the UW-CDR dataset). Under the assumption that such false-positives and false-negatives are clinically valid predictions, the actual performance of AOFI models as a risk estimator for acute organ failure onset could be higher than what we evaluated according to the annotation criteria in Table 2.6. Changes in F1 scores, by accepting these false predictions as true predictions, are provided in Table 2.10.

	MIMIC-3				**UW-CDR**			
Positive Criteria	**Original Criteria**	**Accepting FP**	**Accepting FN**	**Accepting Both**	**Original Criteria**	**Accepting FP**	**Accepting FN**	**Accepting Both**
AHF	0.2518	0.2636	0.3010	0.3146	0.2891	0.3364	0.3631	0.4195
ALI	0.6758	0.6823	0.7293	0.7360	0.4858	0.5105	0.5752	0.6027
AKI	0.4569	0.4881	0.5184	0.5524	0.4583	0.4836	0.5271	0.5548
ALF	0.2202	0.2257	0.3087	0.3160	0.0906	0.0906	0.1111	0.1111

Table 2.10. Reevaluated F1 scores with modified criteria.

2.3.4.2. Patient-level Analysis

During the patient-level error analysis, we observed that an average of 57.49% of false-positive patients were discharged with other acute organ failures on both datasets (Table 2.11). This indicated that some physiological changes that resulted from non-target acute organ failures might increase predicted risks in the target AOFI models. For example, it is known that AKI has a distant effect on other organ systems [74], [75], and ALI aggravates hepatic functionality [76]. Therefore, patients discharged with ALI and AKI might present with an adverse physiology of ALF during their hospital stay, thereby yielding high predicted probabilities of ALF, despite the fact that these patients were not discharged with ALF.

	MIMIC-3				UW-CDR			
Predicted discharge status	**AHF**	**ALI**	**AKI**	**ALF**	**AHF**	**ALI**	**AKI**	**ALF**
# of FP	79	120	68	37	15	79	72	35
Discharge Status of False Positive Patients								
No AOF	17	73	30	4	4	65	47	10
AHF		15	15	2		4	7	4
ALI	44		32	26	11		24	22
AKI	57	40		30	11	12		20
ALF	10	7	2		3	2	1	

Table 2.11. The number of frequently observed discharge diagnoses from false-positive patients in both the MIMIC-3 and UW-CDR datasets; a patient can be discharged with more than one acute organ failure.

2.4. Limitations

This study is based on clinical observations from ICUs, where the physiology of patients is worse compared to overall patient population. Moreover, the two datasets used for the study were collected from only three tertiary hospitals, so it is possible that each dataset reflects hospital-specific characteristics, which yielded different characteristics on annotation criteria between the MIMIC-3 and UW-CDR datasets. As a result, the study population and annotation criteria may not reflect general patient characteristics.

Moreover, the proposed annotation strategy relies on automatically selected interventions to annotate the potential acute organ failure onset. However, some of the measures that manage acute organ failure might be documented outside of the data sources considered in this study (e.g., participating in late-phase clinical trials). Therefore, this approach might miss some of the potential acute organ failure onset cases with other identifiable precursors. Also, the medical coding criteria might vary not only by medical coders but also by institution. Although we showed the feasibility of using the timing of intervention administration and discharge diagnoses as a simple annotation strategy for acute organ failure onset, this annotation strategy might rely too heavily on institution-specific diagnosis coding practices and treatment guidelines.

2.5. Conclusion

In this study, we demonstrated that statistical testing with discharge diagnoses was able to systematically identify clinical interventions that were frequently administered to patients discharged with the target acute organ failure. Moreover, by using the identified clinical interventions and the target discharge diagnoses as labels, trained models were able to show that the probability of receiving relevant interventions in the near future and being discharged with the target diseases could be used as a risk estimate for developing the target acute organ failure in the near future. Our approach produced reasonable prediction accuracies, particularly from the

ALI and AKI AOFI models across two datasets. Moreover, our error analyses indicated that false-positive and false-negative predictions in AOFI models that performed worse than other AOFI models, AHF and ALF AOFI models in both datasets, still might be clinically valid so the performance of predicted probabilities as an acute organ failure risk estimate could be higher than what was presented in this study. The conducted experiments also demonstrated that inferred probabilities tend to be well-aligned with known lab tests which are used to diagnose acute organ failure onset in practice. The presented automatic annotation strategy was able to derive risk prediction models for selected acute-onset diseases without human annotators based on the transparent annotation criteria, which can be further refined following physician evaluation.

Chapter 3. Extending Expert Scores of Patient Risk with Probabilistic Temporal Models

3.1. Introduction

Forecasting patient progress and improving patient outcomes by using predictive analysis has been one of the widely studied research areas in medicine, and early warning scores (EWS) aim to quantify patients' risk of developing clinical events of interest (e.g., sepsis) in advance [25], [77]–[79]. With frequently measured clinical variables taken from existing clinical workflows, these scores use a simple calculation to quickly evaluate a patient's risk with only a minimal chance of calculation error in time-critical care settings. These scores are proposed by practitioners, and they are designed as clinical decision support tools. For example, when high risk scores are observed, practitioners may initiate the order of high sensitivity diagnostic tests or increase the frequency of patient monitoring. Moreover, because these scores often require only a few clinical variables, the calculated risks are easy to compute and to interpret by caregivers.

EWS summarize multiple physiological variables into a few quantities that are aimed at providing a measure of the likelihood of clinical events of interest. Therefore, they are often used to screen high-risk patients for a target event by creating alerts when the patient's EWS exceed a predefined threshold [36], [37]. When alerted, the screening often involves additional reviews from physicians, which are performed by verifying risk factors of the target event. Most EWS only focus on the specific clinical event and do not consider the temporal trends of physiology due to their static nature. Therefore, this procedure could fail to generate proper alerts when 1) the patient is close to, but does not meet, the screening criteria for extended periods; that is, the patient is still developing the target disease because of the sustained abnormal physiology or 2) the EWS for the target event cannot be calculated because the required variables are undocumented. To cover the cases above and improve the performance of EWS-based screening, we can leverage these expert-driven risk estimates while accounting for aspects that are not considered by the criteria.

In the present study, we aimed to improve the screening performance of the widely adopted EWS that quantifies the general risk of organ failures, the Multiple Organ Dysfunction Score (MODS)[25], to identify patients with a high risk of each of the four acute organ-system failure (AOF) onsets: acute heart failure (AHF), acute lung injury (ALI), acute kidney injury (AKI), and acute liver failure (ALF). We hypothesized that the performance of EWS-based screenings can be improved by considering EWS on potentially relevant clinical events as well as their trajectories along with the EWS on the target event. To do this, we used the Hidden Markov Model (HMM) to integrate the trajectories of MODS subscores on the four target organ systems (cardiovascular, respiratory, renal, and hepatic) to generate risk estimates for the target AOF onset during a patient's intensive care unit (ICU) stay. Gold standard information about the timing of each AOF onset is not available from most of the electronic health records (EHRs). Thus, we trained the model based on patients' discharge statuses. In our approach, each latent state in the HMM was designed to be mapped to disjoint AOF risk states so that the likelihood of each latent state could be utilized as the risk estimates for the corresponding AOF risk states. After model training, we conducted a comparative analysis of the performance between the threshold-based screening using the MODS and the proposed method. We also analyzed how the trained HMM describes the general AOF prognoses on the dataset.

3.2. Related Work

3.2.1. Classifier Fusion

Classifier fusion is an ensemble approach that merges the predictions from submodels through algebraic combinations or a probabilistic framework. To do so, classifier fusion often treats predictions from each submodel as a support or a conditional posterior probability for each class [80], [81]. We designed our approach as a classifier fusion regarding the simple threshold-based screening criteria by using MODS for each organ system as weak classifiers evaluating patients' risk of AOF onsets. We hypothesized that classifier fusion could reduce potential false negative predictions when 1) the MODS for the target organ system was not available given the absence of required variables or 2) the MODSs were lower than the threshold from patients discharged with the target AOF. In such cases, calculated scores on other relevant organ systems based on the available clinical variables were assumed to be used as additional evidence via classifier fusion.

3.2.2. HMM for Clinical Event Prediction

For the current study, HMM was selected as the framework for integrating the calculated MODS trajectories during a patient's ICU stay. In general, HMM assumes the temporal transition among latent states and the sequence of observations as emitted evidence from these latent states. Because the model simplifies temporal state transitions and provides a probabilistic description of its behavior, HMMs have been widely used to illustrate patient prognoses [82]–[85]. Moreover, certain studies have leveraged the descriptive capacity of HMM to extract temporal physiological patterns that are frequently observed from target events. For example, Sukkar et al. [86] trained an HMM to derive latent representations on time-series biomarker observations related to Alzheimer's disease with unsupervised training, and analyzed the disease progression as described through the estimated likelihood of latent states. We also leveraged the descriptive power of HMM to understand how the model describes patients' prognoses for the four AOFs. By mapping each latent state to disjoint AOF risk states, we expected the HMM's probabilistic components to facilitate the clinical evaluation, such as evaluating the general transition patterns among AOF risk states, which might be challenging with other sequence models based on neural architectures.

3.3. Methods

3.3.1. Background

In this study, we aimed to improve the performance of AOF onset screening in the ICU setting using the widely adopted EWS, the MODS. The MODS quantifies a patient's general severity for each of the following organ systems based on relevant clinical variables: the cardiovascular, respiratory, renal, hepatic, hematologic, and neurologic systems[3]. The severity score for each organ system ranges from 0 to 4, with 4 being assigned to the most severe state. For the baseline screening criteria of the target AOF, we assumed that the MODS of the corresponding organ system was used to screen high-risk patients (e.g., using cardiovascular MODS for the AHF screening). The detailed MODS scoring criteria are provided in Table 3.1.

Organ System	Severity Score				
	0	1	2	3	4
Cardiovascular (PAR*)	< 10.1	10.1 − 15.0	15.1 − 20.0	20.1 − 30.0	> 30
Respiratory (PaO_2/FiO_2)	> 300	226 − 300	151 − 225	75 − 150	≤ 75
Renal (Creatinine,μmol/L)	< 100	101 − 200	201 − 350	351 − 500	> 500
Hepatic (Bilirubin,μmol/L)	≤ 20	21 − 60	61 − 120	121 − 240	> 240

Table 3.1. MODS for the cardiovascular, respiratory, renal and hepatic systems [25]; * Pressure-adjusted Heart Rate (PAR) = [Heart Rate] × [Central Venous Pressure]/[Mean Arterial Pressure]

3.3.2. Prediction Problem

The HMM for the presented study was designed to take the MODS history (#1 in Figure 3.1), including the evaluation time window (t_i in Figure 3.1), and estimate the risk of the target AOF onset during the evaluation time window. For each day, the trained HMM generated the risk estimates for the onset of AHF, ALI, AKI, and ALF based on the patient's MODS history.

The HMM modeled each patient's prognosis regarding the four AOFs. To describe the model, we assumed that there exists a latent state that holds the true information of the four target AOF onsets not directly available from the EHR, and we considered the calculated MODS for the patient's cardiovascular, respiratory, renal, and hepatic systems as *evidence emitted by the latent states*. To directly compare the performance changes due to trajectory consideration of relevant organ systems with the threshold-based baseline, we did not include the hematologic and neurologic MODS for this study. Moreover, to measure the contribution of trajectory information compared with the latest evidence observed at time t_i during the likelihood estimation, we merged risk estimates from trajectory information and the latest evidence (#2 in Figure 3.1) with different weights.

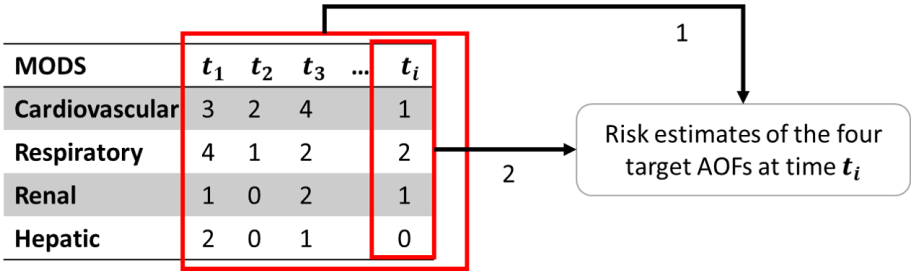

Figure 3.1. HMM design evaluating the AOF risks at the evaluation time window t_i; 1. Trajectory information from time t_1 to t_i, 2. Latest evidence at time t_i.

3.3.3. Input

To train the HMM, calculated MODS subscores for the four organ systems—cardiovascular, lung, renal, and hepatic systems—were used as an input. To provide a general description of the design, we noted the number of organ systems as *k=4*. We denoted the calculated MODS on the j-th organ system at time t_i as $risk_{j,t_i} = MODS_j(x_{pts,[t_i,t_i+1)})$, which took patients' clinical observations from time t_i to t_{i+1}, $x_{pts,[t_i,t_i+1)}$, and calculated the highest MODS, the worst potential patient risk, for the target organ system. The vector representing the calculated MODSs

for all organ systems during $[t_i, t_{i+1}]$ was noted as $risk_{t_i} =< risk_{1,t_i}, \dots, risk_{k,t_i} > \in \{0,1,2,3,4\}^k$ for brevity.

3.3.4. Target Labels

The HMM took the discharge diagnoses as gold standard information for training and evaluation.

We noted the terminal outcomes for the patient *pts* according to the discharge diagnoses available in the EHR as $terminaloutcome_{pts} =< terminal_1, \dots, terminal_k >$, where $terminal_j \in \{0,1\}$ for all $j \in \{1, \dots, k\}$: $terminal_j$ was a binary variable indicating the patient's discharge status on the j-th AOF. For each $terminal_j$, a positive label was assigned when the patient was discharged with the target AOF, and a negative label was assigned otherwise.

3.3.5. Trajectory Modeling

The HMM modeled patients' AOF prognoses by considering MODSs at time t_i for *k* target organ systems, $risk_{t_i}$, as emitted evidence from latent states during the evaluation time window $[t_i, t_{i+1}]$. Each latent state $state_{t_i} \in StateSet$ was assumed to have information about the true onset of AOFs during $[t_i, t_{i+1}]$, which was not available from the dataset. We assigned disjoint AOF states, comprising *k* binary variables indicating each AOF onset, to each latent state, thereby generating $2^k (= |StateSet|)$ possible latent states for each evaluation time window.

To describe the HMM, we estimated the prior probabilities of each latent state, the transition probability between latent states, and the likelihood of observing $risk_{t_i}$ given each state $s \in StateSet$. We noted the prior probability of each *state* as *P*(*state*), and optimized the vector $\overrightarrow{prior}, \overrightarrow{prior} =< prior_{StateSet[1]}, \dots, prior_{StateSet[2^k]} > \in [0,1]^{2^k}$ where $\sum_{s \in StateSet} prior_s = 1$, to estimate the probabilities. The transition matrix was noted as $TM \in [0,1]^{2^k \times 2^k}$, and $TM[from, to]$ represents the transition probability from $StateSet[from]$ to $StateSet[to]$, $P(StateSet[from] \rightarrow StateSet[to])$, where $\sum_{l \in StateSet} TM[from, l] = 1$. The parametric function f_ψ estimating the likelihood of $risk_{t_i}$ for each *state* was implemented using a neural network with a single hidden layer to meet the following property: $P(risk_{t_i} | state) = f_\psi(risk_{t_i})$ where $f_\psi: \{0, \dots, 4\}^k \rightarrow [0,1]^{2^k}$.

Using the probabilistic components above, the HMM estimated the likelihood of each $state_{t_i}$ given the emitted evidence from t_1 to t_i, $P(state_{t_i} | risk_{t_1}, \dots, risk_{t_i})$, as follows:

$$P(state_{t_i} | risk_{t_1}, \dots, risk_{t_i}) = \frac{P(risk_{t_1}, \dots, risk_{t_i} | state_{t_i}) P(state_{t_i})}{\sum_{s \in StateSet} P(risk_{t_1}, \dots, risk_{t_i} | state_{t_i} = s) P(state_{t_i} = s)}$$

Under the Markov assumption, we could rewrite $P(risk_{t_1}, \dots, risk_{t_i} | state_{t_i})$ as follows:

$$P(risk_{t_1}, \dots, risk_{t_i} | state_{t_i}) = \sum_{s \in StateSet} \begin{bmatrix} P(risk_{t_1}, \dots, risk_{t_{i-1}} | state_{t_1}, \dots, state_{t_{i-1}} = s) \\ \times P(state = s) P(s \rightarrow state_{t_i}) P(risk_{t_i} | state_{t_i}) \end{bmatrix}$$

For the calculation, $P(risk_{t_1}, \dots, risk_{t_i - 1} | state_{t_1}, \dots, state_{t_{i-1}} = s)$ was estimated using recursion, and $P(state = s)$, $P(s \rightarrow state_{t_i})$, and $P(risk_{t_i} | state_{t_i})$ were estimated using the

prior probability ($P(state)$), the transition matrix (*TM*), and the likelihood estimator (f_ψ), respectively.

During the likelihood estimation, we assumed that some *state*s could be predicted more accurately with trajectory information while other *state*s relied more on the latest evidence. To measure how the importance of trajectory information varies for the likelihood estimation on each state, we combined two likelihood estimates for $state_{t_i}$—one using latest evidence, $P(state_{t_i}|risk_{t_i})$, and another using trajectory information, $P(state_{t_i}|risk_{t_1}, \dots, risk_{t_i})$—as a weighted sum with $\alpha_{state_{t_i}} \in [0,1]$:

$$P_{total}(state_{t_i}|risk_{t_1}, \dots, risk_{t_i}) = \left(1 - \alpha_{state_{t_i}}\right) P(state_{t_i}|risk_{t_i}) + \alpha_{state_{t_i}} P(state_{t_i}|risk_{t_1}, \dots, risk_{t_i})$$

3.3.6. Model Optimization

Although the forward-backward [87] and Viterbi algorithms [88] are common approaches for HMM training, they were not applicable in the current study because the labels for the predicted *state*s were not available. Instead, we compared the estimated state likelihoods from the time of admission (t_{adm}), $P_{total}(state_{t_{adm}}|risk_{t_{adm}})$, to the time of discharge (t_{disch}), $P_{total}(state_{t_{disch}}|risk_{t_{adm}}, \dots, risk_{t_{disch}})$, with the binary labels in $terminaloutcome_{pts}$ by aggregating the estimated state likelihood sequence into a scalar. We then optimized the model with convex optimization. During the probability aggregation, the state without any AOF onset, $ZeroState = \{0\}^k$, was treated separately because the patients who were discharged with no AOF should remain in this $ZeroState$ throughout their ICU stay; otherwise, at least one AOF would be documented in their discharge diagnoses. The likelihood of patients being assigned to the latent state $s \in StateSet$ at least once throughout their ICU stay was calculated as follows:

$$\beta_{pts,s} = \begin{cases} 1 - \left[\prod_t \left(1 - P_{total}(state_t = s|risk_{t_{adm}}, \dots, risk_t)\right)\right] & if\ s \neq ZeroState \\ \prod_t P_{total}(state_t = s|risk_{t_{adm}}, \dots, risk_t) & if\ s = ZeroState \end{cases} \in [0,1]$$

Because the prevalence of each AOF varied, the AOF with comparably lower prevalence, such as ALF, would be presented with an even lower prevalence if we expanded $terminaloutcome_{pts}$ on the outcome basis, $|terminaloutcome_{pts}| = k$, into the state basis, $|StateSet| = 2^k$. Therefore, we compared β_{pts} and $terminaloutcome_{pts}$ on the outcome basis by further aggregating β_{pts} into $\gamma^+_{pts,j}$ and $\gamma^-_{pts,j}$, in which each quantified the probability of the patient being assigned to either the AOF-positive or AOF-negative states on the j-th AOF at least once throughout their ICU stay, respectively:

$$\gamma^+_{pts,j} = 1 - \left(\prod_{s \in \{\tilde{s} | \tilde{s} \in StateSet, \tilde{s}[j] = 1\}} \left(1 - \beta_{pts,s}\right)\right),$$

$$\gamma^{-}_{pts,j} = 1 - \left(\prod_{s \in \{\tilde{s} | \tilde{s} \in StateSet, \tilde{s}[j] = 0\}} (1 - \beta_{pts,s}) \right)$$

Finally, the following loss function was used to train the HMM:

$$loss(\lambda, \psi) = \sum_{i \in pts} \sum_{j \in \{1,\dots,k\}} \begin{bmatrix} CrossEntropy\left(\gamma^{+}_{i,j}, terminaloutcome_i[j], weight[j]\right) \\ +CrossEntropy\left(\gamma^{-}_{i,j}, 1 - terminaloutcome_i[j], weight[j]^{-1}\right) \\ +regularizer(\lambda) \end{bmatrix},$$

where the weighted cross entropy $CrossEntropy(p, y, \phi)$ was calculated on the probability p, the label y, and the weight on the positive prediction for the class φ. Tensorflow library [89] was used for the implementation.

3.3.7. Evaluation

To evaluate our approach, we predicted the patient's discharge status based on state predictions P_{total} throughout the patient's ICU stay. By aggregating the positive predictions on the predicted state $\widehat{state}_{t_i} = \left[argmax_s P_{total}\left(state_{t_i} = s | risk_{t_{adm}}, \dots, risk_{t_i}\right)\right] \in \{0,1\}^k$, we defined the predicted discharge state for each patient as follows:

$$DischState_{pts} = < \max_{t \in \{t_{adm},\dots,t_{disch}\}} \widehat{state}_t[1], \dots, \max_{t \in \{t_{adm},\dots,t_{disch}\}} \widehat{state}_t[k] >$$

Then, we compared this with $\text{terminaloutcome}_{\text{pts}}$ on the outcome basis. For example, for the predicted $\text{DischState}_{\text{pts}} = < 1,0,0,1 >$ and $terminaloutcome_{pts} = < 1,1,0,1 >$, *pts* was counted as a true-positive during the evaluation on $terminal_1$ and a false-negative during the evaluation on $terminal_2$ and so on. From the models trained with various hyperparameters, we selected the model that achieved the highest micro-F1 score in the validation set because we wanted to minimize the total number of false AOF predictions.

Using the test set, the screening performance of the selected model was evaluated based on three different deployment settings: 1) MODS only (MODS), 2) HMM only (HMM), and 3) MODS and HMM together (MODS+HMM). For each setting, we assumed a patient was called out for additional review when: 1) the MODS for the target organ system was ≥ 3, 2) the HMM made a positive prediction for the target AOF, and 3) either the HMM made a positive prediction or the MODS for the target organ system was ≥ 3, respectively.

3.4. Results

We used the MIMIC-3 dataset [65] in the current study. The dataset comprises the clinical observations of patients admitted to ICUs at the Beth Israel Hospital in Boston, MA. From the dataset, we included patients who were found to have at least one of the required clinical variables to calculate their MODSs during their ICU stays. Patients under the age of 18 were excluded because the normal range of physiological variables differs between children and adults. A brief summary of the patient demographics was presented in Table 3.2.

Descriptor	
Number of hospital admissions	27,769
Number of patient-days	160,980
Age*	63.73 ± 17.30
Admission type: Number of hospital admissions	Elective — 4,002 Emergency — 22,777 Urgent — 990
Number of in-hospital deaths	3,492 (12.57%)

Table 3.2. Patient demographics from the dataset; * the ages of 2,616 patients older than 90 years were randomly adjusted for de-identification purposes.

To train and test the HMM, we divided a total of 27,769 hospital admissions (a total of 160,980 patient-days) into training, validation, and test sets at a 7:2:1 ratio. For the terminal outcomes, we considered the four different AOFs: AHF, ALI, AKI, and ALF. For the terminal states terminal_j of each AOF, we assigned a positive label when a patient was discharged with the target AOF and a negative label otherwise. Selected ICD-9 (International Classification of Diseases, 9th revision) codes for each AOF terminal outcome labeling were presented in Table 3.3.

For the HMM training, the MODSs for the four organ systems on each patient-day were considered as evidence emitted from the latent states, which are assumed to hold information about true AOF onsets. We assigned 0 MODS to an organ system when the score could not be calculated because at least one of the required variables was not available on the patient-day from the EHR. The overall distribution of MODSs in the test set was presented in Table 3.4. We described the results using predictions of the patient's terminal outcome ($DischState_{pts}$; patient-level) and predictions of each patient-day ($\widehat{\text{state}}_{t_1}$; instance-level).

Target Acute Organ Failure	ICD-9 Diagnosis Code	
AHF	428.21	Acute systolic heart failure
	428.23	Acute on chronic systolic heart failure
	428.31	Acute diastolic heart failure
	428.33	Acute on chronic diastolic heart failure
	428.41	Acute combined systolic and diastolic heart failure
	428.43	Acute on chronic combined systolic and diastolic heart failure
ALI	518.81	Acute respiratory failure
	518.51	Acute respiratory failure following trauma and surgery
	518.84	Acute on chronic respiratory failure
	518.53	Acute on chronic respiratory failure following trauma and surgery
AKI	584.9	Acute kidney failure, unspecified
	584.6	Acute kidney failure with lesion of cortical necrosis
	584.7	Acute kidney failure with lesion of medullary necrosis
	584.5	Acute kidney failure with lesion of tubular necrosis
	584.8	Acute kidney failure with specified pathology NEC
ALF	570	Acute and subacute necrosis of liver

Table 3.3. Selected ICD-9 codes for terminal outcome labeling.

Maximum score during ICU stay	**Discharged with AOF**				**Population**			
MODS on the target organ	**AHF**	**ALI**	**AKI**	**ALF**	**Cardiovascular**	**Respiratory**	**Renal**	**Hepatic**
0	63	77	140	7	1,737	1,461	1,685	682
1	1	21	182	0	79	125	654	241
2	2	66	130	3	158	293	213	1,003
3	5	140	56	14	304	596	74	508
4	19	125	50	28	494	297	146	338
Total	90	429	558	52		2,772		

Table 3.4. Per-patient maximum MODS distribution in the test set.

3.4.1. Performance Comparison on Patient-level Predictions

The patient-level prediction performance is provided in Table 3.5. The HMM showed a higher micro-F1 score (0.374 vs. 0.225) with a significantly improved micro-recall (0.661 vs. 0.387) as compared to the MODS setting, which indicated that considering the MODS trajectories of the target and potentially relevant organ systems yielded more accurate predictions in high-risk patient screening tasks overall. The performance improvement was mainly achieved by correctly identifying high-risk patients who were not detected in the MODS setting. The MODS+HMM

setting also showed a higher micro-F1 score as compared with the MODS setting (0.299 vs. 0.225), which showed the potential of the HMM to complement the threshold-based screening when it cannot be deployed by itself.

Deployment Setting	MODS			HMM			HMM+MODS		
Performance	Rec	Prec	F1	Rec	Prec	F1	Rec	Prec	F1
AHF	0.267	0.030	0.054	0.411	0.076	0.128	0.644	0.048	0.090
ALI	0.618	0.297	0.401	0.620	0.256	0.362	0.786	0.239	0.366
AKI	0.190	0.482	0.273	0.753	0.374	0.500	0.755	0.367	0.494
ALF	0.808	0.050	0.094	0.442	0.109	0.175	0.808	0.049	0.093
Micro measures	0.387	0.159	0.225	0.661	0.261	0.374	0.760	0.186	0.299

Table 3.5. Patient-level performance comparison for different deployment settings.

With respect to the performance for each AOF onset screening, AHF, AKI, and ALF showed the best F1 scores of 0.128, 0.500, and 0.175 in the HMM setting, respectively, while ALI showed the best performance in the MODS setting (F1: 0.401). The ALI screening result from the HMM did not show a higher F1 score and presented a similar recall with decreased precision as compared to the MODS setting. There could be two explanations for this observation. First, considering the MODS trajectories on other organ systems might not provide additional information for the high-risk ALI patient screening task. Compared to the MODS of other organ systems, the respiratory MODS showed more suitable characteristics as a screening tool; it showed a comparably higher F1 score with a higher recall in the MODS setting. Specifically, 61% of patients discharged with ALI had a respiratory MODS greater than or equal to 3 although only 891 patient-days (20.14% from patients discharged with ALI; 5.79% from all test patients) were called out during the MODS-based ALI screening.

Second, the HMM might make a positive prediction for patients at a high risk of developing ALI even though they did not develop the disease during their ICU stays. Among the 1,715 patients classified as true-negative patient-level ALI predictions in the MODS setting, 446 patients were classified as false-positive in the HMM. From those 446 false-positive patients, we observed at least one of the following diseases that are potentially relevant in ALI from 309 patients (69.28%) on their discharge diagnoses: unspecified congestive heart failure, unspecified essential hypertension, and unspecified kidney failure. We believe the decreased precision due to newly identified false-positive patients with the above discharge diagnoses might indicate that the HMM identified patients at a high risk of ALI as positive [90], even though they did not develop ALI during their ICU stays.

3.4.2. Coverage of Missed Patients from the MODS-based Screening

For the current study, we did not have gold standard information about the time of the target AOF onsets, but each patient's terminal outcomes were available from the discharge diagnoses in the EHR. Therefore, we evaluated instance-level predictions from the HMM, $\widehat{\text{state}}_{t_1}$, by focusing on the HMM's prediction from patients discharged with the target AOF under the following

conditions: 1) when 0 MODS was imputed due to the missing observation and 2) when their MODSs were lower than the threshold (< 3) throughout their ICU stays.

First, for instances (a patient-day) when the HMM made a positive prediction while 0 MODS was imputed for the target organ system, we examined how the next-day MODS of the target organ system was changed. To accomplish this, we analyzed instances where 1) a 0 MODS was imputed for the target organ system, 2) a positive prediction was made by the HMM on the corresponding AOF onset, and 3) the MODS for the target organ was calculated the day after with all the required variables available (Figure 3.2). For instances that met the aforementioned conditions, the next-day MODS on 33.32% of such instances were above the MODS-based screening threshold, ≥ 3 (Table 3.6). This indicates that the HMM identified these high-risk patients earlier than the MODS setting by using the MODS history of other organ systems even though the MODS of the target organ system was unavailable. Compared to other organ systems, the renal MODS showed that the majority of the next-day MODSs were lower than the threshold (< 3). Literature review revealed that the clinical community defines AKI onset by the relative increase in creatinine levels compared to the baseline level instead of the absolute level [68], which may reflect a different next-day MODS distribution from that in other organ systems. Moreover, the patient-level prediction results on AKI in Table 3.5 indicated the most improved recall as compared to other AOFs in the HMM setting while showing a higher F1 score when compared to the MODS baseline. Therefore, we suspect that renal MODS is sensitive, but not specific, for AKI screening.

MODS	...	t_i	t_{i+1}
Cardiovascular	...	0	4
Respiratory	...	1	...
Renal	...	3	...
Hepatic	...	2	...
HMM AHF Prediction	...	+	...

Figure 3.2. The imputed current-day and calculated next-day cardiovascular MODSs with the positive HMM prediction on AHF.

Calculated MODS	AHF	ALI	AKI	ALF
0	34	111	417	1
1	67	110	183	136
2	74	172	54	357
3	126	238	19	134
4	124	96	36	81

Table 3.6. Calculated next-day MODSs after the positive prediction by the HMM.

We also examined how well the HMM detects patients who were discharged with the target AOF but did not meet the MODS-based screening criteria during their ICU stays. The test set included 621 patients discharged with at least one of the four target AOFs with the maximum MODS below the threshold for the target organ system, thereby classifying them as false negatives in the MODS setting. Among these patients, the HMM correctly classified 379 patients as positives. To

compare the proportion of patients missed by the MODS baseline with the proportion of patients missed by the HMM, we calculated the false omission rate (FOR) by calculating the proportion of false-negative patients from all the negatively predicted patients. Table 3.7 shows that, overall, the FOR was lower in the HMM when compared to the MODS-based screening, while the ALF prediction did not show noticeable changes. As predictions from the HMM showed lower FORs with a higher micro-F1 score, we believe predictions from the HMM are more suitable as a high-risk patient screening tool compared to the MODS-based screening.

	AHF	**ALI**	**AKI**	**ALF**
# of patients (predicted negative in the MODS)	1,974	1,879	2,552	1,926
# of patients discharged with the AOF	66	164	452	8
FOR on the MODS	0.0334	0.0873	0.1771	0.0042
FOR on the HMM	0.0204	0.0676	0.0843	0.0042

Table 3.7. FOR for patients with the maximum MODS < 3.

3.5. Discussion

We interpreted MODS as estimates of the probability of AOF onsets. We employed an HMM to describe patients' risk trajectories for the four AOFs using a probabilistic framework, and showed a higher micro-F1 score in detecting patients discharged with the four AOFs when compared to the MODS-based screening. Moreover, the probabilistic components trained in the HMM allowed us to further analyze how the model described the prognoses of the four AOFs.

The learned α_s value for each state s quantifies to what extent each state's likelihood estimation relies on trajectory information when compared with the latest evidence. During the model implementation, we designed $\alpha_s \in [0,1]$ to be higher (≈ 1) when it fully relies on trajectory information. As presented in Table 3.8, the states with only one positive AOF onset showed a lower dependency on trajectory information ($\alpha_s = 0.3315 \pm 0.3861$) when compared with other states with two or more AOF onsets ($\alpha_s = 0.7893 \pm 0.3280$). Moreover, the *ZeroState*—the state without any AOF onsets—showed that the model relies on the latest evidence and trajectory information with similar weights. This illustrates that the risk estimation of more severe AOF states should take trajectory information into account more than the risk estimation of less severe AOF states.

$State$	<0,0,0,0>	<1,0,0,0>	<0,1,0,0>	<0,0,1,0>	<0,0,0,1>	
α_{state}	0.5308	0.0741	0.8863	0.3024	0.0632	
$State$	<1,0,0,1>	<1,0,1,0>	<0,1,0,1>	<1,1,0,0>	<0,1,1,0>	<0,0,1,1>
α_{state}	0.9877	0.9993	0.0003	0.9984	0.9463	0.9778
$State$	<1,1,1,0>	<1,1,0,1>	<1,0,1,1>	<0,1,1,1>	<1,1,1,1>	
α_{state}	0.9654	0.8371	0.9754	0.6195	0.3727	

Table 3.8. αstate on each state likelihood estimation; State: <AHF,ALI,AKI,ALF>.

The estimated transition probabilities between states, *TM*, allowed us to evaluate the overall cascading patterns of AOFs during patients' ICU stays in terms of the number of positively

predicted AOFs (Table 3.9). The objective of the trained HMM was to describe patients' AOF prognoses observed from ICUs. Therefore, it tended to predict more AOF onsets if no AOF was predicted on the previous day. This is because patients with comparably better prognoses are often transferred from an ICU to a stepdown unit that manages less severe patients because the overall management cost is expensive in ICUs [91]. Moreover, when all four AOF onsets were predicted as positive, the patient's recovery on the following day would be limited because the probability of transitioning to states with two or more AOF onsets was estimated as 0.7052.

Previous\Next	0	1	2	3	4
0	0.0103	0.4320	0.3967	0.1169	0.0442
1	0.0120	0.4537	0.2083	0.2959	0.0301
2	0.0088	0.5119	0.1358	0.2523	0.0912
3	0.0485	0.4226	0.2630	0.1102	0.1557
4	0.0054	0.2894	0.4430	0.2544	0.0078

Table 3.9. Transition trends regarding the number of predicted AOFs (Prev: Number of positively predicted AOFs at time t_i; Next: Number of positively predicted AOFs at time t_{i+1}).

Finally, there is medical consensus on the notion that patients' in-hospital mortality have a positive correlation with the number of organ failures developed during their ICU stay [92]. By comparing the in-hospital mortality with the number of positively predicted AOF onsets throughout the ICU stay, $DischState_{pts}$, we verified that the HMM predictions agreed with the consensus in showing increasing in-hospital mortality rate when more AOFs were predicted throughout a patient's ICU stay (Table 3.10a). We observed a similar trend between in-hospital mortality rate and the number of developed AOFs that were documented in discharge diagnoses (Table 3.10b).

(a) In-hospital mortality vs. number of predicted AOF throughout the ICU stay.

# of predicted AOF during ICU stay	# of patients	# of expired patients	In-hospital mortality
0	1,593	123	0.0772
1	57	4	0.0702
2	619	92	0.1486
3	446	121	0.2713
4	57	19	0.3333

(b) In-hospital mortality vs. number of documented AOF from discharge diagnoses.

# of documented AOF	# of patients	# of expired patients	In-hospital mortality
0	1,914	126	0.0658
1	622	136	0.2186
2	202	84	0.4158
3	33	13	0.3939
4	1	0	0.0000

Table 3.10. In-hospital mortality rate based on the number of (a) predicted AOF throughout the ICU stay, and (b) developed AOF from discharge diagnoses.

3.6. Case Study – Trajectory Model Training with Risk Estimates from AOFI Models

The analyses presented above showed that considering trajectory information improved the micro-F1 score of AOF prediction tasks compared to the baseline method—a threshold-based prediction using the MODS subscore for the target organ system. The performance improvement observed in the micro-recall was mainly attributed to improved micro-F1 score with similar micro-precision. This indicates that the model was able to identify cases that were missed in the baseline method while not significantly increasing the number of false-positive predictions. Therefore, the additional experiment was conducted to verify whether this trend could also be seen when estimates from AOFI models were considered as emitted evidence from latent states. For the experiment, the definition of the latent state and the terminal outcome were unchanged compared to the MODS experiment settings in previous sections.

3.6.1. Prediction Problem

The HMM aims to predict the risk of target acute organ failure onset during $[t_i, t_{i+1}]$ by considering the predicted risks from four AOFI models—AHF, ALI, AKI, and ALF AOFI models—as evidence emitted from latent states. The HMM used predicted risks from the day after the ICU admission, $risk_{t_{adm}+1}$, to time t_{i+1}, $risk_{t_i}$, as evidence (i.e., the estimated probability of j-th AOF onset $risk_{j,t_i}$ is now equal to $\mathrm{p}^{AOF[j]}_{\mathrm{pts},[\mathrm{t_i},t_{i+1})} = f_{AOF[j]}(x_{pts,[t_{i-1},t_i)})$, where $AOF[j]$ indicates the j-th AOF of the interest). Similar to the MODS experiment setting, four different acute organ failures (AHF, ALI, AKI, and ALF) were considered as the terminal outcome, which yielded 16 different latent state $state_{t_i}$ for each time window $[t_i, t_{i+1}]$.

Without modifying the HMM structure described in Section 3.3.5, we also examined whether changes in the likelihood estimator f_ψ would improve prediction performance. First, more complex neural networks were examined to evaluate whether the model could learn more patterns for the state likelihood estimation if additional hidden layers are trained. Second, we examined whether keeping the orders learned from the AOFI models would improve prediction performance or not. AOFI models, similar to the MODS, rely on myopic physiological evidence, which is patient physiology measured up to 24 hours before the time of prediction, when predicting the target AOF onset. We hypothesized that if a higher likelihood of target events (i.e., proxy events which receive relevant clinical interventions and being discharged with the target AOF) would indicate a higher risk of the target AOF onset, constraining the predictions of the HMM to follow the orders of the AOFI model estimates would improve prediction performance. For example, if one instance had higher risk levels from the AHF AOFI model than another instance, the former instance should also provide a higher or equal likelihood of states with AHF onset, regardless of the status of other AOFs. To implement this condition, submodularity, on the proposed modeling framework, only non-negative values were used as weights on the neural network.

3.6.2. Results

To directly compare performance between the baseline model, risk estimates from the four AOFI models, and the HMM, we used the same training, validation, and testing patients as Aim 1. During the Aim 1 experiment, we excluded groups of patients in the training dataset who were discharged with the target AOF but did not receive any of the relevant interventions. Therefore, there was no estimated probability of instances from these patients, and we only considered instances from patients who had estimated probabilities from all four AOFI models, thereby yielding fewer training instances compared to the Aim 1 experiment (15,855 hospital admissions to 5,611 from MIMIC-3;10,445 hospital admissions to 7,793 from UW-CDR).

The best model was selected when the model achieved the highest micro-F1 score on the validation set based on patient-level predictions, and the model trained by the single hidden layer with submodularity constraint was selected. This indicates that the HMM showed the best performance when assuming higher risk estimates from the AOFI model present a higher likelihood of latent states with the target AOF onset, while increased complexity was not able to learn additional information for the likelihood estimation. As Table 3.11 shows, the HMM presented higher micro-recall compared to the AOFI models for patient-level predictions with lower micro-precision in both datasets. For individual AOF prediction task, higher recalls and lower precisions were also observed in both datasets compared to the predictions from AOFI models.

(a)

	Original AOFI			HMM w/ AOFI		
MIMIC-3	Prec	Rec	F1	Prec	Rec	F1
AHF	**0.3780**	0.2211	**0.2791**	0.2608	**0.5324**	0.2608
ALI	**0.6552**	0.6080	**0.6307**	0.4717	**0.8000**	0.4717
AKI	**0.7302**	0.3644	0.4861	0.4928	**0.6152**	**0.4928**
ALF	**0.2745**	0.3111	**0.2917**	0.1702	**0.3556**	0.1702
Micro	**0.6093**	0.4151	0.4938	0.4114	**0.6502**	**0.5039**

(b)

	Original AOFI			HMM w/ AOFI		
UW-CDR	Prec	Rec	F1	Prec	Rec	F1
AHF	**0.3600**	0.1765	**0.2369**	0.2778	**0.1961**	0.2299
ALI	**0.5434**	0.5529	**0.5481**	0.4435	**0.6236**	0.5184
AKI	**0.5313**	0.5152	**0.5231**	0.3821	**0.5697**	0.4574
ALF	**0.2000**	0.3000	**0.2400**	0.1028	**0.3667**	0.1606
Micro	**0.4889**	0.4736	**0.4811**	0.3519	**0.5313**	0.4234

Table 3.11. Performance comparison of patient-level predictions between AOFI models and HMM; (a) MIMIC-3 and (b) UW-CDR.

To examine whether these changes were due to the changes in distribution, or in the decision boundary, we adjusted the threshold used in the AOFI patient-level predictions to match the recall achieved by the HMM models, then compared the precision after the adjustment. Moreover, in order to quantify the disagreement of the patient-level predictions for each patient, we calculated a kappa score for each target AOF. After the adjustment, the HMM trained with the MIMIC-3 dataset showed a similar precision level compared to the AOFI models, while the model trained with UW-CDR dataset still showed a lower precision level (Table 3.12). Moreover, we observed higher kappa score from AOFs with higher performance in the AOFI model (ALI and AKI) compared to the AOFs with lower performance (AHF and ALF), which indicates more disagreement were observed in AOFs that showed lower performance from AOFI models, thereby indicating more changes were made for AHF and ALF prediction from the HMM compared to the ALI and AKI predictions. In addition, patient-level predictions tend to agree more in patients discharged with the target AOF, except for AKI from the MIMIC-3 dataset. Lastly, the HMM made fewer positive instance-level predictions in all AOFs but it also achieved higher recalls in patient-level predictions in all AOFs from both datasets.

(a) MIMIC-3

	Adjusted AOFI			HMM w/ AOFI			Kappa score			# pos predicted instances; AOFI vs. HMM*
MIMIC-3	Prec	Rec	F1	Prec	Rec	F1	All	Discharged w/ AOF	Not Discharged w/ AOF	
AHF	0.2342	0.5324	0.2342	**0.2608**	0.5324	**0.2608**	0.55	0.52	0.52	971 vs. 719
ALI	**0.4769**	0.8000	**0.4769**	0.4717	0.8000	0.4717	0.89	0.92	0.83	2,661 vs. 2,933
AKI	0.4752	0.6152	0.4752	**0.4928**	0.6152	**0.4928**	0.68	0.59	0.63	1,548 vs. 1,278
ALF	**0.2353**	0.3556	**0.2353**	0.1702	0.3556	0.1702	0.44	0.61	0.37	334 vs. 131
Micro	0.4023	0.6502	0.4971	**0.4114**	0.6502	**0.5039**	N/A			5,514 vs. 5,063

(b) UW-CDR

	Adjusted AOFI			HMM w/ AOFI			Kappa score			# pos predicted instances; AOFI vs. HMM**
UW-CDR	Prec	Rec	F1	Prec	Rec	F1	All	Discharged w/ AOF	Not Discharged w/ AOF	
AHF	**0.3226**	0.1961	0.2439	0.2778	0.1961	0.2299	0.65	0.88	0.55	207 vs. 132
ALI	**0.4818**	0.6235	0.5436	0.4435	0.6236	0.5184	0.67	0.72	0.53	1,344 vs. 1,024
AKI	**0.4772**	0.5697	0.5194	0.3821	0.5697	0.4574	0.70	0.75	0.60	1,175 vs. 1,070
ALF	**0.1964**	0.3667	0.2558	0.1028	0.3667	0.1606	0.50	0.56	0.47	196 vs. 144
Micro	**0.4385**	0.5313	0.4805	0.3519	0.5313	0.4234	N/A			2,922 vs. 2,370

Table 3.12. Performance comparison between the HMM predictions and threshold-adjusted AOFI predictions on patient-level; (a) MIMIC-3 dataset, (b) UW-CDR dataset. *The number of positively predicted instance-level predictions before the threshold adjustment.

To conduct clinical validation on instance-level predictions, we compared the summary statistics of lab tests frequently used to confirm the AOF onset for positive and negative instance-level predictions by using the framework introduced in Section 2.3.3.1 (See Figure 2.3). As expected, Table 3.13 showed worse prognoses when the HMM made a positive instance-level prediction on the target AOF in both datasets, except for the pCO2 level on ALI in the UW-CDR dataset (average pCO2 level was slightly higher in the patients predicted negative).

(a) MIMIC-3

MIMIC Lab Test	Instance-level Prediction	HMM			AOFI			Target AOF
		Average	Std. Dev	# Obs	Average	Std. Dev	# Obs	
NTproBNP	0	8424.59	11702.90	55	7402.82	11027.56	52	AHF
NTproBNP	1	9897.00	5666.75	3	19435.60	11783.58	6	AHF
pCO2	0	41.27	10.92	3612	41.20	10.71	3927	ALI
pCO2	1	41.96	9.98	5815	42.05	10.07	5500	ALI
pO2	0	117.16	63.77	3614	116.87	63.29	3929	ALI
pO2	1	111.73	46.55	5815	111.63	45.84	5500	ALI
Urea Nitrogen	0	29.82	25.37	8905	27.32	22.80	8143	AKI
Urea Nitrogen	1	45.25	29.31	1691	48.75	31.43	2453	AKI
Total Bilirubin	0	3.80	6.77	2479	3.29	6.21	2301	ALF
Total Bilirubin	1	10.92	14.90	84	10.55	11.59	262	ALF

(b) UW-CDR

UW-CDR Lab Test	Instance-level Prediction	HMM			AOFI			Target AOF
		Average	Std. Dev	# Obs	Average	Std. Dev	# Obs	
NTproBNP	0	1199.59	1931.84	91	1219.14	1961.77	88	AHF
NTproBNP	1	1497.57	806.87	7	1236.20	784.50	10	AHF
pCO2	0	40.30	10.74	3281	40.16	10.78	2654	ALI
pCO2	1	40.08	10.74	1188	40.36	10.68	1815	ALI
pO2	0	113.77	61.93	3281	118.74	67.61	2651	ALI
pO2	1	111.34	59.97	1188	104.93	49.94	1818	ALI
Urea Nitrogen	0	24.68	21.45	5136	24.81	22.53	4970	AKI
Urea Nitrogen	1	45.20	29.70	1781	43.12	27.75	1947	AKI
Total Bilirubin	0	4.63	8.57	1398	4.41	8.01	1353	ALF
Total Bilirubin	1	8.66	9.73	83	9.58	13.12	128	ALF

Table 3.13. Comparison between gold-standard lab tests and instance-level predictions from the HMM; (a) MIMIC-3 and (b) UW-CDR.

The transition matrix learned from the HMM showed similar patterns to those we observed in the MODS experiment (Table 3.14). First, the estimated probability of developing one or more AOFs on the next day when the patient was predicted to have no AOF was unignorable (0.7074 from the MIMIC-3; 0.3686 from the UW-CDR). Second, when all AOFs were predicted positive for the previous day, patients' recovery on the following day is limited as the transition probability to two or more AOFs is higher than the probability of the other states (0.7710 vs. 0.2290 in the MIMIC-3; 0.5871 vs. 0.4129 in the UW-CDR dataset). We also observed increased in-hospital mortality when more AOFs were predicted during a patient's hospital stay (Table 3.15), similar to the pattern observed in the MODS experiment.

MIMIC-3

From\To	0	1	2	3	4
0	0.2926	0.2047	0.3188	0.1832	0.0007
1	0.0828	0.0216	0.3304	0.5579	0.0073
2	0.1155	0.2059	0.3453	0.3300	0.0033
3	0.6979	0.0162	0.0398	0.2361	0.0100
4	0.0199	0.2091	0.5746	0.1956	0.0008

UW-CDR

From\To	0	1	2	3	4
0	0.6314	0.1757	0.1742	0.0177	0.0010
1	0.3781	0.1063	0.3855	0.1119	0.0182
2	0.3992	0.1712	0.2612	0.1195	0.0489
3	0.1927	0.1625	0.4819	0.1288	0.0341
4	0.3042	0.1087	0.4258	0.1563	0.0050

Table 3.14. Transition probability based on the number of predicted AOFs from the previous day to the next day.

(a) In-hospital mortality by number of predicted AOFs.

# of predicted AOF	MIMIC-3			UW-CDR		
	Total	**Expired**	**Proportion**	**Total**	**Expired**	**Proportion**
0	1000	26	0.0260	1045	140	0.1340
1	762	101	0.1325	53	15	0.2830
2	263	41	0.1559	176	89	0.5057
3	142	37	0.2606	61	43	0.7049
4	20	6	0.3000	10	8	0.8000

(b) In-hospital mortality by number of documented AOF discharge diagnoses.

# of diagnosed AOF	MIMIC-3			UW-CDR		
	Total	**Expired**	**Proportion**	**Total**	**Expired**	**Proportion**
0	1386	76	0.0548	1082	162	0.1497
1	523	64	0.1224	151	62	0.4106
2	224	54	0.2411	78	45	0.5769
3	52	17	0.3269	27	20	0.7407
4	2	0	0.0000	7	6	0.8571

Table 3.15. (a) Number of predicted AOFs during patients' hospital stay vs. in-hospital mortality; (b) Number of diagnosed AOFs vs. in-hospital mortality.

Through aforementioned analyses, we verified the instance-level predictions generally agreed with the clinical consensus regarding patients' AOF onset. Moreover, as the HMM achieved higher recall in the patient-level predictions with fewer instance-level positive predictions compared to the AOFI models, we conclude considering the trajectory information has the potential to refine estimated risks based on myopic evidence. If we take the labor required to verify positive instance-level prediction into account when the proposed risk prediction models are deployed, the predictions from the HMM might be more suitable as a high-risk AOF patient screening tool compared to those from the AOFI models.

3.6.3. Potential Measures to Improve Performance Using Expert Knowledge

The proposed HMM framework was able to show a higher micro-F1 score when MODS were used as evidence for patients' AOF onsets. Although the model trained with risk estimates from AOFI was not able to show higher micro-F1 scores, the results from both datasets indicate that the trajectory consideration improved the recall for the patient-level predictions while generating fewer positive instance-level predictions. Therefore, we believe the HMM was able to refine the

instance-level predictions regarding patients' risk of AOF onset by using risk estimates based on myopic evidence.

Since the HMM describes the model's behavior based on interpretable components, including a transition matrix between latent states, prior probabilities for each state, and the dependency of the latest evidence compared to the trajectory information, the prediction performance could be further improved with expert knowledge by providing additional constraints during the training.

First, α_{state} was designed to quantify the level of dependency on the trajectory information compared to the latest evidence; a higher $\alpha_{state}(\cong 1)$ indicates a higher dependency on likelihood estimates based on the trajectory information, while a lower $\alpha_{state}(\cong 0)$ indicates more reliance on the latest evidence. To incorporate expert knowledge, we could specify the condition that these estimates should follow, focusing either more or less on the trajectory information (Figure 3.3). This constraint could be implemented in the loss function as follows:

$$loss' = loss(\lambda, \psi) + CrossEntropy(\alpha_{state}, knowledge)$$

where $knowledge \in \{0,1\}^{2^k}$ consists of 0 if the states should put more weight on the latest evidence, while 1 indicates the states should focus more on the trajectory. For states where such a constraint is not available, it can be removed from the loss calculation.

State <AHF, ALI, AKI, ALF>	Dependency on trajectory
(0, 0, 0, 0)	0.4701
(0, 0, 0, 1)	0.5718
(0, 0, 1, 0)	0.7738
(0, 0, 1, 1)	0.1257
(0, 1, 0, 0)	0.4016
(0, 1, 0, 1)	0.5335
(0, 1, 1, 0)	0.3577
(0, 1, 1, 1)	0.5165
(1, 0, 0, 0)	0.3587
(1, 0, 0, 1)	0.5435
(1, 0, 1, 0)	0.6947
(1, 0, 1, 1)	0.2815
(1, 1, 0, 0)	0.6588
(1, 1, 0, 1)	0.6114
(1, 1, 1, 0)	0.5275
(1, 1, 1, 1)	0.6713

Should hepatorenal syndrome rely more on trajectory information rather than the evidence observed in the previous 24 hours?
- Add cross-entropy loss to the loss function to penalize this value when it is lower than 0.5

Figure 3.3. Examples of α_{state} adjustment.

Second, the prior probabilities of each state were also estimated directly from the dataset. If experts could identify groups of states that are clinically invalid, such prior probabilities could be clamped to 0 by modifying the prior probabilities as follows:

$$p(state)' = p(state) \odot mask_{prior}$$

where $\odot$ indicates element-wise production, and $mask_{prior} \in R^{2^k}$ consists of 0 when the state is clinically invalid, and 1 otherwise (Figure 3.4a).

Lastly, the transition probabilities between states could also be masked similarly to the estimated prior probability mentioned above (Figure 3.4b). For example, when experts could identify per-day transitions that are clinically invalid in ICUs, the transition matrix $TM = (softmax(raw_{TM}))$ could also be changed as follows:

$$TM' = softmax(raw_{TM} \odot mask_{transition})$$

where $mask_{transition} \in R^{2^k \times 2^k}$ consists of 0 when the state transition is clinically invalid, and 1 otherwise.

(a) Example of masking the prior probability.

State	Prior Probability
(0, 0, 0, 0)	0.8610
(0, 0, 0, 1)	0.0024
(0, 0, 1, 0)	0.0517
(0, 0, 1, 1)	0.0108
(0, 1, 0, 0)	0.0293
(0, 1, 0, 1)	0.0062
(0, 1, 1, 0)	0.0006
(0, 1, 1, 1)	0.0016
(1, 0, 0, 0)	0.0118
(1, 0, 0, 1)	0.0075
(1, 0, 1, 0)	0.0030
(1, 0, 1, 1)	0.0005
(1, 1, 0, 0)	0.0097
(1, 1, 0, 1)	0.0000
(1, 1, 1, 0)	0.0039
(1, 1, 1, 1)	0.0001

Is it impossible to have only three active AOFs instead of four AOFs simultaneously?
- Clamp the prior probability to 0

(b) Example of masking the transition probability.

If it is impossible for patients without any AOF onset to develop only ALF, or for patients with hepatorenal syndrome to develop AHF without ALI on the next day, clamp the state transition probability to 0

From\To	(0, 0, 0, 0)	(0, 0, 0, 1)	(0, 0, 1, 0)	(0, 0, 1, 1)	(0, 1, 0, 0)	(0, 1, 0, 1)	(0, 1, 1, 0)	(0, 1, 1, 1)	(1, 0, 0, 0)	(1, 0, 0, 1)	(1, 0, 1, 0)	(1, 0, 1, 1)	(1, 1, 0, 0)	(1, 1, 0, 1)	(1, 1, 1, 0)	(1, 1, 1, 1)
(0, 0, 0, 0)	0.2926	0.1058	0.0098	0.0411	0.0266	0.0848	0.1621	0.0753	0.0625	0.0045	0.0032	0.0066	0.0232	0.1011	0.0002	0.0007
(0, 0, 0, 1)	0.0195	0.0520	0.0003	0.2332	0.0012	0.0106	0.0421	0.3395	0.0014	0.0475	0.0220	0.1214	0.0217	0.0773	0.0034	0.0068
(0, 0, 1, 0)	0.0619	0.0133	0.0037	0.0023	0.0007	0.0004	0.0081	0.0441	0.0004	0.0012	0.0246	0.0033	0.0011	0.8256	0.0014	0.0080
(0, 0, 1, 1)	0.0096	0.0028	0.0001	0.0201	0.0002	0.0000	0.0009	0.9576	0.0001	0.0019	0.0005	0.0003	0.0000	0.0058	0.0001	0.0001
(0, 1, 0, 0)	0.0504	0.0002	0.0000	0.0001	0.0051	0.0001	0.9220	0.0044	0.0017	0.0040	0.0002	0.0001	0.0014	0.0058	0.0041	0.0005
(0, 1, 0, 1)	0.0321	0.0016	0.0007	0.8679	0.0001	0.0031	0.0037	0.0425	0.0028	0.0093	0.0029	0.0181	0.0016	0.0063	0.0037	0.0037
(0, 1, 1, 0)	0.6277	0.0001	0.0001	0.0001	0.0003	0.0000	0.2251	0.1263	0.0001	0.0000	0.0006	0.0000	0.0013	0.0173	0.0010	0.0000
(0, 1, 1, 1)	0.9976	0.0003	0.0001	0.0003	0.0000	0.0002	0.0003	0.0008	0.0000	0.0000	0.0000	0.0001	0.0001	0.0000	0.0000	0.0000
(1, 0, 0, 0)	0.2678	0.0542	0.0000	0.0017	0.0003	0.0001	0.0056	0.0107	0.0125	0.0229	0.0867	0.0490	0.0051	0.4127	0.0496	0.0210
(1, 0, 0, 1)	0.0997	0.0143	0.0294	0.1516	0.0639	0.0374	0.0395	0.0081	0.5166	0.0035	0.0023	0.0053	0.0132	0.0095	0.0049	0.0008
(1, 0, 1, 0)	0.8197	0.0035	0.0000	0.0001	0.0004	0.0001	0.0011	0.0030	0.0032	0.0032	0.1109	0.0116	0.0015	0.0084	0.0275	0.0060
(1, 0, 1, 1)	0.1633	0.1188	0.0011	0.0481	0.0003	0.0057	0.0055	0.0616	0.0034	0.1344	0.0040	0.3703	0.0054	0.0645	0.0041	0.0096
(1, 1, 0, 0)	0.0484	0.0207	0.0003	0.0010	0.0042	0.0012	0.0540	0.0126	0.2870	0.0254	0.4295	0.0250	0.0008	0.0630	0.0189	0.0080
(1, 1, 0, 1)	0.0425	0.0657	0.0010	0.0123	0.0010	0.0002	0.0185	0.7669	0.0011	0.0233	0.0043	0.0096	0.0004	0.0093	0.0031	0.0409
(1, 1, 1, 0)	0.6466	0.0058	0.0000	0.0001	0.0014	0.0002	0.0017	0.0232	0.0017	0.0030	0.0245	0.0217	0.0056	0.1247	0.1261	0.0139
(1, 1, 1, 1)	0.0199	0.1877	0.0059	0.0080	0.0019	0.0014	0.5472	0.0175	0.0136	0.0040	0.0138	0.0030	0.0003	0.1736	0.0015	0.0008

Figure 3.4. Examples of adding constraints on probabilistic estimates from the HMM.

3.7. Conclusion

We have showed that the performance of EWS-based screening could be improved by integrating EWS trajectories on relevant clinical events into a unifying probabilistic framework. Moreover, the probabilistic formulation of AOF prognoses provided interpretable components, which allowed us to conduct a clinical evaluation on the estimates. In the experiment conducted with risk estimates from AOFI models, we also found a potential of the proposed approach for refining AOF risk predictions based on estimates derived from myopic evidence. The evaluation showed that the parameters estimated by the HMM generally agreed with the medical consensus. Lastly, the proposed framework showed the potential of HMM training with temporal risk estimates of acute-onset diseases and terminal outcomes when gold standard information about the exact time of event onsets is unavailable. When expert-driven EWS are deployed for high-risk patient screenings, the proposed HMM framework will enable physicians to review the trend of risk state transitions and calibrate the parameters to improve the screening performance.

Chapter 4. Counterfactual Analysis of Organ Toxicity of Clinical Interventions

4.1. Introduction

Hospitalized patients receive a wide array of clinical interventions during their hospital stay. Although the aim of such clinical intervention is to treat physiological abnormalities in patients, the prevalence of adverse events due to planned clinical interventions is not insignificant. When administered medications are focused on, the Smith et al. study suggested that the incidence of adverse drug reaction (ADR) could be as high as 14.7% of all patient cases from United Kingdom national hospital system wards, and half of such events were either definitely or possibly avoidable [93].

Although there are many measures evaluating the potential toxicity of medications, including preapproval clinical trials from manufactures or randomized clinical trials after the product is released [94], they are cost-intensive, which makes them more difficult to apply in all suspected cases. Moreover, although there are some reactive measures to warn about suspected high-risk medications—through adverse drug experience reports to manufactures [95], post-marketing observational studies [96], or market withdrawal orders from governing agencies (e.g., Food and Drug Administration)—most of them are conducted after a sufficient number of formally reported ADR incidents; the earlier study insisted that the gap between the first ADR event to the market withdrawal could be as long as six years [97].

In practice settings, foreseeing ADRs due to clinical intervention is challenging because the risk varies by patient and considering relevant risk factors is solely dependent on the expertise of each caregiver; this may result in variations in the quality of care. Although many dosing guidelines regarding controversial medication administration have been implemented by expert groups to decrease such variances, they are mostly focused on medications that are widely used in practice. For medications that are either recently released or less utilized, however, such information is not readily available. Therefore, if a clinical decision support system can provide information about the potential risks of candidate clinical interventions for each patient, which might be too complex to be analyzed in time-critical settings, such as ICUs, it would allow caregivers to either avoid clinical interventions with high potential risk of ADR when alternatives are available, or prepare for the onset of ADR events if there is no alternative for patient management.

There are two possible directions for implementing clinical decision support on clinical intervention choices: 1) providing a list of applicable interventions based on the patient's status, or 2) providing the quantified risk of ADRs for queried interventions. The former was the focus of expert systems in the mid-1970s, and most of them were implemented as a rule-based system. Shortliffle et al. aimed to provide a list of applicable antibiotics based on a patient's symptoms entered by caregivers [98]. However, such a system was not successfully integrated into the clinical workflow because the benefits of having the list of system-generated antibiotics suggestion were not clear. As the system encoded the general rules of antibiotic selection criteria, it tended to provide either choices that were too obvious for physicians or not applicable in cases that were not considered during the model implementation. Moreover, as the legal boundary has not been clearly drawn regarding the responsibility of such model-aided clinical practice [99]–

[101], model-generated suggestions on applicable interventions might have some limitations as a clinical decision support even in the current practice setting.

In contrast, a system that could provide a quantified risk of potential ADR on queried interventions would be more applicable clinical decision support tools, especially where intervention selections tend to rely more on the empirical aspects of individual experts. Such a system could be also helpful when potential side effects of interventions are either too complex to be evaluated for individual patients in time-critical care settings or the current understanding of the clinical intervention and the patient's physiological disturbance is still limited. For instance, antibiotic administration is a core part of treating patients with infectious diseases, and the treatment choices are often empirically made due to the success rate of identifying responsible pathogens and the delay between the presentation of symptoms and the identification process [102].

When treating patients with suspected infections, wide-spectrum antibiotics are often administered as the first response before transitioning to narrow-spectrum antibiotics once causative pathogens are identified. Moreover, it is widely accepted that some antibiotics (e.g., aminoglycosides or vancomycin) are responsible for drug-induced nephrotoxicity. At the same time, balancing the trade-off between the efficacy of infection treatment and the risk of potential nephrotoxicity relies heavily on a physician's specialty and experience [103]. Therefore, many clinical guidelines have been implemented to provide antibiotic dosing information on patients with renal impairment to reduce inter-practitioner variance with regards to antibiotic-induced nephrotoxicity. In addition, the nephrotoxicity of antibiotics is an active research area, which could indicate that the current understanding remains limited [104], [105]. If data-driven clinical decision support can provide quantified risks of potential nephrotoxicity on candidate antibiotics at the time of decision-making in time-critical settings, it would serve as an additional safety measure preventing antibiotic-induced renal impairment.

Using observed serum creatinine level (SCr) measurements from EHRs, we propose a counterfactual modeling framework that can explain the observed kidney function trajectory based on different antibiotic choices by jointly learning about 1) the baseline kidney trajectory, and 2) the renal impairment due to administered antibiotics. To achieve this, we trained the effect-free model and the response model jointly; the effect-free model focuses on illustrating a patient's baseline kidney trajectory without the influence of planned antibiotic administration, while the response model focuses on describing the nephrotoxicity of the planned administration of antibiotics through potential SCr increments. With the estimated nephrotoxicity of each antibiotic administration from the trained model, we conducted quantitative analyses to evaluate the clinical validity of such estimates using literature review.

4.2. Related Works

For administered medications, their efficacies are often quantified through drug half-life, the time required to decrease the concentration of the medication half in the body. The drug half-life mostly depends on total body clearance [106], and this is defined as the sum of the renal and non-renal clearance. When a patient's renal clearance is impaired and the administered drugs are known to be excreted through the kidney, the resulting drug accumulation often causes various renal complications [107]. Therefore, most clinical institutions have internal renal dosing guidelines to prevent drug-induced renal impairment. Among the different nephrotoxins

currently used in the clinical workflows, the nephrotoxicity of antibiotics is widely known and there exist many ongoing studies evaluating the magnitude of renal impairment caused by potentially nephrotoxic antibiotics [104], [105].

During patient management, patients with suspected infection are often first treated with wide-spectrum antibiotics, then are moved to narrow-spectrum antibiotics once the responsible pathogens are identified through microbiology cultures, which often takes 24 to 48 hours to confirm the pathogens [108]. Therefore, there exists a time gap between a patient's initial symptom presentation and the confirmation of the pathogens, and physicians often need to initiate antibiotic therapy based on institutional antibiotic dosing guidelines and experience alone until the microbiology culture successfully identifies the pathogens. Patients' symptoms serve as evidence for physicians to estimate the potential spectrum of infection they need to cover, and there often exists a number of candidate antibiotic choices for the suspected spectrum. Even after the infected pathogens and a list of susceptible antibiotics were identified from the microbiology study, there is no tool to quantify the potential nephrotoxicity of the candidate antibiotics based on a patient's status quo. Therefore, it would be challenging for physicians to objectively compare the trade-offs between controlling the infection and the nephrotoxic effects due to administered antibiotics. In the proposed study, we aimed to quantify the nephrotoxic effect of different antibiotic administration plans by estimating the resultant SCr increment so that it could serve as clinical decision support, thereby allowing physicians to quantitatively compare such trade-offs.

Counterfactual analysis stems from causal inference, which aims to extract a potential causal relationship between the outcome and the variables presented in the dataset [109]. Recently, Schulam et al. [110] presented promising prediction accuracy of future SCr trajectory prediction based on SCr history and the timing of renal support therapies. They described the SCr trajectory using two separate submodels: one describing baseline kidney trajectory, and another predicting the effect of renal support therapy on the baseline kidney trajectory. Similarly, our work hypothesized that the nephrotoxic effect of antibiotics and baseline kidney function could also be separated from the observed SCr measurements by training two submodels that each explain the baseline kidney functions without the influence of the planned antibiotics administration (the effect-free trajectory model) and the nephrotoxic effect caused by the antibiotics (the response model). Compared to their modeling framework, we assumed the physiological disturbance of each antibiotic administration would vary by the patient's baseline physiology, so we allowed predictions from the effect-free trajectory model to contribute to the nephrotoxicity estimation on the planned administration of antibiotics from the response model. Moreover, there exists a domain consensus regarding antibiotics administrations: higher doses of nephrotoxic antibiotics tend to worsen patient's kidney function. Therefore, we considered such constraints during the model implementation by not letting the response model predict lower nephrotoxicity on higher-dose antibiotics administration. We assumed that this constraint would facilitate clinical interpretation of the model behavior by avoiding counterintuitive results, such as possessing clinically known risk factors (e.g., asthma) being identified as a benevolent feature on the risk of the target outcome (e.g., pneumonia) [111].

4.3. Methods

This study focused on implementing the model predicting the future SCr trajectory of patients admitted to ICUs based on previously measured SCr levels and antibiotics that were

administered during their ICU stays. To do so, we trained a model to explain the factual SCr levels, SCr measurements documented in the EHR, by estimating the SCr level without the influence of planned antibiotic administration (the effect-free trajectory, #1 in Figure 4.1) and the SCr increment due to the planned antibiotic administrations (nephrotoxic response to the planned antibiotic administration, #2 in Figure 4.1). During the model training, we trained two submodels jointly, each estimating the effect-free trajectory and the nephrotoxic response to the planned antibiotic administration. In the following sections, we refer to the antibiotic administrations observed in the dataset as factual antibiotic administrations and to all other potential administration patterns that were not observed in the dataset as counterfactual antibiotic administrations on each patient.

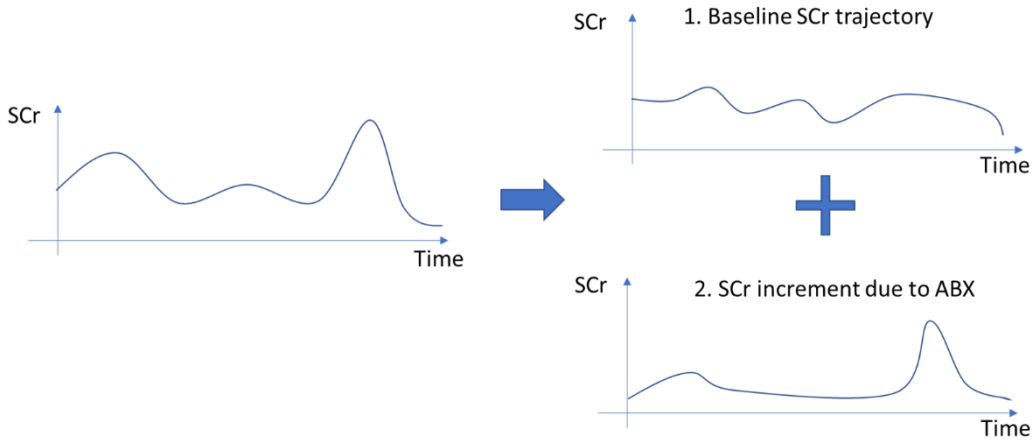

Figure 4.1. Decomposing observed for SCr levels with the effect-free trajectory (#1) and the nephrotoxic response to antibiotics (#2).

By separating the influence of planned antibiotic administration from the observed SCr trajectory, we aimed to quantitatively estimate the nephrotoxic effect of different antibiotics from the factual administration. Therefore, two trained submodels were assumed to facilitate the counterfactual analysis, such as comparing the nephrotoxic response with and without the target antibiotic. The details are discussed in the following sections.

4.3.1. Input

Serum creatinine (SCr) measurements documented in the EHR were used as a *physiological input*, while the time stamp and dosing information of administered antibiotics documented in the EHR were used as an *intervention input* for the study. Both physiological inputs and intervention inputs were summarized as a daily level during preprocessing. For the day d_i, x_{SCr,d_i} represented the distribution of observed SCr, and x_{abx,d_i} represented the administered antibiotics during the day. The daily SCr distribution $x_{SCr,d_i} \in R^4$ was described with min, max, average, and standard deviation based on the SCr measured during the day. For the daily antibiotic administration x_{abx,d_i}, we described the vector with the total amount administered for each antibiotic, similar to renal antibiotic dosing guidelines (e.g., two administrations of 50 mg and 100 mg vancomycin during the day d_i were represented as 150 mg cumulative vancomycin administration in x_{abx,d_i}), where $x_{abx,d_i} \in R^{|ABX|}$ and $|ABX|$ was the number of different kinds of antibiotics considered for the study.

4.3.2. Modeling

For the study, we aimed to decompose the patient's SCr trajectory into the baseline SCr trajectory without the influence of antibiotics and the resultant SCr increment due to the antibiotics administered. By separating the influence of antibiotics from the observed SCr trajectory, we hypothesized that 1) the nephrotoxic effect of different antibiotic administrations could be learned, and 2) the baseline SCr level when no antibiotics are planned to be administered could be estimated.

To describe the baseline SCr trajectory without the influence of planned antibiotic administration and the nephrotoxicity due to the planned administration of antibiotics separately, we trained two submodels jointly: 1) the effect-free trajectory model $f_{EFT,\theta}$, and 2) the response model $g_{RES,\phi}$. The $f_{EFT,\theta}$ aimed to predict the patient's next-day SCr distribution with the assumption that no antibiotics would be administered on the target date. The submodel was implemented using a recurrent neural network (RNN) with a gated recurrent unit (GRU) in addition to a residual connection to the previous-day SCr distribution, $x_{SCr,d_{i-1}}$. Using the predicted latent vector h_{d_i} from the GRU cell for the day d_i, $h_{d_i} = GRU(< x_{SCr,d_{i-1}}, h_{d_{i-1}} >)$, where $h_{d_{i-1}}$ is the cell state forwarded from the previous date, we estimated the effect-free SCr distribution for the next day $N(\widehat{avg_{d_i}}, \widehat{std_{d_i}})$ using the feed-forward layer as follows:

$$< \widehat{avg_{d_i}}, \widehat{std_{d_i}} >= f_{EFT,\theta}(< x_{SCr,d_{i-1}}, h_{d_i} >)$$

The response model $g_{RES,\phi}$ aimed to estimate the nephrotoxicity due to the planned administration of antibiotics x_{abx,d_i} by predicting the resultant SCr increment. Differently from $f_{EFT,\theta}$, we implemented $g_{RES,\phi}$ not to use previous antibiotics administration information, thereby making the model rely solely on the predicted latent state h_{d_i} and the planned administration of antibiotics x_{abx,d_i} with a residual connection to $x_{SCr,d_{i-1}}$ to predict the average increment of SCr on the next day, $\widehat{modif}_{d_i}$:

$$\widehat{modif}_{d_i} = \begin{cases} \max\left(0, g_{RES,\phi}(< x_{abx,d_i}, h_{d_i}, x_{SCr,d_{i-1}} >)\right) \ when \ \left|\left|x_{abx,d_i}\right|\right|_2 \neq 0 \\ 0 \ when \ \left|\left|x_{abx,d_i}\right|\right|_2 = 0 \end{cases}$$

Although it is possible that previously administered antibiotics had a long-term effect on a patient's renal function, we assumed this information should be conveyed through the latent variable describing the effect-free creatinine trajectory, h_{d_i}. For example, aminoglycoside is a class of antibiotics—including tobramycin, gentamicin, and amikacin—with known nephrotoxicity [112], and they are known to be effective when treating gram-negative infections. At the same time, it is also known that some gram-negative pathogens, such as *Staphylococcus*, also cause renal impairment when the kidneys are infected. Therefore, if previous antibiotic administration information was directly provided for the nephrotoxicity estimation, we assumed that the previous administration of antibiotics frequently used to control pathogens potentially damaging the kidneys would confound the true nephrotoxicity estimation of the planned antibiotics x_{abx,d_i}.

During the literature review, we found that most of the clinical guidelines for presumed nephrotoxic antibiotics suggest lower doses for patients with renal impairment. Moreover, the

risk of presenting nephrotoxicity tends to increase when antibiotics from different classes were concurrently administered because drug-induced side effects might cause direct or indirect kidney damage [113]. Therefore, we assumed that the nephrotoxic response would increase or stay the same when either the total dose of a single antibiotic was increased, or different types of antibiotics were additionally administered. Therefore, we implemented $g_{RES,\phi}$ as a submodular function to satisfy the following property:

$$g_{RES,\phi}(< x_{abx,d_i}, h_{d_i}, x_{SCr,d_{i-1}} >) \leq g_{RES,\phi}(< x_{abx,d_i}', h_{d_i}, x_{SCr,d_{i-1}} >),$$

where $\left|\left|x_{abx,d_i}\right|\right|_2 \leq \left|\left|x'_{abx,d_i}\right|\right|_2$ due to a higher dose of antibiotics that had been administered or additional administration of different antibiotics. To implement this, we first transformed the concatenated vector of the hidden state h_{d_i} and $x_{SCr,d_{i-1}}$ using a feed-forward layer with Rectified Linear Units (ReLU) activation in order to generate the positive-value embedding vector describing the patient's kidney function trajectory, $emb^+_{d_i}$. Then, $emb^+_{d_i}$ was concatenated with x_{abx,d_i}, and fed into another feed-forward layer with non-negative weights to allow interactions among features and to conserve the monotonicity. The final SCr increment due to the planned antibiotic administration x_{abx,d_i} was estimated using the additional feed-forward layer with non-negative weights after the transformation with the scaling function. For the scaling function, we compared the performance with logarithmic function $y = \log(x + 1)$, exponential function $y = \exp(x) - 1$, and linear function $y = x$. The architecture of $g_{RES,\phi}$ is presented in Figure 4.2.

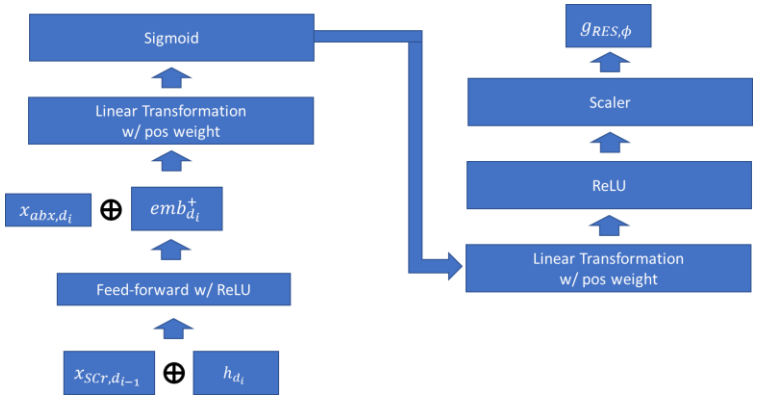

Figure 4.2. The response model architecture; ⊕ indicates vector concatenation.

In clinical trial setting, most of studies defined the target antibiotic as nephrotoxic when the drug was responsible for 0.5 mg/dL increment of the patient's SCr increment. Accordingly, by using the predicted $\widehat{avg_{d_i}}$ and $\widehat{std_{d_i}}$ from $f_{EFT,\theta}$ and $\widehat{modif_{d_i}}$ from $g_{RES,\phi}$, we predicted the next-day SCr distribution under the planned administration of antibiotics as $N\left(\widehat{avg_{d_i}} + \widehat{modif_{d_i}}, \widehat{std_{d_i}}\right)$.

As the presented study used the dataset extracted from the EHR, we did not have gold-standard information to evaluate the predicted $\widehat{avg_{d_i}}$, $\widehat{std_{d_i}}$, and $\widehat{modif_{d_i}}$ separately. Instead, we used the Kullback–Leibler (KL) divergence for evaluating the predicted distribution $\widehat{x_{SCr,d_i}} \sim N\left(\widehat{avg_{d_i}} + \widehat{modif_{d_i}}, \widehat{std_{d_i}}\right)$ given the observed next-day SCr level distribution $x_{SCr,d_i} \sim N\left(avg_{d_i}, std_{d_i}\right)$ to calculate the training loss, $loss(\theta, \phi)$, for the optimization:

$$loss(\theta, \phi) = \sum_{pts} \sum_{i} KL\left(\widehat{x_{SCr,d_i}} || x_{SCr,d_i}\right)$$

$$= \sum_{pts} \sum_{i} \left[\log \frac{\widehat{std_{d_i}}}{std_{d_i}} + \frac{std_{d_i}^2 + \left(\widehat{avg_{d_i}} + \widehat{modif_{d_i}} - avg_{d_i}\right)^2}{2\widehat{std_{d_i}}} - \frac{1}{2} \right]$$

For the regularization, we used dropout on feed-forward layers. The overall architecture of the model is presented in Figure 4.3.

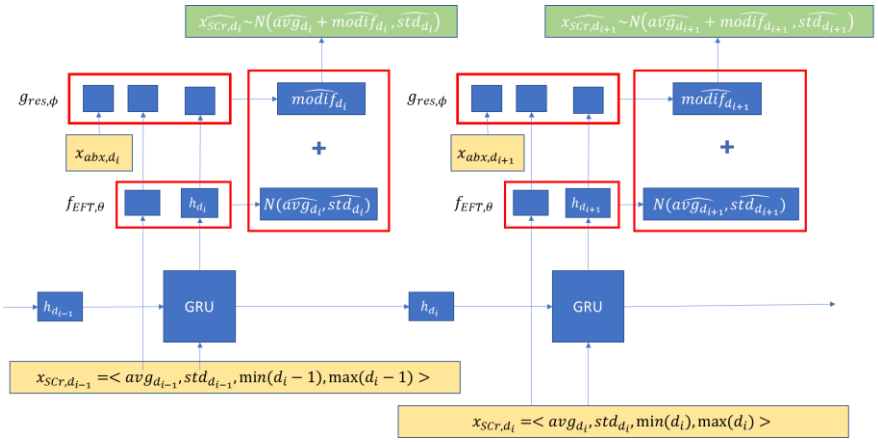

Figure 4.3. The overall model architecture.

4.3.3. Evaluation

From the models trained with different hyperparameters, we selected the model that achieved the smallest KL divergence on the validation set from the factual antibiotic administration. During the analysis, however, we used the mean absolute error of the predicted mean and the observed mean SCr as an accuracy measure to describe the selected model because it is a more straightforward performance metric for showing the accuracy of predicted SCr in the clinical setting. In order to compare the effect of having an additional model explaining the response of antibiotic administrations, we trained a baseline model that predicts the next-day SCr distribution only with $f_{EFT,\theta}$ without the factual antibiotic administrations x_{abx,d_i} using the following loss function:

$$loss_{baseline}(\theta, \phi) = \sum_{pts} \sum_{i} \left[\log \frac{\widehat{std_{d_i}}}{std_{d_i}} + \frac{std_{d_i}^2 + \left(\widehat{avg_{d_i}} - avg_{d_i}\right)^2}{2\widehat{std_{d_i}}} - \frac{1}{2} \right]$$

For patients with infectious diseases in ICU settings, it is common to observe more than two antibiotics administered on the same date to cover the suspected spectrum of infecting pathogens [102]. Therefore, the predicted nephrotoxicity of the factual antibiotic administration from $g_{RES,\phi}$ could not be directly used to evaluate the nephrotoxicity of an individual antibiotic because it entailed a patient's baseline kidney function and the influence of different antibiotics administered at the same time. In order to measure the contribution of each antibiotic on predicted nephrotoxicity from the response model $g_{RES,\phi}$, we borrowed the feature occlusion

method used earlier in Zeiler et al's work [114]. In this study, they trained a computer-vision model to visualize what part of the input image contributes when predicting whether the target object is present or not by providing attention weights from the input image. During their analysis, they conducted a feature occlusion to evaluate how the prediction accuracy changes when the region of high attention was masked to see whether the model was making the prediction based on the object or the background information presented in the image. This evaluation scheme was also used in a clinical informatics study when evaluating the feature contributions of the trained RNN model [24]. Therefore, we also adopted the feature occlusion method to quantify the contribution of each antibiotic to the predicted nephrotoxicity. For all patient-days with the target antibiotic *med* administration, we compared the predicted risk increment based on the factual antibiotics administration, $\widetilde{modif}_{d_i}$, with the counterfactual antibiotics administration without the target antibiotic, $\widetilde{modif}_{d_i}$: $effect_{med} = (\widetilde{modif}_{d_i}, \widetilde{modif}_{d_i,med})$, where $\widetilde{modif}_{d_i,med} = \max\left(0, g_{RES,\phi}\left(< \widetilde{x_{abx,d_i,med}}, h_{d_i}, x_{SCr,d_{i-1}} >\right)\right)$, $\widetilde{x_{abx,d_i,med}} = x_{abx,d_i} \backslash \{med\}$, and $med \in ABX$. For all measured $effect_{med}$ pairs on each factual antibiotic administration, we conducted a paired *t*-test to evaluate whether the estimated nephrotoxicity distribution of the factual administration $\widetilde{modif}_{d_i}$ was statistically significantly higher than the counterfactual administration $\widetilde{modif}_{d_i,med}$.

4.4. Results

For the study, we extracted hospital admissions from the MIMIC-3 dataset [65] to train and evaluate the model, which consist of physiological measurements and information about clinical intervention administrations performed on patients who stayed in ICUs at Beth-Israel Hospital in Boston, MA. By including adult patients who had SCr measured at least once during their ICU stay, a total of 16,332 hospital admissions were used for the study; we further divided hospital admissions into a 7:2:1 ratio for training, validation, and testing set, respectively. For antibiotics, we considered 47 different antibiotics that were observed from the training set as an intervention input. The demographics of the patients are provided in Table 4.1.

Hospital admissions	16,332
Patient-day	144,417
# of Creatinine level measurement	95,762
Creatinine level distribution	1.5470±1.5231 mg/dL
Admission Type	ELECTIVE – 2605 EMERGENCY – 14,227 URGENT – 224
Gender	Female – 7,563 Male – 9,493
Age	64.07±16.89

Table 4.1. Demographics of the MIMIC-3 dataset used for the study.

The trained model aimed to predict SCr distribution by adding two different trajectory estimates: the effect-free trajectory and the response to the planned antibiotics on day d_i. The effect-free trajectory was estimated using observed SCr until $\{x_{SCr,d_0}, \dots, x_{SCr,d_{i-1}}\}$, where d_0 indicates the date of ICU admission, while the renal response to administered antibiotics was estimated based on the antibiotics administration plan on day d_i, x_{abx,d_i}, along with the latent variables describing the effect-free trajectory. When comparing the predicted mean SCr, $\widehat{avg}_{d_i} + \widehat{modif}_{d_i}$, and the observed mean SCr, avg_{d_i}, from the test set, the model showed a mean absolute error of 0.2092 in a total of 10,751 patient-days. The mean absolute error from the baseline model, only considering SCr trajectory without factual antibiotic administrations, was 0.2135 in the same dataset. As the earlier study reported the error range of the SCr measurement from the patient sample was around 0.2 [115], we believe the process of decomposing the influence of antibiotics with the separate response model did not degrade the performance of the SCr trajectory prediction tasks.

The response model $g_{RES,\phi}$ aimed to predict the SCr increment due to the planned antibiotic administrations x_{abx,d_i}, and the contribution of each antibiotic to the estimated increment was evaluated using feature occlusion. In the test set, we observed at least one antibiotic administration on 4,173 patient-days (38.81%), and more than two different kinds of antibiotics were co-administered on 2,429 patient-days. To analyze the influence of individual antibiotic on the predicted SCr increments, we made two nephrotoxicity estimates, SCr increments with and without the target antibiotic, and conducted a paired *t*-test to evaluate whether the target antibiotic was responsible for increased SCr with statistical significance.

Using the dosing information in the factual antibiotic administration, the feature occlusion analysis showed an average 0.0427 mg/dL SCr increment with the standard deviation of 0.1420 from 8,134 individual antibiotic administration events in the test set. From the paired *t*-test, we identified 43 out of 47 antibiotics that showed a statistically significant increase in SCr under the significance level $\alpha = 0.05$ and the power $1 - \beta = 0.8$. However, as both intervention inputs and physiological inputs were extracted from patients under the ongoing clinical management, no antibiotic showed their 95% confidence interval upper bound (95% CI UB) higher than the clinical nephrotoxicity threshold used in other studies, 0.5 mg/dL SCr increment, due to the administered antibiotic [116], [117].

Table 4.2 shows the top-10 frequently administered antibiotics from the test set and their average SCr increments evaluated through feature occlusions. Although the majority of the antibiotics are known nephrotoxins, which require the dosing adjustments for patients with renal impairments, we believe the institutional dosing guideline might prevent us to observe nephrotoxicity on the drugs in the lists. Therefore, the response model was not able to learn the corresponding SCr increments for antibiotics administrations documented in terms of standard dosing units.

Average SCr increment	# of administrations	Documented Units	Antibiotics	*p*-value	Std. Dev.
0.0071	2074	dose	Vancomycin	<0.001	0.0471
0.0457	755	dose	Piperacillin/ Tazobactam (Zosyn)	<0.001	0.1894
0.0301	710	dose	Cefepime	<0.001	0.1217
0.0386	656	dose	Metronidazole*	<0.001	0.0469
0.0536	608	dose	Meropenem	<0.001	0.1007
0.0239	465	dose	Ciprofloxacin	<0.001	0.0986
0.0280	389	dose	Cefazolin	<0.001	0.1399
0.0460	332	dose	Ceftriaxone*	<0.001	0.1531
0.0145	221	dose	Acyclovir	0.067	0.1177
0.0815	220	dose	Piperacillin	<0.001	0.2330

Table 4.2. Average SCr increments predicted from the response model and p-values from paired t-test on factual and counterfactual ABX administration; * indicates that the antibiotic did not require dosing adjustment according to the drug labels.

In addition to vancomycin, aminoglycosides are also known to be nephrotoxic, however, the response model was not able to learn about their nephrotoxicity from the datasets, except for Gentamicin (Table 4.3). From the training dataset, Gentamicin showed various dosing information in different underlying kidney functions compared to Tobramycin and Amikacin (Figure 4.4). Moreover, as the nephrotoxicity of aminoglycosides is already known in practice, the SCr levels on the previous day of antibiotics administration are comparably lower than for the administration of other antibiotics (average previous-day SCr level accompanied by non-aminoglycosides administration was 1.4121 ± 1.4272 and 1.1683 ± 1.8986 on aminoglycoside administration), which indicates it is not common to observe cases where nephrotoxic antibiotics are administered to patients with renal dysfunction. Therefore, we suspect that the availability of the aforementioned cases might limit the model to learn the nephrotoxic effects of the other two aminoglycosides through the counterfactual analysis.

AVG SCr increment	CNT	Documented Units	LABEL	p-val	STD	95% UB
0.2762	50	dose	Gentamicin	<0.001	0.2476	0.3349
0.0046	72	dose	Tobramycin	<0.001	0.0052	0.0056
0.0012	25	dose	Amikacin	<0.001	0.0010	0.0015

Table 4.3. Predicted SCr increments from the response model on aminoglycosides.

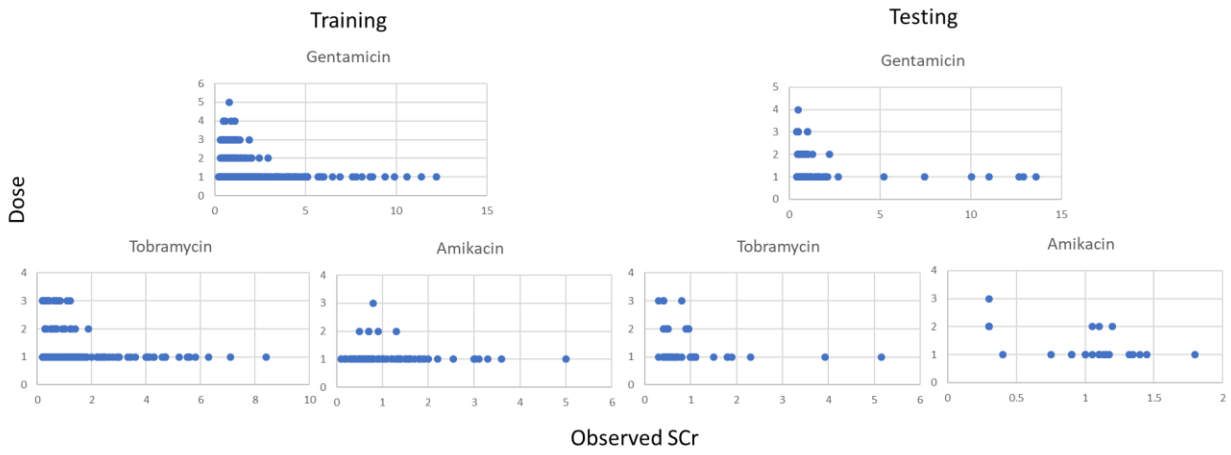

Figure 4.4. Dosing variability presented in the training and test set.

The top five antibiotics with statistical significance in paired *t*-tests were presented in Table 4.2 in descending order of the 95% CI UB. From renal dosing guidelines from three different institutions, including Stanford Medicine[3], Nebraska Medical Center[4], and University of Washington/Harborview Medical Center[5], we found lower dosing suggestions for Penicillin G Potassium, Ceftazidime, and Ambisome. For Ceftrazidime, although the administrations based on standard dosing were not identified as nephrotoxic (average SCr increments were 0.0468 ± 0.0550 on 67 Ceftrazidime administrations documented with standard dosing), the antibiotic administered with gram-based dosings were identified as nephrotoxic. From the test sets, instances with the gram-based dosing documentation was accompanied by standard dosing-based documentation for the same day. Therefore, we suspect that the gram-based dosing documentation might be used to document additional Ceftrazidime administration along with the standard dosing; among instances with gram-based Ceftrazidime dosing information, 26.31% (5/19) of instances also had the standard dose-based administration in the training set, and 40% (2/5) of instances had the information in the test set. Therefore, although the standard-dosing based administration was not responsible for the patient's nephrotoxicity, we believe the dosing practice that exceeds the institutional standard might accounts for the patient's presented nephrotoxicity with Ceftrazidime.

[3] http://med.stanford.edu/bugsanddrugs/dosing-protocols/_jcr_content/main/panel_builder/panel_0/download/file.res/SHC%20ABX%20Dosing%20Guide%202019-02-01.pdf

[4] https://www.nebraskamed.com/sites/default/files/documents/for-providers/asp/antimicrobial-renal-dosing-guidelines.pdf

[5] https://depts.washington.edu/idhmc/wp-content/uploads/2015/11/Antibiotic-dosing.pdf

AVG increments	CNT	Documented Units	LABEL	p-val	STD	95% CI UB
0.3091	12	dose	Penicillin G potassium	<0.001	0.1218	0.3722
0.3431	112	dose	Linezolid*	<0.001	0.1671	0.3693
0.2091	4	grams	Ceftazidime	0.040	0.1197	0.3499
0.2762	50	dose	Gentamicin	<0.001	0.2476	0.3349
0.2068	53	dose	Ambisome	<0.001	0.1914	0.2508

Table 4.4. Top five statistically significant antibiotics based on 95% CI UB; * no dosing adjustment specified from drug labels.

During the analysis, there were 122 individual antibiotic administrations where the response model predicted SCr increments higher than 0.5mg/dL (Table 4.5). The average SCr levels the day before antibiotic administration were 5.9613 mg/dL with a standard deviation of 2.5424, which was higher than cases with predicted SCr increments less than 0.5mg/dL (the average of 1.3382 mg/dL with standard deviation 1.2961). As the majority of the estimated nephrotoxins in Table 4.2 only includes a few of the following, this might indicate that patients' baseline SCr levels have implications on the presented nephrotoxicity, which agrees with the consensus that dose-adjustment should be based on a patient's creatinine clearance.

Antibiotic	# of occurrence	Antibiotic	# of occurrence
Piperacillin/Tazobactam (Zosyn)	27	Ciprofloxacin	4
Cefepime	14	Ampicillin/Sulbactam (Unasyn)	3
Ceftriaxone	12	Gentamicin	3
Piperacillin	9	Acyclovir	1
Ampicillin	8	Imipenem/Cilastatin	1
Linezolid*	8	Gancyclovir	1
Cefazolin	8	Penicillin G potassium	1
Nafcillin	7	Ambisome	1
Meropenem	7	Metronidazole*	1
Vancomycin	6		

Table 4.5. Antibiotics responsible for SCr increment higher than 0.5mg/dL; * no dosing adjustment suggested from drug labels.

Linezolid was identified as potential nephrotoxin in the population and was responsible for clinical nephrotoxicity in eight instances (patient-days) in the test dataset. Linezolid is often administered as an alternative antibiotic when patients receiving Vancomycin show nephrotoxicity as they cover a similar spectrum. However, after the literature review, we found some studies insisting that the nephrotoxicity profile did not show statistical significance compared to the Vancomycin and the drug label did not provide any dosing adjustment suggestions for patients with renal impairment. Because one of the metabolites in the linezolid is cleared by the kidney, the result warrants further analysis on Linezolid's nephrotoxicity.

4.5. Discussion

As shown in the results, the proposed model was able to show promising accuracy in predicting the next-day SCr distribution based on the factual antibiotics administration. Moreover, the estimated nephrotoxicity of administered antibiotics tends to agree with the medical consensus regarding known nephrotoxicity. By assuming the model generated clinically sound estimates of both the effect-free SCr trajectory and the estimated nephrotoxicity due to the planned administration of antibiotics, we further analyzed how patients progressed when the response model predicted comparatively higher nephrotoxicity from the factual antibiotic administrations. During the analysis, we evaluated patient prognoses regarding cumulated nephrotoxicity estimates $\sum_i \widehat{modif}_{d_i}$ throughout their ICU stay. Higher cumulated SCr increment due to administered antibiotics was expected to present adverse renal prognoses compared to patients with less cumulated SCr increments, and the difference was assumed to be observed through their discharge diagnoses. The average cumulated SCr increment was 0.4192 mg/dL with the standard deviation of 1.4094 from the 1,057 hospital admissions accompanied with at least one antibiotic administration during the patient's ICU stay in the test set.

We examined whether patients discharged with kidney-related diseases would show higher cumulated SCr increments due to the planned administration of antibiotics compared to the population. By conducting a one-tailed *t*-test under the significance level $\alpha = 0.05$ and the power $1 - \beta = 0.8$, acute kidney failure with tubular necrosis and hypertensive chronic kidney disease showed statistical significance (Table 4.6). For patients discharged with acute kidney injury with tubular necrosis, the earlier study showed that the majority of drug-induced renal impairment is presented as either acute interstitial nephritis (AIN) or acute tubular necrosis (ATN) [107], where the etiology of AIN tends to be allergic reactions due to the choice of medication, while ATN is more dose-dependent [118]. Therefore, we suspect that dose-related renal impairment might be the more prevalent cause of antibiotic-induced nephrotoxicity in the dataset. For patients discharged with hypotensive chronic kidney diseases (CKD), it could be possible that a patient's existing renal impairment might be worsened during antibiotic therapies. Moreover, as CKD patients receive dialysis regularly, an earlier study [119] reported that they found a higher prevalence of gram-negative infections compared to the population due to 1) a higher frequency of cefazolin use [120], 2) clustering in hemodialysis units, and 3) the presence of in-dwelling catheters [121], thereby having higher propensity for gram-negative infections and receiving the antibiotics targeted to the pathogen.

ICD-9 CODE	Avg. Cumulated Risk	Std. Dev.	# of admission	Diagnosis	p-value
Population	0.4192	1.4094	1057		
584.5	1.6811	2.2758	78	Acute kidney failure with lesion of tubular necrosis	<0.001
403.91	1.2212	1.8788	38	Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage V or end-stage renal disease	<0.001
584.9	0.5877	1.6943	177	Acute kidney failure, unspecified	<0.001
403.9	0.4737	1.0035	96	Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage I through stage IV, or unspecified	<0.001

Table 4.6. Kidney-related ICD-9 codes with statistically significant cumulated SCr increments.

4.6. Case study – Applying design to identify organ-toxic interventions using risk estimates from AOFI models

In the sections above, the proposed framework was able to demonstrate the potential of decomposing the effect of selected clinical interventions (i.e., antibiotics) from the physiological observations (i.e., serum creatinine level). Moreover, the comparative analysis on models trained with and without the consideration of antibiotics therapy showed similar mean absolute error when predicting next-day SCr level, which indicates the training of the response model did not degrade the overall prediction performance. Lastly, the predictions from the response model tend to agree with the medical consensus regarding the nephrotoxicity of antibiotics.

By extending the idea, we examined whether the framework could also decompose the potential adverse effect of frequently administered clinical interventions from the risk estimated by AOFI models. The trajectory modeling approach discussed in Chapter 3 showed the potential for improving the performance of FTR event prediction when risks driven from myopic evidence are used as an input. Therefore, we also aimed to evaluate whether a more complex trajectory modeling architecture based on the attention-based, single-headed transformer [122] could improve the performance of predicting AOF onset. In the experiment, we aimed to quantify the potential toxicity of clinical interventions frequently documented for patients admitted to ICUs. To compare the performance with other modeling approaches presented in the earlier chapters, we used the same training, validation, and test dataset as Chapter 2 from the MIMIC-3 dataset.

To implement the framework to take the risk estimates from AOFI models as an input, it required modification of the design proposed in Section 4.3.2. First, risk estimates from the AOFI models are probabilistic quantities while the physiological inputs (i.e., SCr levels) were used during the nephrotoxicity analysis. Therefore, the predicted risks from both the effect-free model and the response model should also be bounded within [0,1] range in the experiment. Moreover, to simplify the current experiment, we only estimated potential adverse effects of clinical intervention administrations without considering the dosing-dependent response. Therefore, intervention inputs were represented as a vector of binary variables. Accordingly, outputs of both the effect-free and the response model should be merged probabilistically compared to the simple addition we used in the nephrotoxicity analysis. The details of the modification are discussed in the following sections.

4.6.1. Method

4.6.1.1. Input

In the experiment, patient risk estimated from AOFI models on the day d_i was noted as $x_{AOF,d_i} \in [0,1]^4$. For the intervention input, we considered the 100 most commonly documented clinical interventions from patients admitted to the ICUs, and noted them as $x_{intv,d_i} \in \{0,1\}^{100}$ by representing whether each clinical intervention was administered to the patient on the day d_i or not. As mentioned above, we did not consider the dosing-dependent response of each clinical intervention.

4.6.1.2. Changes in the likelihood of AOF onset with and without clinical interventions

Due to the changes in both physiologic and intervention inputs, estimates from the effect-free and the response model should be represented as a probability. First, outputs of the effect-free model, $f^{AOFI}_{EFT,\theta'}$, should be bounded within [0,1]. Therefore, we used the sigmoid activation function on the output nodes for the implementation. The probability of developing each of the four AOFs without any clinical intervention, $P_{AOF,baseline,d_i}$, could be written as follows:

$$P_{AOF,baseline,d_i} = f^{AOFI}_{EFT,\theta'}(< x_{AOF,d_{i-1}}, h_{d_i} >) \in [0,1]^4$$

In addition, as the estimates from the response model should be represented as a probability as well, the output of the response model, $g^{AOFI}_{RES,\phi'}$, should also be bounded within [0,1]. Therefore, the sigmoid activation function on the output nodes was used for the implementation as well. With the submodularity constraint mentioned in Section 4.3.2, the probability of developing each AOF due to the clinical intervention x_{intv,d_i}, $P_{AOF,intv,d_i}$, can be written as follows:

$$P_{AOF,intv,d_i} = \begin{cases} g^{AOFI}_{RES,\phi'}(< x_{intv,d_i}, h_{d_i}, x_{AOF,d_{i-1}} >) \in [0,1]^4 \; when \; \left|\left|x_{intv,d_i}\right|\right|_2 \neq 0 \\ 0 \; when \; \left|\left|x_{intv,d_i}\right|\right|_2 = 0 \end{cases}$$

4.6.1.3. Merging two risk estimates—baseline AOF risks and risk increments due to clinical interventions

In the nephrotoxicity experiment, we assumed that the nephrotoxicity incurred by the antibiotics could be represented as an additive effect on the baseline estimates. Therefore, estimates from the effect-free model and the response model were added to estimate of the next-day SCr distribution. Since the current experiment has two probabilistic estimates to calculate the risk of

AOF onsets on the next day, $P_{AOF,intv,d_i}$ and $P_{AOF,baseline,d_i}$, they should be merged within the probabilistic framework to calculate the final probability $P_{AOF,final,d_i}$.

To estimate $P_{AOF,final,d_i}$, we assumed there exists a prior probability of the target AOF caused by clinical interventions, and estimated $P_{AOF,byIntv} \in [0,1]^4$ during the training. Then, we also hypothesized that the patient's AOF onset could be either explained through the baseline estimates, $(1 - P_{AOF,byIntv}) * P_{AOF,baseline,d_i}$, or by the clinical interventions scheduled to be administered, $P_{AOF,byIntv} * P_{AOF,intv,d_i}$. By merging two estimates as a complement event, the $P_{AOF,final,d_i}$ was defined as follows:

$$P_{AOF,final,d_i} = \frac{1 - \left(1 - P_{AOF,byIntv} * P_{AOF,baseline,d_i}\right)\left(1 - \left(1 - P_{AOF,byIntv}\right) * P_{AOF,intv,d_i}\right)}{\left(1 - P_{AOF,byIntv} * \left(1 - P_{AOF,byIntv}\right)\right)}$$

4.6.1.4. Loss function

In the nephrotoxicity analysis, the loss was calculated based on the observed next-day SCr distribution with KL-divergence. However, there was no gold standard information to evaluate the instance-level prediction $P_{AOF,final,d_i}$ in the current experiment. Therefore, as we presented in Section 2.2.5, we used discharge diagnoses as a gold standard instead and defined the loss to compare the aggregated probability using instance-level risk estimates $P_{AOF,final,d_i}$ and the discharge diagnoses. The aggregated probability $P_{AOF,agg,intv+baseline}$ represented the probability of a patient developing the target AOF at least once during their ICU stay with the consideration of the adverse effect of clinical interventions, and was represented as follows:

$$P_{AOF,agg,intv+baseline} = 1 - \prod_i \left(1 - P_{AOF,final,d_i}\right) \in [0,1]^4$$

Moreover, we also designed the estimated instance-level probabilities from the effect-free model, $P_{AOF,baseline,d_i}$, to follow the gold-standard information as well with the assumption that the intervention-induced AOF is comparably less than the AOF induced by a patient's baseline physiology. Therefore, the aggregated probability only using the effect-free estimates $P_{AOF,agg,baseline}$ was also defined and considered into the loss function:

$$P_{AOF,agg,baseline} = 1 - \prod_i \left(1 - P_{AOF,baseline,d_i}\right) \in [0,1]^4$$

Finally, the loss function was defined as follows:

$$\begin{aligned} loss(\theta', \phi', \lambda, weight_{final}, weight_{baseline}) \\ = \sum_{pts} [&WeightedCrossEntropy(y_{pts}, P_{AOF,agg,intv+baseline}, weight_{final}) \\ &+ WeightedCrossEntropy(y_{pts}, P_{AOF,agg,baseline}, weight_{baseline})] \\ &+ regularizer(\lambda) \end{aligned}$$

where $y_{pts} \in \{0,1\}^4$ represents the discharge diagnoses that we used as gold-standard information, $weight_{baseline}$ and $weight_{final}$ quantify to what extent the model should

emphasize learning $P_{AOF,agg,baseline}$ and $P_{AOF,agg,intv+baseline}$, respectively, and θ' and ϕ' are parameters of the effect-free model f_{EFT}^{AOFI} and the response model g_{RES}^{AOFI}, respectively. Dropout on the f_{EFT}^{AOFI} and g_{RES}^{AOFI} was used as a regularizer for the training.

4.6.1.5. Evaluation

We assumed that the adverse influence of each clinical intervention could be quantified for each instance using the later component of the $P_{AOF,final,d_i}$ calculation, $P_{AOF,byIntv} * P_{AOF,intv,d_i}$. In cases where the intervention was responsible for the target AOF onset, we assumed that the risk estimated for the factual clinical intervention $P_{AOF,byIntv} * P_{AOF,intv,d_i}$ would be significantly higher than the risk estimated without the target intervention $P_{AOF,byIntv} * P_{AOF,intv',d_i}$, where $intv' = intv/\{TargetIntervention\}$. To quantify the potential harm caused by the target intervention for each instance, the probability of the patient developing AOF due to the target clinical intervention was estimated as the product event of 1) the probability of the patient developing the target AOF based on the factual clinical intervention, and 2) the probability of the patient not developing the target AOF when the target intervention is not administered. Therefore, we defined the aforementioned quantity as a risk score, $RiskScore_{d_i,TargetIntervention}$, written as follows:

$$RiskScore_{d_i,TargetIntervention} = \left(P_{AOF,byIntv} * P_{AOF,intv,d_i}\right) * \left(1 - P_{AOF,byIntv} * P_{AOF,intv',d_i}\right)$$

During the analysis, we first analyzed the overall risk trends of each $TargetIntervention$ estimated in $RiskScore_{d_i,TargetIntervention}$ from the population level through summary statistics. Then, we demonstrated how $RiskScore_{d_i,TargetIntervention}$ estimated on each instance and intervention could be used to identify cases where the administered clinical interventions can explain the increased risk of the target AOF onset.

4.6.2. Results

To evaluate the performance changes with and without the decomposition, we also trained another prediction model solely based on the risk estimates from AOFI models without considering the effect of interventions. In order to train only the effect-free model $f_{EFT,\theta_{\neg intv}}^{AOFI}$, we used the following loss function:

$$loss_{baseline}(\theta_{\neg intv}, \lambda') = \sum_{pts}\left[CrossEntropy\left(y_{pts}, P_{AOF,agg,baseline}\right)\right] + regularizer(\lambda')$$

The best model on the proposed framework and the best $f_{EFT,\theta_{\neg intv}}^{AOFI}$ model were selected where they achieved the highest micro-F1 score in the validation set. The performance of predicting the AOF onsets based on patient-level predictions from each model is presented in Table 4.7.

(a) Patient-level predictions from different trajectory modeling settings

	f_{EFT}^{AOFI}			f_{EFT}^{AOFI} and g_{RES}^{AOFI} (proposed)			$f_{EFT,\theta_{\neg intv}}^{AOFI}$ (baseline)		
	PREC	REC	F1	PREC	REC	F1	PREC	REC	F1
AHF	0.3697	0.2919	0.3262	**0.4161**	**0.3206**	**0.3622**	0.3694	0.2775	0.3169
ALI	0.6311	0.6525	0.6417	**0.6657**	**0.6808**	**0.6732**	0.6543	0.6469	0.6506
AKI	0.6171	0.4746	0.5365	0.5724	**0.5360**	0.5536	**0.6286**	0.5021	**0.5583**
ALF	0.3182	0.1707	0.2222	**0.3200**	**0.1951**	**0.2424**	0.2571	0.2195	0.2368
Micro	0.5710	0.4861	0.5251	0.5747	**0.5288**	**0.5508**	**0.5780**	0.4954	0.5343

(b) Patient-level prediction performance on modeling approaches presented in the earlier sections

	Original AOFI			HMM w/ AOFI			Adjusted AOFI		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
AHF	**0.3780**	0.2211	**0.2791**	0.2608	**0.5324**	0.2608	0.2342	0.5324	0.2342
ALI	**0.6552**	0.6080	**0.6307**	0.4717	**0.8000**	0.4717	0.4769	0.8000	0.4769
AKI	**0.7302**	0.3644	0.4861	0.4928	**0.6152**	**0.4928**	0.4752	0.6152	0.4752
ALF	**0.2745**	0.3111	**0.2917**	0.1702	**0.3556**	0.1702	0.2353	0.3556	0.2353
Micro	**0.6093**	0.4151	0.4938	0.4114	**0.6502**	**0.5039**	0.4023	0.6502	0.4971

Table 4.7. Patient-level performance comparison between (a) the proposed model and $f_{EFT,\theta_{\neg intv}}^{AOFI}$, (b) the HMM and the AOFI models.

As the result shows, the model with consideration of the clinical intervention (considering both f_{EFT}^{AOFI} and g_{RES}^{AOFI} to predict the likelihood of AOF onsets) showed a higher micro-F1 score compared to the model only based on the risk estimates from AOFI models ($f_{EFT,\theta_{\neg intv}}^{AOFI}$). The trained model estimated prior probabilities of developing AOF due to the clinical interventions for each AOF as following: 0.5723 for AHF, 0.5831 for ALI, 0.5133 for AKI and 0.0151 for ALF. Accordingly, the model relied on the baseline physiology and the estimates from both physiology and the clinical interventions with similar weights when predicting the target AOF onsets while putting more weight on the latter, except in the case of ALF.

4.6.3. Discussion

We first calculated $RiskScore_{d_i,TargetIntervention}$ on each instance to quantify how much risk increment does each intervention could explain. Table 4.8 shows five intervention-AOF pairs that showed the highest risk scores in descending order.

AOF	Average $RiskScore$	Std. Dev. $RiskScore$	Clinical intervention
AKI	0.3315	0.1186	Heparin Sodium (Prophylaxis)
ALI	0.1528	0.1349	Heparin Sodium (Prophylaxis)
ALI	0.1327	0.1270	5% Dextrose 0.45% Normal Saline
ALI	0.1326	0.1269	Ranitidine (Prophylaxis)
ALI	0.0888	0.1093	Pantoprazole (Protonix)

Table 4.8. The list of clinical intervention-AOF pairs with higher average risk score from the test set

Compared to the nephrotoxicity analysis presented in earlier sections, we cannot strictly assume that the intervention with a higher risk score directly implies its potential organ toxicity. Instead, we can consider that these interventions could explain the risk increment that cannot be explained via physiologic evidence by one of the following: 1) the intervention exerted a deleterious effect regarding the target AOF onset, 2) the intervention was administered to the comorbidity of the target AOF that develops acutely, or 3) the intervention was relevant to non-physiological events (e.g., initiation or maintenance procedures). For example, the model estimated a high average risk score on prophylactic heparin sodium administration for AKI and ALI. After the literature review, we found that there should be special caution on patients with renal impairment during anticoagulation therapy because the therapy can increase the bleeding risk [123]. In contrast, for the patients with high risk of ALI, prior studies insist that prophylactic heparin sodium administered to them with in nebulized form improved the outcomes [124]. If this is the case, the intervention administration might indicate that physicians are already aware of the patient's high risk of ALI. For isotonic 5% dextrose and 0.45% normal saline, although it is often used as a maintenance fluid for patients admitted to ICUs, the recent study suggests that patients with a high risk of ALI showed a better prognosis when hypertonic fluid was used compared to fluids with other osmolality, including isotonic and hypotonic fluids [125]. Therefore, the risk increment due to the choice of maintenance fluid on patients with high risk of ALI could be reviewed to verify whether the osmolality could be the risk factor for the increased risk of ALI. Lastly, ranitidine and pantoprazole are frequently administered to prevent stress ulcer in the ICU setting. Although stress ulcer prophylaxis is recommended in high-risk patients to prevent stress-related mucosal disease [126], those interventions could also alter the stomach pH of patients, and the disturbed gut flora due to the change of acidity level might be related to the increased risk of infections [127]. Therefore, it might be worth reviewing intervention-AOF pairs with high average risk scores.

The risk score estimated for each intervention and each patient could be used to identify instances for additional reviews. Similar to the analysis conducted in the population level, instance-intervention pairs with higher risk score could be manually reviewed to clarify the causal relationship between them. Table 4.9 shows 20 intervention-instance pairs that showed the highest risk score for the corresponding AOF. Similar to the population-level analysis, the risk estimated for the clinical interventions can be evaluated with the assumption of 1) actual risk incurred by the clinical intervention administration (e.g., ALI and pantoprazole administration), 2) risk incurred by other clinical events that necessitates the identified clinical interventions (e.g., 16-gauge needle insertion for either transfusion or bolus therapy, nasal swab for the suspected infections, or fentanyl for relieving a patient's ongoing severe pain), or 3) non-physiological interventions potentially indicating the patient's sudden and severe deterioration of the target AOF (e.g., family updates by the RN and OR received).

Counterfactual risk estimates	Factual risk estimates	*Riskscore*	AOF	Intervention
0.0023	0.5819	0.5805	ALI	Family updated by RN
0.0081	0.5831	0.5783	ALI	Nasal Swab
0.0107	0.5826	0.5764	ALI	OR Received
0.0076	0.5742	0.5698	ALI	Family updated by RN
0.0153	0.5783	0.5695	ALI	Family updated by RN
0.0219	0.5820	0.5693	ALI	Fentanyl
0.0070	0.5731	0.5691	ALI	Urine Culture
0.0174	0.5786	0.5685	ALI	16 Gauge
0.0074	0.5714	0.5672	ALI	Family updated by RN
0.0237	0.5803	0.5665	ALI	Fentanyl
0.0023	0.5651	0.5638	ALI	Family updated by RN
0.0118	0.5699	0.5632	ALI	Family updated by RN
0.0264	0.5781	0.5628	ALI	Magnesium Sulfate
0.0217	0.5724	0.5600	ALI	Fentanyl
0.0320	0.5756	0.5572	ALI	Fentanyl
0.0310	0.5750	0.5571	ALI	Magnesium Sulfate
0.0102	0.5610	0.5552	ALI	Family updated by RN
0.0079	0.5572	0.5528	ALI	Nasal Swab
0.0042	0.5545	0.5522	ALI	Family updated by RN
0.0346	0.5715	0.5517	ALI	Pantoprazole (Protonix)

Table 4.9. Top-20 instance-intervention pairs with high predicted risks.

4.6.4. Potential measures to improve performance with expert knowledge

The proposed modeling framework showed improved prediction performance of patient-level AOF onsets, and the model was able to identify a few clinical interventions that could explain the additional risk increment which could not be explained through AOFI models. Similar to the suggestions provided in Section 3.6.3, we believe the model's performance can be improved through manual review from experts, and the predictions from the proposed framework would facilitate the reviewing process.

First, the summary statistics of the risk scores for each clinical intervention can be used as a screener identifying intervention-AOF pairs for further analysis. For example, if the review reveals that prophylactic treatments with higher risk scores, such as heparin sodium or ranitidine, are not relevant to the target AOF onset, physicians could re-train the model after excluding these interventions from the candidate intervention list. In case the specific physiological states that require such clinical interventions can be defined, additional indicator variable could be added during expert annotation process. In the prediction time, such information could be provided to the model as background information so that the model could adjust the risk estimated by myopic evidence to derive a refined estimation.

Second, as the framework was able to derive the quantitative estimates of the potential adverse effect of each intervention on each instance, physicians could focus on reviewing instances with higher risk score from the target clinical intervention. If the manual review could differentiate two possible explanations of the predicted high risk, either from the patient's baseline physiology

or from the combined effect of the administered clinical interventions and the patient's baseline physiology, the corresponding annotation can be integrated into the loss function as follows:

$$loss' = loss(\theta', \phi', \lambda, weight_{final}, weight_{baseline}) + \sum_{instance} CrossEntropy(y_{instance,baseline}, P_{AOF,baseline,d_i}) + \sum_{instance} CrossEntropy(y_{instance,intv+baseline}, P_{AOF,final,d_i})$$

, where $y_{instance,baseline}$ represents gold-standard information for the target AOF onset based on baseline physiology, and $y_{instance,intv+baseline}$ is an indicator variable representing whether the AOF was incurred due to the clinical interventions and/or the baseline physiology.

4.7. Limitations

The study did not have gold-standard information regarding the nephrotoxicity of individual antibiotic from factual administrations. Therefore, we instead maximized the likelihood of predicted SCr distribution with the factual antibiotic administrations given the observed next-day SCr distribution. Although we validated the trends of predicted SCr increments from potentially nephrotoxic antibiotics through literature review, the prediction accuracy of the factual SCr trajectory and the corresponding nephrotoxicity estimates could be improved with gold-standard information on the nephrotoxicity of the factual antibiotic administration. By adding additional penalty terms comparing predicted nephrotoxicity from $g_{res,\phi}$ with the gold-standard information, it might be possible to improve the accuracy of estimates on both effect-free estimates and outputs from the response model.

4.8. Conclusion

In this study, we showed that the nephrotoxic effect of antibiotics and the baseline renal function trajectory could be decomposed by jointly training two submodels explaining the effect-free trajectory and the response of the intervention. The model showed promising accuracy for predicting the next-day SCr distribution based on the factual antibiotic administration plans and patients' previous SCr level measurements. Moreover, the response model was able to provide the list of antibiotics that are potentially nephrotoxic with quantitative estimate from the dataset. We also verified that the pattern learned from the model agrees with medical consensus; the trained model predicted higher cumulated nephrotoxicity in patients with chronic renal impairment or patients who developed acute tubular necrosis during their hospital stay. Although there could be cases where controlling a patient's infection outweighs managing a patient's renal prognosis within intensive care settings, we believe risk estimates on antibiotic-induced nephrotoxicity in the context of a patient's current renal function trajectory will allow physicians to quantitatively compare candidate antibiotic treatment options.

Chapter 5. Conclusion

In the previous chapters, I presented the framework for deriving a risk estimation model for acute-onset diseases by only specifying the corresponding discharge diagnoses, the framework that can improve risk estimates based on myopic evidence by considering trajectory on relevant FTR events altogether, and the framework that can quantify the potential adverse effect of clinical interventions administered to patients. Although the results presented in each chapter showed the potential of the proposed frameworks, the findings still emphasize the importance of expert knowledge in order to derive a more mature risk estimation model. Still, risks estimated from the proposed frameworks were able to demonstrate that they followed clinical consensus on the risk of target acute-onset diseases. Therefore, in case physicians are involved in manually reviewing the predictions for clinical validation, those predictions could be utilized as building blocks for implementing risk prediction models with better accuracy. In the following, I will summarize the findings from each aim and how the proposed framework could be used in today's clinical practice and clinical research.

5.1. Summary of the research findings

Aim 1: In Chapter 2, I provided a framework that can derive the risk estimates by only specifying the groups of discharge diagnoses that are relevant to the target acute-onset disease. The framework first identified relevant clinical interventions. By using the timing of clinical interventions and the discharge diagnoses as a proxy event of the target disease onset, we trained a model to predict the proxy events in a supervised setting. Analyses of the risks estimated by the AOFI models showed that the likelihood of receiving relevant clinical intervention within the next 24 hours and being discharged with the target disease could be used as a risk estimate for the disease.

Aim 2: In Chapter 3, I presented the modeling approach that can integrate risk trajectories of potentially relevant clinical outcomes, which were estimated based on myopic evidence, to the target acute-onset disease. The results showed that the screening performance of existing EWSs could be improved by considering temporal trends of risk estimates of the relevant clinical outcomes. Moreover, although the approach was not able to show improved micro-F1 scores when risk estimates from the AOFI model were considered as an EWS, it still showed the potential to improve the prediction performance of high-risk AOF patients with improved patient-level recall and less positive prediction on instance-level. The clinical validation showed that the estimated risks agreed with the clinical consensus by showing a worse prognosis when the model made a positive instance-level prediction compared to the other instances based on biomarkers used to diagnose the target AOFs.

Aim 3: In Chapter 4, I demonstrated that both physiologic variables and risk estimates from patients could be decomposed into the baseline physiology, a patient's physiologic state without any clinical interventions, and the adverse effect due to the clinical interventions. Moreover, the adverse effect estimated by the model was able to show not only the trend of risk exerted from each clinical intervention in the population level but also the potential of identifying cases that might develop the AOFs due to the clinical intervention.

5.2. Potential measures to improve the accuracy of the trained model

From all risk estimates driven by models discussed in this dissertation, results indicate that expert knowledge should be integrated in order to generate more clinically sound risk estimations. As each of the proposed frameworks has a different modeling architecture, there could be different avenues through which such knowledge could be incorporated.

Aim 1: The framework only requires discharge diagnoses for the target disease to derive the AOFI model. After the training, the training procedure provides the list of clinical interventions that were used as proxy events. Therefore, physicians could modify the list of interventions by adding new clinical interventions that are commonly administered to patients with the disease or by removing the intervention if it is irrelevant to the target disease. Moreover, clinicians may review the feature importance learned by the model (e.g., information gain) and remove the clinical variables that are either not routinely measured or unintuitive in the current workflow. Lastly, the proxy labels on the training dataset could be refined through additional manual annotation from experts. Instead of annotating all instance-level observations, experts can focus on reviewing instances where the current model showed higher uncertainty [128]. This would allow practitioners to have a risk estimation model with higher accuracy and less annotation labor compared to the conventional methods.

Aim 2: As mentioned in Section 3.6.3, the trained HMM delivered quantitative estimates on the prior probability, transition probabilities, and the dependency of the trajectory information for each FTR event state specified for the training. Therefore, physicians may review such quantitative estimates and refine them by adding additional penalty terms for the re-training. This would allow models to utilize expert knowledge so that they can train the model with not only better performance but also more straightforward to the current medical understanding.

Aim 3: In the nephrotoxicity study presented in Chapter 4, the trained model estimates the next-day SCr level without any clinical intervention and the SCr increment due to the administered ABXs, then combine the two quantities to estimate the next-day SCr distribution. Moreover, the model trained with AOFI risk estimates was aimed at representing the baseline risk of AOFs with the potential risk increments resulting from the clinical interventions frequently administered in ICU settings. As the model decomposes a patient's prognosis in terms of the baseline physiology and the influence of clinical interventions, experts could focus on validating the instances with higher predicted risks from either the response model or the final risk estimation. If the expert annotation could clarify the potential FTR incidents caused by the clinical interventions, the model's performance could be improved through iterative training.

5.3. Potential usage of the model presented in the dissertation

Aim 1: As the model only requires the discharge diagnoses for the model derivation, it could serve as a baseline modeling approach to quickly evaluate the predictability of the target disease onsets based on the available dataset, and the resulting model could be improved by with active learning methods [128]. Since the proposed approach is able to prioritize patients using estimated risks from the model, the framework would be able to provide cases that require additional review based on expert knowledge. By using the proposed approach, this would allow

practitioners to implement a model with better accuracy and less annotation cost for the event of interest compared to the current practice on deriving risk prediction models.

The simple derivation of risk model from the disease of interest could also be utilized to characterize both risk factors and biomarkers for the disease. As the framework only requires the group of patients and the disease of interest, researchers could derive a separate AOFI model for the target disease on different cohorts, then compare how the risk profile varies in order to further their understanding of the difference in the prognosis of each cohort. As the framework could provide the feature contribution of the model for the prediction, it can be used as a tool for biomarker discovery for the disease of interest. This could also serve as a list of candidate physiological variables that experts need to evaluate when determining EWSs for the target disease.

After the analysis of the risks from AOFI models, we found prior studies indicating that the quality of the current discharge diagnoses coding practice is sub-optimal. Lo Re et al. reported that the accuracy of ALF coding was too low after their manual chart review, where the predicted positive value on the disease ranged from 5% to 15% [129]. As the prediction results from ALF AOFI models in both datasets showed, the patient-level prediction accuracy is significantly lower than the performance observed in other AOFs, while risk estimates from the model showed a positive correlation with the gold-standard variable, total bilirubin level. Therefore, we believe this could indicate that either the quality of ALF discharge diagnoses would limit the model's optimization process, or the performance of the ALF AOFI model was significantly underestimated. Therefore, predictions from the AOFI models could be used in the current setting by comparing the discharge diagnoses assigned to patients at the time of discharge and the patient's risk estimated by AOFI models. If the institution could implement additional verification procedures for cases that do not agree, it would improve the quality of discharge diagnoses, which could lend itself to maximizing the reimbursement rate and maintaining the quality of discharge diagnoses that could be used for the further research, such as EHR-based cohort analysis. With the better quality of discharge diagnoses, practitioners could also re-train the model to achieve better accuracy in the prediction tasks.

Aim 2: The HMM discussed in Chapter 3 was able to show that the screening performance of the existing EWS could be improved by considering trajectory information for relevant FTR events. When institutions deploy different, but potentially relevant, EWSs for their own surveillance purposes, the proposed HMM model would allow them to improve the performance of screening for patients at high-risk. Moreover, as the transition matrix and the prior probability would allow them to fine-tune models based on expert knowledge, it would generate a more straightforward risk estimation for the practitioners. In cases where the predictions from the HMM cannot be directly attributed to regulation issues, the predictions from the HMM can be used along with the EWS screening criteria to decrease the chance of FTR events. Moreover, predictions with improved performance could be used to automate the lab test orders by adding gold-standard tests for the disease for which the model predicted a high risk. By doing so, physicians could review the test results when it is necessary and decrease the delay between the onset of the disease and the confirmation process.

Aim 3: The quantitative estimates of the potential adverse effect of clinical interventions would serve as a tool for researchers focusing on drug safety to provide a candidate list of interventions they could evaluate. If the model predicted higher risk increments for the clinical intervention

across the population, the causal relationship between the clinical intervention and the disease could be further investigated. Moreover, if higher-risk increments were estimated for clinical intervention for the specific cohorts, researchers could further investigate what risk factors would be involved for the potential side effect of the intervention and derive a clinical guideline for intervention administration, if there were any. If the model could achieve the desired accuracy for the clinical practice, it could serve as an alerting tool for physicians so that they can either avoid potentially harmful clinical intervention in the patient's status quo or prepare reactive measures for the potential side effects. If such interventions should be applied, physicians could document the reasoning for the administration so that it can be reviewed when needed.

References

[1] J. van den Bos, K. Rustagi, T. Gray, M. Halford, E. Ziemkiewicz, and J. Shreve, "The $17.1 billion problem: The annual cost of measurable medical errors," Health Aff., vol. 30, no. 4, pp. 596–603, 2011.

[2] M. A. Makary and M. Daniel, "Medical error-the third leading cause of death in the US," BMJ, vol. 353, 2016.

[3] I. of M. (US) C. on Q. of H. C. in America, L. T. Kohn, J. M. Corrigan, and M. S. Donaldson, *To Err is Human*. 2000.

[4] A. B. Haynes *et al.*, "A Surgical Safety Checklist to Reduce Morbidity and Mortality in a Global Population," *N. Engl. J. Med.*, vol. 360, no. 5, pp. 491–499, Jan. 2009.

[5] "Never Events CMS."

[6] F. H. Morriss *et al.*, "Effectiveness of a barcode medication administration system in reducing preventable adverse drug events in a neonatal intensive care unit: a prospective cohort study.," *J. Pediatr.*, vol. 154, no. 3, pp. 363–8, 368.e1, Mar. 2009.

[7] "Effect of Barcode-assisted Medication Administration on Emergency Department Medication Errors."

[8] "Tragic Errors: Usability and Electronic Health Records."

[9] S. K. Aberegg, E. F. Haponik, and P. B. Terry, "Omission Bias and Decision Making in Pulmonary and Critical Care Medicine," *Chest*, vol. 128, no. 3, pp. 1497–1505, Sep. 2005.

[10] B. S. Dean, E. L. Allan, N. D. Barber, and K. N. Barker, "Comparison of medication errors in an American and a British hospital," *Am. J. Heal. Pharm.*, vol. 52, no. 22, pp. 2543–2549, Nov. 1995.

[11] "Hospital and Patient Characteristics Associated with Death after Surgery: A Study of Adverse Occurrence and Failure to Rescue on JSTOR." [Online]. Available: https://www.jstor.org/stable/3765780?seq=1#metadata_info_tab_contents. [Accessed: 19-Sep-2019].

[12] J. H. Silber, P. S. Romano, A. K. Rosen, Y. Wang, O. Even-Shoshan, and K. G. Volpp, "Failure-to-rescue: comparing definitions to measure quality of care.," *Med. Care*, vol. 45, no. 10, pp. 918–25, 2007.

[13] M. Farquhar, "AHRQ Quality Indicators," in *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*, 2008, p. 1403.

[14] T. Isaac and A. K. Jha, "Are Patient Safety Indicators Related to Widely Used Measures of Hospital Quality?," *J. Gen. Intern. Med.*, vol. 23, no. 9, pp. 1373–1378, Sep. 2008.

[15] M. E. Warfield, "A Cost-Effectiveness Analysis of Early Intervention Services in Massachusetts: Implications for Policy," *Educ. Eval. Policy Anal.*, vol. 16, no. 1, pp. 87–99, Mar. 1994.

[16] V. a Ferraris, M. Bolanos, J. T. Martin, A. Mahan, and S. P. Saha, "Identification of Patients With Postoperative Complications Who Are at Risk for Failure to Rescue," *JAMA Surg.*, vol. 149, no. 11, pp. 1103–1108, 2014.

[17] M. Yamamoto, S. Ishikawa, and K. Makita, "Medication errors in anesthesia: an 8-year retrospective analysis at an urban university hospital," *J. Anesth.*, vol. 22, no. 3, pp. 248–

252, Aug. 2008.

[18] J. Callen, J. McIntosh, and J. Li, “Accuracy of medication documentation in hospital discharge summaries: A retrospective analysis of medication transcription errors in manual and electronic discharge summaries,” *Int. J. Med. Inform.*, vol. 79, no. 1, pp. 58–64, Jan. 2010.

[19] G. Teasdale and B. Jennett, “ASSESSMENT OF COMA AND IMPAIRED CONSCIOUSNESS. A Practical Scale,” *Lancet*, vol. 304, no. 7872, pp. 81–84, 1974.

[20] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, “Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today’s critically ill patients*,” *Crit. Care Med.*, vol. 34, no. 5, pp. 1297–1310, 2006.

[21] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, “Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records,” *Sci. Rep.*, vol. 6, no. 1, p. 26094, 2016.

[22] C. A. Bejan, F. Xia, L. Vanderwende, M. M. Wurfel, and M. Yetisgen-Yildiz, “Pneumonia identification using statistical feature selection,” *J. Am. Med. Informatics Assoc.*, vol. 19, no. 5, pp. 817–823, Sep. 2012.

[23] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, “A targeted real-time early warning score (TREWScore) for septic shock,” *Sci. Transl. Med.*, vol. 7, no. 299, pp. 299ra122-299ra122, Aug. 2015.

[24] H. Suresh, N. Hunt, A. Johnson, L. A. Celi, P. Szolovits, and M. Ghassemi, “CLINICAL EVENT PREDICTION AND UNDERSTANDING USING DEEP NETWORKS Clinical Intervention Prediction and Understanding using Deep Networks.”

[25] J. C. Marshall, D. J. Cook, N. V. Christou, G. R. Bernard, C. L. Sprung, and W. J. Sibbald, “Multiple organ dysfunction score: A reliable descriptor of a complex clinical outcome,” *Critical Care Medicine*. 1995.

[26] J. H. Silber, S. V Williams, H. Krakauer, and J. S. Schwartz, “Hospital and patient characteristics associated with death after surgery. A study of adverse occurrence and failure to rescue.,” *Med. Care*, vol. 30, no. 7, pp. 615–29, 1992.

[27] R. Paterson *et al.*, “Prediction of in-hospital mortality and length of stay using an early warning scoring system: Clinical audit,” *Clin. Med. J. R. Coll. Physicians London*, 2006.

[28] J.-R. LE GALL *et al.*, “A simplified acute physiology score for ICU patients,” *Crit. Care Med.*, vol. 12, no. 11, pp. 975–977, Nov. 1984.

[29] J. L. Vincent *et al.*, “The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure,” *Intensive Care Med.*, vol. 22, no. 7, pp. 707–710, 1996.

[30] F. Gao, T. Melody, D. F. Daniels, S. Giles, and S. Fox, “The impact of compliance with 6-hour and 24-hour sepsis bundles on hospital mortality in patients with severe sepsis: a prospective observational study.,” *Crit. Care*, 2005.

[31] A. Kortgen, P. Niederprüm, and M. Bauer, “Implementation of an evidence-based ‘standard operating procedure’ and outcome in septic shock,” *Crit. Care Med.*, 2006.

[32] A. M. Sawyer *et al.*, “Implementation of a real-time computerized sepsis alert in nonintensive care unit patients,” *Crit. Care Med.*, 2011.

[33] T. Desautels *et al.*, “Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach.,” *JMIR Med. informatics*, vol. 4, no. 3, p. e28, 2016.

[34] M. E. Hock Ong *et al.*, “Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score,” *Crit. Care*, 2012.

[35] M. Singer *et al.*, “The third international consensus definitions for sepsis and septic shock (sepsis-3),” *JAMA - Journal of the American Medical Association*. 2016.

[36] C. P. Subbe, R. G. Davies, E. Williams, P. Rutherford, and L. Gemmell, “Effect of introducing the Modified Early Warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions,” *Anaesthesia*, 2003.

[37] J. Gardner-Thorpe, N. Love, J. Wrightson, S. Walsh, and N. Keeling, “The value of Modified Early Warning Score (MEWS) in surgical in-patients: A prospective observational study,” *Ann. R. Coll. Surg. Engl.*, 2006.

[38] D. S. Char, N. H. Shah, and D. Magnus, “Implementing Machine Learning in Health Care — Addressing Ethical Challenges,” *N. Engl. J. Med.*, 2018.

[39] B. Min, R. Grishman, L. Wan, C. Wang, and D. Gondek, “Distant Supervision for Relation Extraction with an Incomplete Knowledge Base,” in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013.

[40] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP ’09*, 2009.

[41] A. Go, R. Bhayani, and L. Huang, “Twitter Sentiment Classification using Distant Supervision,” *Processing*, 2009.

[42] M. Aczon *et al.*, “Dynamic Mortality Risk Predictions in Pediatric Critical Care Using Recurrent Neural Networks,” 2017.

[43] M. Ghassemi, T. Naumann, T. Brennan, D. a Clifton, and P. Szolovits, “A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse , Heterogeneous Clinical Data,” *Proc. Twenty-Ninth AAAI Conf. Artif. Intell.*, pp. 446–453, 2015.

[44] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, “Learning to Diagnose with LSTM Recurrent Neural Networks,” Nov. 2015.

[45] N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, “Simultaneous Modeling of Multiple Diseases for Mortality Prediction in Acute Hospital Care,” *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD ’15*, pp. 855–864, 2015.

[46] J. Yoon, A. Alaa, S. Hu, and M. Schaar, “ForecastICU: A Prognostic Decision Support System for Timely Prediction of Intensive Care Unit Admission,” *Proc. 33rd Int. Conf. Mach. Learn.*, 2016.

[47] J. Wiens and E. S. Shenoy, “Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology,” *Clin. Infect. Dis.*, vol. 66, no. 1, pp. 149–153, Jan.

2018.

[48] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients,” *Health Aff.*, vol. 33, no. 7, pp. 1123–1131, Jul. 2014.

[49] J. R. Le Gall *et al.*, “Customized probability models for early severe sepsis in adult intensive care patients. Intensive Care Unit Scoring Group.,” *JAMA*, 1995.

[50] F. A. Masoudi *et al.*, “Gender, age, and heart failure with preserved left ventricular systolic function,” *J. Am. Coll. Cardiol.*, 2003.

[51] I. Mehmood, N. Ejaz, M. Sajjad, and S. W. Baik, “Prioritization of brain MRI volumes using medical image perception model and tumor region segmentation,” *Comput. Biol. Med.*, 2013.

[52] R. S. Olson, W. La Cava, Z. Mustahsan, A. Varik, and J. H. Moore, “Data-driven advice for applying machine learning to bioinformatics problems.”

[53] “Allstate Claims Severity | Kaggle.” [Online]. Available: https://www.kaggle.com/c/allstate-claims-severity. [Accessed: 25-Jan-2019].

[54] T. K. Ho, “The random subspace method for constructing decision forests,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998.

[55] R. E. Schapire, “The Boosting Approach to Machine Learning: An Overview,” Springer, New York, NY, 2003, pp. 149–171.

[56] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?,” in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP ’02*, 2002.

[57] D. H. Lee and E. Horvitz, “Predicting Mortality of Intensive Care Patients via Learning about Hazard,” *Proc. 31th Conf. Artif. Intell. (AAAI 2017)*, pp. 4953–4954, 2017.

[58] J. McGaughey, F. Alderdice, R. Fowler, a Kapila, a Mayhew, and M. Moutray, “Outreach and Early Warning Systems (EWS) for the prevention of intensive care admission and death of critically ill adult patients on general hospital wards.,” *Cochrane Database Syst. Rev.*, no. 3, p. CD005529, 2007.

[59] C. for D. C. and P. (Cdc), “ICD-9-CM official guidelines for coding and reporting,” Atlanta, GA, 2011.

[60] E. Shantsila and G. Y. H. Lip, “Thrombotic Complications in Heart Failure,” *Circulation*, vol. 130, no. 5, pp. 387–389, Jul. 2014.

[61] S. M. Hardman and M. R. Cowie, “Fortnightly review: anticoagulation in heart disease.,” *BMJ*, vol. 318, no. 7178, pp. 238–44, Jan. 1999.

[62] “Heparin Sodium, for intravenous use[package insert], Fresenius Kabi, LakeZurich, IL,” 2017.

[63] X. Meng *et al.*, “[seminal] MLlib: Machine Learning in Apache Spark,” *J. Mach. Learn. Res.*, vol. 17, pp. 1–7, 2016.

[64] J. Davis and M. Goadrich, “The relationship between Precision-Recall and ROC curves,” in *Proceedings of the 23rd international conference on Machine learning - ICML ’06*, 2006.

[65] A. E. W. Johnson *et al.*, “MIMIC-III, a freely accessible critical care database,” *Sci. Data*,

vol. 3, 2016.

[66] C. W. Yancy *et al.*, “2017 ACC/AHA/HFSA Focused Update of the 2013 ACCF/AHA Guideline for the Management of Heart Failure,” *J. Am. Coll. Cardiol.*, vol. 70, no. 6, pp. 776–803, 2017.

[67] J. Devaquet *et al.*, “Effects of inspiratory pause on CO2 elimination and arterial PCO2 in acute lung injury.,” *J. Appl. Physiol.*, 2008.

[68] J. a Kellum *et al.*, “KDIGO Clinical Practice Guideline for Acute Kidney Injury,” *Kidney Int. Suppl.*, 2012.

[69] “EASL Clinical Practical Guidelines on the management of acute (fulminant) liver failure,” *J. Hepatol.*, 2017.

[70] R. C. Bone *et al.*, “Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis,” in *Chest*, 1992.

[71] G. Garcia-Tsao, C. R. Parikh, and A. Viola, “Acute kidney injury in cirrhosis,” *Hepatology*, vol. 48, no. 6, pp. 2064–2077, Dec. 2008.

[72] S. K. Asrani, D. A. Simonetto, and P. S. Kamath, “Acute-on-Chronic Liver Failure.,” *Clin. Gastroenterol. Hepatol.*, vol. 13, no. 12, pp. 2128–39, Nov. 2015.

[73] L. Christou, G. Pappas, and M. E. Falagas, “Bacterial Infection-Related Morbidity and Mortality in Cirrhosis,” *Am. J. Gastroenterol.*, vol. 102, no. 7, pp. 1510–1517, Jul. 2007.

[74] M. E. Grams and H. Rabb, “The distant organ effects of acute kidney injury,” *Kidney Int.*, 2012.

[75] G. M. Chertow, E. Burdick, M. Honour, J. V. Bonventre, and D. W. Bates, “Acute Kidney Injury, Mortality, Length of Stay, and Costs in Hospitalized Patients,” *J Am Soc Nephrol*, 2005.

[76] M. Karcz, B. Bankey, D. Schwaiberger, B. Lachmann, and P. J. Papadakos, “Acute respiratory failure complicating advanced liver disease,” *Semin. Respir. Crit. Care Med.*, 2012.

[77] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence, “APACHE-acute physiology and chronic health evaluation: a physiologically based classification system.,” *Crit. Care Med.*, vol. 9, no. 8, pp. 591–7, Aug. 1981.

[78] J. R. Le Gall *et al.*, “A simplified acute physiology score for ICU patients.,” *Crit. Care Med.*, 1984.

[79] J.-L. Vincent *et al.*, “The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure,” *Intensive Care Med.*, 2002.

[80] M. Muhlbaier, A. Topalis, and R. Polikar, “Ensemble Confidence Estimates Posterior Probability,” 2005.

[81] L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin, “Decision templates for multiple classifier fusion: an experimental comparison,” *Pattern Recognit.*, 2001.

[82] H. Jung *et al.*, “Development of a Novel Markov Chain Model for the Prediction of Head and Neck Squamous Cell Carcinoma Dissemination.,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2016, pp. 1832–1839, 2016.

[83] H. Uğuz, A. Arslan, and İ. Türkoğlu, “A biomedical system based on hidden Markov

model for diagnosis of the heart valve diseases," *Pattern Recognit. Lett.*, vol. 28, no. 4, pp. 395–404, Mar. 2007.

[84] C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto, "Multistate Markov models for disease progression with classification error," *J. R. Stat. Soc. Ser. D Stat.*, 2003.

[85] R. V Andreao, B. Dorizzi, J. Boudy, and J. Mota, "ST-Segment Analysis Using Hidden Markov Model Beat Segmentation: Application to Ischemia Detection," 2004.

[86] R. Sukkar, E. Katz, Y. Zhang, D. Raunig, and B. T. Wyman, "Disease progression modeling using Hidden Markov Models," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2012.

[87] J. D. Freguson, "Variable duration models for speech," in *Proc. Symposium on the Application of Hidden Markov Models to Text and Speech*, pp. 143–179.

[88] G. D. Forney, "The viterbi algorithm," *Proc. IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[89] M. Abadi *et al.*, "TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, 2016, pp. 265–284.

[90] F. Husain-Syed *et al.*, "Cardio-pulmonary-renal interactions: A multidisciplinary approach," *Journal of the American College of Cardiology*. 2015.

[91] J. F. Dasta, T. P. McLaughlin, S. H. Mody, and C. T. Piech, "Daily cost of an intensive care unit day: the contribution of mechanical ventilation.," *Crit. Care Med.*, 2005.

[92] E. Bilevicius, D. Dragosavac, S. Dragosavac, S. Araújo, A. L. E. Falcão, and R. G. G. Terzi, "Multiple organ failure in septic patients," *Brazilian J. Infect. Dis.*, vol. 5, no. 3, pp. 103–110, Jun. 2001.

[93] E. C. Davies, C. F. Green, S. Taylor, P. R. Williamson, and D. R. Mottram, "Adverse Drug Reactions in Hospital In-Patients: A Prospective Analysis of 3695 Patient-Episodes," *PLoS One*, vol. 4, no. 2, p. 4439, 2009.

[94] C. D. Furberg, "Understanding drug safety and how to maximize it for patients.," *JAAPA*, vol. 24, no. 11, p. 16, Nov. 2011.

[95] S. J.W., "A pharmaceutical manufacturer's perspective on reporting adverse drug experiences," *Am. J. Hosp. Pharm.*, 1990.

[96] G. R. Venning, "Identification of adverse reactions to new drugs. II (continued): How were 18 important adverse reactions discovered and with what delays?," *Br. Med. J. (Clin. Res. Ed).*, vol. 286, no. 6362, pp. 365–8, Jan. 1983.

[97] I. J. Onakpoya, C. J. Heneghan, and J. K. Aronson, "Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature," 2016.

[98] E. H. Shortliffe and B. G. Buchanan, "A model of inexact reasoning in medicine," *Math. Biosci.*, 1975.

[99] J. Fox and R. Thomson, "Clinical decision support systems: a discussion of quality, safety and legal liability issues.," *Proceedings. AMIA Symp.*, 2002.

[100] D. Bates, "Clinical Decision Support and the Law - The big Picture," *Louis UJ Heal.*, 2011.

[101] M. Greenberg and M. S. Ridgely, “Clinical Decision Support and Malpractice Risk,” *JAMA*, vol. 306, no. 1, pp. 90–91, Jul. 2011.

[102] S. Leekha, C. L. Terrell, and R. S. Edson, “General principles of antimicrobial therapy,” in *Mayo Clinic Proceedings*, 2011.

[103] L. A. Mandell *et al.*, “Infectious Diseases Society of America/American Thoracic Society Consensus Guidelines on the Management of Community-Acquired Pneumonia in Adults,” *Clin. Infect. Dis.*, 2007.

[104] S. L. Goldstein, “Automated/integrated real-time clinical decision support in acute kidney injury.,” *Curr. Opin. Crit. Care*, vol. 21, no. 6, pp. 485–9, Dec. 2015.

[105] “Solutions For Patient Safety | Children’s Hospitals Working Together to Eliminate Harm.” [Online]. Available: https://www.solutionsforpatientsafety.org/. [Accessed: 02-Jul-2019].

[106] T. P. Gibson, “Renal Disease and Drug Metabolism: An Overview,” *Am. J. Kidney Dis.*, vol. 8, no. 1, pp. 7–17, Jul. 1986.

[107] G. S. Markowitz and M. A. Perazella, “Drug-induced renal failure: A focus on tubulointerstitial disease,” *Clinica Chimica Acta*. 2005.

[108] K. Chapin and P. Murray, “Principles of stains and media,” in *Manual of Clinical Microbiology*, 2007, pp. 258–261.

[109] J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed, “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9,” *Stat. Sci.*, 1990.

[110] P. Schulam and S. Saria, “Reliable Decision Support using Counterfactual Models,” Mar. 2017.

[111] R. Ambrosino, B. G. Buchanan, G. F. Cooper, and M. J. Fine, “The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies,” *Ninet. Annu. Symp. Comput. Appl. Med. Care. Towar. Cost-Effective Clin. Comput. Proc.*, 1995.

[112] P. M. Tulkens, “Nephrotoxicity of aminoglycoside antibiotics,” *Toxicol. Lett.*, 1989.

[113] M. J. Rybak and B. J. McGrath, “Combination Antimicrobial Therapy for Bacterial Infections,” *Drugs*, vol. 52, no. 3, pp. 390–405, Sep. 1996.

[114] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks arXiv:1311.2901v3 [cs.CV] 28 Nov 2013,” in *Proceedings of European conference on computer vision–ECCV 2014*, 2014.

[115] M. Joffe *et al.*, “Variability of creatinine measurements in clinical laboratories: results from the CRIC study.,” *Am. J. Nephrol.*, vol. 31, no. 5, pp. 426–34, 2010.

[116] L. Awdishu and R. L. Mehta, “The 6R’s of drug induced nephrotoxicity,” *BMC Nephrology*. 2017.

[117] E. Minejima, J. Choi, P. Beringer, M. Lou, E. Tse, and A. Wong-Beringer, “Applying New Diagnostic Criteria for Acute Kidney Injury To Facilitate Early Identification of Nephrotoxicity in Vancomycin-Treated Patients,” *Antimicrob. Agents Chemother.*, 2011.

[118] S. Rosen and I. E. Stillman, “Acute Tubular Necrosis Is a Syndrome of Physiologic and Pathologic Dissociation,” 2008.

[119] S. J. Berman, E. W. Johnson, C. Nakatsu, M. Alkan, R. Chen, and J. LeDuc, “Burden of Infection in Patients with End-Stage Renal Disease Requiring Long-Term Dialysis,” *Clin. Infect. Dis.*, 2004.

[120] M. Finland, “Changing ecology of bacterial infections as related to antibacterial therapy,” *J. Infect. Dis.*, 1970.

[121] D. G. Maki, “Nosocomial bacteremia. An epidemiologic overview,” *Am. J. Med.*, 1981.

[122] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017.

[123] J. Lutz, K. Jurk, and H. Schinzel, “Direct oral anticoagulants in patients with chronic kidney disease: Patient selection and special considerations,” *International Journal of Nephrology and Renovascular Disease*. 2017.

[124] G. J. Glas *et al.*, “Nebulized heparin for patients under mechanical ventilation: an individual patient data meta-analysis,” *Annals of Intensive Care*. 2016.

[125] A. Roch, C. Guervilly, and L. Papazian, “Fluid management in acute lung injury and ards,” *Ann. Intensive Care*, vol. 1, no. 1, p. 16, Dec. 2011.

[126] N. Stollman and D. C. Metz, “Pathophysiology and prophylaxis of stress ulcer in intensive care unit patients,” *Journal of Critical Care*. 2005.

[127] J. L. Vincent *et al.*, “The Prevalence of Nosocomial Infection in Intensive Care Units in Europe: Results of the European Prevalence of Infection in Intensive Care (EPIC) Study,” *JAMA J. Am. Med. Assoc.*, 1995.

[128] B. Settles, “Active Learning Literature Survey,” *Mach. Learn.*, 2010.

[129] V. Lo Re *et al.*, “Validity of diagnostic codes and laboratory tests of liver dysfunction to identify acute liver failure events,” *Pharmacoepidemiol. Drug Saf.*, 2015.