Department of Intelligence Science and
Technology
Graduate School of Informatics
KYOTO UNIVERSITY

Doctor Thesis

Machine Learning Approaches for Personalized Clinical Risk Modeling

Nozomi Nori
則 のぞみ

Supervisor: Professor Hisashi Kashima

February 2017

Abstract

Healthcare systems are no exception to the diverse fields affected by the era of big data. Clinicians are already compelled to make sense of the vast amounts of clinical data that constantly result from routine daily medical practice and patient monitoring; as often as not, this quantity of data exceeds short-term human cognitive processing capacities. While the increased availability of clinical data poses both opportunities and challenges, the ideal use of these diverse clinical data would allow for its effective applications to improvements in care, such as early detection of adverse clinical events, optimization of resource allocation and triage, cost reduction, and the creation of new medical knowledge.

Of these applications, clinical risk modeling, i.e., predictive modeling of relevant risks in medical treatment, has been a major subject of research. A typical example of the use of clinical risk modeling is severity assessment in acute hospital care. Accurate assessment of the severity of a patient's condition plays a fundamental role in acute hospital care, such as that provided in an intensive care unit (ICU), where clinicians intensively attend to critically ill patients. ICU clinicians are required to make sense of a large amount of clinical data in a limited time to estimate the severity of a patient's condition, which ultimately leads to the planning of appropriate care. To enhance severity assessments, patient mortality risk is often used as a surrogate to describe the patient's condition; there have been numerous studies for modeling the mortality risk of ICU patients. Overall, clinical risk modeling makes two major clinical contributions: improving the efficiency and quality of clinical care by, for example, appropriate preventive actions based on prediction and knowledge discovery through data analysis with learned models.

Despite the potentially significant implications of clinical risk modeling, personalized clinical risk modeling, that is, addressing the individuality of target patients while exploiting the common structure shared among the population to obtain patient-specific or patient-dependent risk models, has been a relatively understudied problem. The working hypothesis behind personalized modeling is that the target of interest might have some specific features in addition to the common structure shared by all the samples at hand; therefore, developing a customized model for the target that reflects its specificity

as well as a common structure might improve the predictive modeling.

In this study, we investigate clinical risk modeling for patients in ICUs, where clinicians attend to a heterogeneous group of patients with various disease types, all of whom are severely ill, and address the mortality risk prediction problem to enhance the severity assessments. Despite the diversity of ICU patients, risk modeling for ICU patients has typically been carried out by developing one common predictive model that is shared for all patients. We present several formulations for the personalized risk modeling of ICU patients to address the diversity of this population. First, we address patient specificity in terms of the disease the patient is associated with by simultaneously modeling multiple diseases. Specifically, we formulate risk prediction as a multitask learning problem, in which a task corresponds to a disease. To deal with data paucity resulting from disease-based customization and data sparseness associated with electronic medical records, we develop a method that integrates medical domain knowledge via graph Laplacians, which is introduced as an inductive bias in the learning process. Second, we address one critical issue in personalized clinical risk modeling: determining which unit should be used to model patient specificity, by learning the task unit in a multitask learning framework. Specifically, we assume a small number of latent basis tasks, where each latent task is associated with its own parameter vector; a parameter vector for a specific patient is constructed as a linear combination of these. The latent representation of a patient, namely, the coefficients of the combination, is learned based on the collection of diseases associated with the patient. Our method could be considered a multitask learning method in which latent tasks are learned based on the collection of diseases. Third, the data paucity problem is especially distinct for infrequent diseases, that is, diseases whose number of patients is relatively small in the entire population. To further address data paucity for such diseases, we explore the use of transfer learning for the risk modeling of infrequent diseases.

We demonstrate the effectiveness of our proposed methods using several real-world datasets collected from hospitals. Our method achieved higher predictive performance compared with a single-task learning method, the de facto standard, and several standard multitask learning methods. Furthermore, we expect that our proposed methods could be used not only for predictions, but also for understanding disease-specific contexts and patient-specificity from different viewpoints.

Acknowledgements

First, I would like to express my deep gratitude to my advisor, Professor Hisashi Kashima, who supported my doctoral research with great patience and tolerance. He had maximum respect for my independence in research, which further enhanced my independence throughout this project. He has also shown much consideration for me in various situations.

I would also like to thank all my collaborators in medicine, including Dr. Kazuto Yamashita, Professor Hiroshi Ikai, Professor Susumu Kunisawa, and Professor Yuchi Imanaka. Discussions with them have inspired me to elaborate my research topics.

Through discussions with Dr. Kazuto Yamashita, I deepened my understanding of the background of contemporary clinical research, which enabled me to conduct research on clinical risk modeling during my doctoral studies. He also contributed significantly to analyzing the learned models in our research.

Finally, special gratitude goes to my family, who have respected for my liberty throughout my life. While respecting for my independence, they gave me unstinting help whenever I got into difficulties.

Contents

Chapter 1 Introduction 1
1.1 The Data Revolution in Healthcare 1
1.2 Clinical Risk Modeling 1
1.3 Addressing Patient Characteristics in Clinical Risk Modeling 2
1.4 Personalized Clinical Risk Modeling for Acute Hospital Care 3
1.5 Solutions 5
1.5.1 Simultaneous Risk Modeling of Multiple Diseases with Medical Domain Knowledge (Chapter 3) 5
1.5.2 Learning Implicit Tasks via Mapping from Disease Space to Latent Space (Chapter 4) 5
1.5.3 Transfer Learning for Infrequent Diseases (Chapter 5) 5
1.6 Roadmap 6

Chapter 2 Related Work 7
2.1 Mortality Modeling for ICU Patients 7
2.2 Multitask and Transfer Learning 8
2.3 Multitask and Transfer Learning for Clinical Risk Modeling 9

Chapter 3 Simultaneous Risk Modeling of Multiple Diseases 11
3.1 Introduction 11
3.2 Simultaneous Modeling of Multiple Diseases 16
3.2.1 Problem Definition 16
3.2.2 Cross-regularized MTL 17
3.3 Empirical Study 19

3.3.1 Experimental Setup 20
3.3.2 Results and Discussion 24
3.4 Related Work 30
3.4.1 MTL for Clinical Problems 30
3.5 Summary 31

Chapter 4 Learning Implicit Tasks for Patient-Specific Risk Modeling 32
4.1 Introduction 32
4.2 Related Work 34
4.2.1 Patient-Specific Modeling 34
4.2.2 Mortality Modeling in the ICU 35
4.2.3 Multitask Learning 36
4.3 Learning Patient-Specific Risk Models 37
4.3.1 Problem Definition 37
4.3.2 Optimization Problem 39
4.4 Empirical Study 41
4.4.1 Experimental Setup 41
4.4.2 Results and Discussion 44
4.5 Summary 49

Chapter 5 Transfer Learning for Infrequent Diseases 50
5.1 Introduction 50
5.2 Related Work 51
5.2.1 Transfer Learning 51
5.2.2 Transfer Learning for Healthcare Problems . . . 52
5.3 Disease-specific Risk Modeling via Distribution Match . . 53
5.3.1 Problem Definition 53
5.3.2 Approach 54
5.4 Empirical Study 55
5.4.1 Experimental Setup 55
5.4.2 Results and Discussion 59
5.5 Summary 62

Chapter 6 Conclusion 66

Publications 68

References 69

List of Figures

3.1 A disease-patients distribution plot for an ICU dataset. The horizontal axis represents disease indices based on 3-digit ICD-10 codes. The vertical axis represents the number of corresponding patients. Most of the diseases only have a few patients. 13

3.2 An example of multitask learning formulation for multiple disease types. Patients are grouped into a task by their main disease as defined by the third level of the ICD-10 hierarchy. We define the similarities among diseases from this hierarchical structure to jointly learn prediction models for different diseases via multitask learning. 14

3.3 Comparison of AUCs for each disease among the Proposed, STL (common), and MTL (Mean) methods in the prediction setting of day 1. 29

4.1 Histogram of disease combinations in an ICU dataset. Many combinations are specific to a single patient. 33

4.2 Our MTL model for patients with multiple diseases. 35

4.3 Visualization of similarities among latent tasks via MDS using $\tilde{\boldsymbol{S}}$. Only one latent task ($k = 1$) was positioned as an outlier, while all others aligned with one dimension. 47

4.4 Histogram of the ratio of high-mortality diseases in the top-10 high-risk predictive features. For a latent task ($k = 1$), the highest-risk predictive features are composed of high-mortality diseases, whereas this tendency is not observed for other latent tasks. 48

4.5 Histogram of the ratio of low-mortality diseases in the top-10 low-risk predictive features. For a latent task ($k = 1$), the lowest-risk predictive features are composed of low-mortality diseases, whereas this tendency is not observed for other latent tasks. 49

5.1 Classification accuracy, i.e., the accuracy of the estimate of Equation (5.4), of each disease. We observe that for many diseases, classification accuracy is high compared to that of random prediction (accuracy of 0.5). 62

List of Tables

3.1 Notation and descriptions. 16

3.2 List of the diseases used in our study, along with the number of cases (patients) and the death rate for each disease. 22

3.3 Statistics of the dataset. 23

3.4 Comparison methods used in our experiment. 25

3.5 Comparison of averaged AUCs in each prediction setting. We performed a Wilcoxon signed-rank test for each pair of the method with the highest AUC and another method in each prediction setting. We confirmed that our proposed method statistically significantly ($p < 0.05$) outperformed all other methods in each setting, except for the results in bold font, if any. Regarding the boldface results, there was no statistically significant ($p < 0.05$) improvement using our proposed method over the comparison methods. There was no method that performed equally to or better than our proposed method throughout all the settings. 25

3.6 Comparison of averaged recalls in each prediction setting. We conducted a Wilcoxon signed-rank test for each pair of the method with the highest recall and another method in each prediction setting. We confirmed that our proposed method statistically significantly ($p < 0.05$) outperformed all other methods in each setting, except for the results in bold font. Regarding the boldface results, there was no statistically significant ($p < 0.05$) improvement by our proposed method over the comparison method, although there was no method that performed equally to or better than our proposed method throughout all the settings. 26

3.7 Comparison of averaged precisions in each prediction setting. We conducted a Wilcoxon signed-rank test for each pair of the method with the highest precision and another method in each prediction setting. We confirmed that our proposed method statistically significantly ($p < 0.05$) outperformed all other methods in each setting, except for the results shown in bold. Of the results in bold, there was no statistically significant ($p < 0.05$) improvement using our proposed method over the comparison method. Only *Proposed-dz* performed equally to or better than our proposed method throughout all the settings. 27

3.8 Comparison of averaged specificities in each prediction setting. We conducted a Wilcoxon signed-rank test for each pair of the method with the highest specificity and another method in each prediction setting. The results in bold font outperformed all nonboldface results. There was no method that performed equally to or better than MTL ($\ell_{2,1}$) and MTL (Trace) throughout all settings in terms of specificity. . 27

3.9 Effect of the similarity between diseases in the prediction setting of day 1. We conducted a Wilcoxon signed-rank test for the pair of our proposed methods that incorporate similarity with and without considering the root of the tree. The boldface results indicate better performance than the nonboldface results. We confirmed that considering the root of the tree resulted in statistically significant ($p < 0.05$) improvement over not considering it, except in specificity, for which there was no statistically significant difference between the two methods. 29

4.1 Notation and descriptions. 37

4.2 Statistics of the dataset. 42

4.3 Comparison of various methods used in our experiment. . . . 43

4.4 Comparison of averaged AUCs. We conducted a Wilcoxon signed rank-test for each pair of the method with the highest AUC and another method in each prediction setting. The results shown in bold indicate the method statistically significantly ($p < 0.05$) outperformed all the other methods in the relevant setting. For prediction at ICU admission, MTL (DM) method outperformed all other methods. For prediction at ICU discharge, our proposed method outperformed all other methods. 44

4.5 Comparison of averaged recalls. We conducted a Wilcoxon signed-rank test for each pair of the method with the highest recall and another method in each prediction setting. The result given in bold indicates the method statistically significantly ($p < 0.05$) outperformed all other methods in the concerned setting. In both settings, Proposed-w/o-A method outperformed all other methods. 44

4.6 Comparison of averaged precisions. We conducted a Wilcoxon signed-rank test for each pair of the method with the highest precision and another method in each prediction setting. The bold result, with the highest precision, indicates that the method statistically significantly ($p < 0.05$) outperformed all nonboldface methods in the concerned setting. For prediction at ICU admission, STL (common) and MTL (DM) outperformed all other methods. For prediction at ICU discharge, MTL (DM) outperformed all other methods. 45

4.7 Comparison of averaged specificities. We performed a Wilcoxon signed-rank test for each pair of the method with the highest specificity and another method in each prediction setting. The result in bold, with the highest specificity, indicates that the method statistically significantly ($p < 0.05$) outperformed all the other nonboldface methods in the concerned setting. For prediction at ICU admission, Proposed-w/o-pop and STL (common) outperformed all other methods. For prediction at ICU discharge, MTL (DM) method outperformed all other methods. 46

4.8 Example of top-10 predictive features for a disease-mortality-related latent task ($k = 1$). 46

5.1 Statistics of the dataset. 56
5.2 List of diseases, along with category, based on the patient population of the disease. 57
5.3 List of diseases used as target diseases, along with category, based on the patient population of the disease. 58
5.4 Comparison of recall for each pair of the transfer model and common, separate, and MTL-DM model, respectively, for each disease. For each pair of transfer and comparison model in each disease, statistically significant ($p < 0.05$) differences were confirmed via a Wilcoxon signed-rank test. For most diseases, recall significantly improved with the use of our transfer method. 63
5.5 Comparison of precision for each pair of the transfer model and common, separate, and MTL-DM model, respectively, for each disease. For each pair of transfer and comparison model in each disease, statistically significant ($p < 0.05$) differences were confirmed via a Wilcoxon signed-rank test. For most diseases, precision significantly improved with the use of our transfer method. 64
5.6 Comparison of specificity for each pair of the transfer model and common, separate, and MTL-DM model, respectively, for each disease. For each pair of the transfer and comparison model in each disease, statistically significant ($p < 0.05$) differences were confirmed via a Wilcoxon signed-rank test. For most diseases, comparison model showed significantly better specificity than the transfer model. 65

Chapter 1

Introduction

1.1 The Data Revolution in Healthcare

Healthcare systems have not been excluded from the wide range of domains affected by the revolutionary era of big data. Clinical staffs are already obliged to make sense of a large, ever-increasing amount of clinical data that is incessantly resulting from daily routine clinical practice and patient monitoring, the large amount and rapid timing of which typically exceeds human cognitive capacities. While the increased availability of clinical data poses both opportunities and challenges, the optimal use of this vast amount of diverse clinical data would enable beneficial applications for the improved care, such as early detection of adverse events, optimization of resource allocation and triage, cost reduction, and creation of new knowledge to deepen our understanding of medicine [1, 2, 3].

1.2 Clinical Risk Modeling

Among the promising healthcare applications, clinical risk modeling [4], i.e., predictive modeling of relevant risks in medical treatment, is one of the major research subjects. An accurate prediction of clinical risks, such as a patient mortality [5], hospital readmission [6, 7], and adverse reactions to medications [8], could lead to appropriate preventive actions by, for example, medical alerts. Besides, the accurate prediction of clinical risks could enhance the assessment of a patient's condition, which is a crucial step for clinicians to plan appropriate care and better manage patients. A representative exam-

ple is acute hospital care, as provided in an intensive care unit (ICU), where clinicians must intensively attend to critically ill patients. An accurate assessment of the severity of a patient's condition plays a fundamental role in acute hospital care; to improve clinicians' severity assessments, patient mortality risk is often used as a surrogate to describe the severity of a patient's condition [6, 9, 10, 11, 12]. Clinical risk modeling has also been utilized for performance evaluation for institutions or surgeries by comparing an observed patient outcome with predicted risk [13, 14, 15]. Furthermore, accurate clinical risk modeling could lead to the discovery of medical knowledge through learning from data.

1.3 Addressing Patient Characteristics in Clinical Risk Modeling

A crucial point in conventional clinical risk modeling, although it is not often noted explicitly but rather placed as an implicit assumption, is to capture patient characteristics adequately so that an appropriate set of patients is identified for which an effective common model can be developed. For example, since the patient population differs by institution, general models developed by using data from multiple institutions or a different institution might tend to perform relatively poorly for a specific institution [16]. The adequacy of characterization depends on the objective of clinical risk modeling, and how such patient characteristics should be addressed has been noted for individual problems in the medical literature. For the assessment or comparison of performance for cardiac surgeries performed by institutions, where explanatory variables are mainly composed of the procedures performed, preoperative patient conditions should be comparable with those of the original environments in which the model was developed [17]. For instance, an institution might only admit relatively young, healthy, elective patients, whereas patients in another institution might be mostly composed of elderly, non-elective, clinically ill patients; in such a case, even if same the surgeries are performed, the surgery patients' outcomes would be highly varied depending on the institution, making a model developed from the former institution ineffective for use in the latter. Several studies in the medical domain have reported that considering patient characteristics in clinical risk modeling, by, for example,

accounting for hospital-specific characteristics, can improve prediction performance over those that do not [16, 17].

Such patient characteristics are usually addressed based on the knowledge of the analysts conducting clinical risk modeling; analysts typically identify and preselect relevant samples from the candidate pools in advance, then conduct careful feature engineering so that patient characteristics are reflected by, for example, introducing specific features, such as institution-specific features. These processes are basically hand-operated and heavily depend on the analysts. On the other hand, the mechanization of the process for capturing patient characteristics and exploiting the inherent relationships among different types of patient groups, is in its infancy in clinical risk modeling.

In fact, addressing both the *specificity* and the *commonality* among a patient population — that is, capturing the specific characteristics of the target patients while exploiting the common structure that is shared among the population — has been one of the most fundamental issues in recent clinical research. While obtaining clinically useful models compels the model to be customized to the target of the analysis, limiting the data to those that are exactly relevant to the target often results in an impractically small sample size. We define clinical risk modeling that addresses patient characteristics while exploiting common structure shared among a population as personalized clinical risk modeling, which is investigated in this study.

1.4 Personalized Clinical Risk Modeling for Acute Hospital Care

In this study, we focus on clinical risk modeling for acute hospital care, such as that provided in the ICU, and present several formulations for personalized modeling. In acute hospital care, patient mortality risk is often used as a surrogate to describe the severity of the condition of the patient. Accurate prediction of the mortality risks would enable appropriate preventive actions such as medical alarms; in addition, it would enable optimization of resource allocation, thereby improving the quality and efficiency of the care delivery.

One of the most salient features of the ICU is the diversity of the patients: ICU clinicians are faced with diverse patients with a wide variety of diseases, in contrast with situations in which clinicians face patients with similar dis-

eases in a specific treatment department. For instance, some patients are admitted to ICU due to infectious diseases such as sepsis and pneumonia, whereas others are postoperative patients admitted after some major surgery. Furthermore, each patient is typically associated with multiple different types of diseases, making the clinical states of ICU patients rather complicated.

Nevertheless, risk modeling for ICU patients has been typically made by developing one common predictive model that is shared with all the patients. Consequently, we must address the following characteristics of ICU patients: (1) disease-specific characteristics, and (2) patient-specific characteristics that are captured in the collection of the diseases associated with each patient.

The significance of addressing the first characteristics, i.e., disease-specific characteristics, is illustrated as follows. Each disease might have a specific prediction rule that accounts for the patient's risk; therefore, developing disease-dependent risk model would enhance the predictive modeling. For example, a kind of stomach medicine could be administered to patients who have received artificial respiration for the prevention of gastric ulcer because artificial respiration often causes gastric ulcer; on the other hand, the same stomach medicine could be administered to remedy severe gastric ulcer with bleeding. The corresponding prediction rule would be different depending on which type of disease a patient has. As a result, building a specific model to each disease would improve the predictive modeling for ICU patients. Similarly, the significance of addressing the second characteristics, i.e., the collection of diseases associated with each patient, is supported by the fact that each patient is typically associated with multiple different types of diseases, and the clinical state of the patient is different depending on the combination of the diseases.

However, we must deal with several issues when addressing the above-mentioned characteristics. First, as is often the case with clinical data, most of the diseases only have a small number of patients in acute hospital care. Therefore, information about a specific disease is fairly limited, making it difficult to develop a customized model reflecting the characteristics of a specific disease. Second, although some patient characteristics might be captured by the collection of the diseases associated with each patient, how exactly such patient characteristics should be treated and obtained from the various collections of diseases is not obvious. One simple way would be to create a model for each combination of diseases associated with a patient; however, such a

naïve approach is complicated by the combinational explosion of diseases.

1.5 Solutions

We present methods for personalized clinical risk modeling in acute hospital care to address each of the aforementioned issues.

1.5.1 Simultaneous Risk Modeling of Multiple Diseases with Medical Domain Knowledge (Chapter 3)

To cope with the data paucity of each disease in considering disease-specific characteristics, we present a multitask learning method in which a task corresponds to a disease; data paucity is mitigated by introducing inductive biases using graph Laplacians encoding medical domain knowledge related to the similarities among diseases and among electronic medical records.

1.5.2 Learning Implicit Tasks via Mapping from Disease Space to Latent Space (Chapter 4)

To address the second issue, that is, to account for the patient characteristics that reflect the collection of diseases each patient is associated with, we present a multitask learning method in which latents tasks are learned based on a collection of diseases. Specifically, we assume a small number of latent basis tasks, where each latent task is associated with its own parameter vector; a parameter vector for a specific patient is constructed as a linear combination of these. The latent representation of a patient (the coefficients of the combination), is learned based on the collection of diseases associated with the patient. Consequently, our method produces patient-specific risk models that reflect the collection of diseases associated with each patient.

1.5.3 Transfer Learning for Infrequent Diseases (Chapter 5)

Data paucity is especially noticeable for infrequent diseases, whose number of patients is relatively small compared to the entire population. For infre-

quent diseases, both training data and medical domain knowledge are quite limited. The above-mentioned proposed method, which exploits medical domain knowledge, would be effective in some situations; however, there should also be some inherent relationship among diseases that is learned from data and could be exploited for the risk modeling of infrequent diseases. To further address the data paucity problem for infrequent diseases, we explore the use of transfer learning for infrequent diseases risk modeling.

1.6 Roadmap

This thesis is organized into a related work section (Chapter 2), three main research results (Chapters 3, 4, and 5), and conclusions (Chapter 6). We first present several related works to provide background for the studies described in this thesis in Chapter 2. Chapter 3 addresses the issue of data paucity in considering disease-specific characteristics for clinical risk modeling in acute hospital care, which was published in the Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [18], and IEICE TRANSACTIONS on Information and Systems [19][*1]. Chapter 4 addresses the issue of learning implicit tasks for patient-specific risk modeling in acute hospital care, which was published in the Proceedings of the 31st AAAI Conference on Artificial Intelligence [20][*2]. Chapter 5 presents a feasibility study of transfer learning for risk modeling of infrequent diseases.

[*1] Chapter 3 is based on these papers [18, 19], by the same authors, which appeared in the Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, and IEICE TRANSACTIONS on Information and Systems, Copyright (C) 2017 ACM, IEICE.

[*2] Chapter 4 is based on the paper [20], by the same authors, which appeared in the Proceedings of the 31st AAAI Conference on Artificial Intelligence, Copyright (C) 2017 AAAI.

Chapter 2

Related Work

2.1 Mortality Modeling for ICU Patients

Traditionally, mortality modeling for the ICU patients has been conducted via scoring systems, such as SAPS (simplified acute physiology score) and APACHE (acute physiology and chronic health evaluation), both of which use fixed clinical decision rules based mainly on physiological data [5]. However, it should be noted that these ICU scoring systems are only used in rather limited situations. Specifically, a study reported that they were used for 10–15% of ICU patients in the US as of 2012 [13]. With the increased availability of varied data from hospital electronic medical records (EMRs), the feasibility of data-driven mortality predictive modeling based on EMRs has been explored extensively in the clinical domains [9, 21, 22, 23]. These studies demonstrate that EMRs can be used to generate clinically plausible mortality prediction models with superior discrimination.

Many of these studies can be viewed as feature engineering. Hug and Szolovits [21] conducted exhaustive feature engineering to explore the feasibility of real-time mortality risk prediction by using various nurse-charted observations, such as vital signs, lab results, and medications, as well as several carefully designed features suggested by the medical literature as being useful for prediction. The use of unstructured information (e.g., clinical notes) has been explored in recent years; for example, Lehman [23] et al. applied hierarchical Dirichlet processes to unstructured clinical notes to develop risk stratifications for ICU patients. Ghassemi et al. [10] examined the use of latent variable models to obtain meaningful features from unstructured clinical

notes. Both papers reported that the features extracted from clinical notes were predictive yet interpretable. Another direction that has recently been explored is time-series modeling; Ghassemi et al. [11] examined the use of time-series modeling via a multitask Gaussian process to forecast the severity of illness in ICU patients.

One aspect of the ICU, which is fundamental yet left largely uninvestigated in mortality modeling, is the diversity of patients: ICUs are composed of patients with a wide variety of disease types. In fact, Sionitis et al. [5] pointed out in their review of 94 studies related to ICU mortality modeling that there was a large variability of prediction accuracy across various diseases and population subgroups. This suggests that a promising direction of ICU mortality modeling would be to address the diversity of patients. In this thesis, we provide several methods to address patient diversity in mortality modeling for ICU patients.

2.2 Multitask and Transfer Learning

Multitask learning is a learning paradigm whose primary goal it to improve generalization performance by leveraging information from *related tasks* as an inductive bias in the learning process [24]. In multitask learning, multiple related tasks are learned simultaneously by sharing information across different tasks. To date, there have been numerous studies on multitask learning [24, 25, 26, 27, 28, 29, 30, 31, 32, 33].

Transfer learning is a closely related concept; in a sense, transfer learning could be considered a broader concept than multitask learning. In traditional machine learning methods, the training and future data are assumed to lie in the same feature space and have the same distribution. In the transfer learning framework, these assumptions are violated; the data for the task of interest and data from other related tasks come from different distributions. There are two main concepts in the transfer learning framework: *domain* and *task*. A *domain* comprises two parts: a feature space X and a marginal probability distribution $P(X)$. Given a specific domain, a *task* consists of two components: a label space Y and an objective predictive function $f(\cdot)$. Based on these definitions, transfer learning can be categorized into three types [34]: inductive, transductive, and unsupervised transfer learning. Multitask learning is closely related to inductive transfer learning, in which the source and

target tasks are different and the source and target domains may be either the same or different.

The main difference between multitask and transfer learning is that multitask learning aims to improve the overall performance in all tasks, whereas the aim of transfer learning is to improve performance for only the target task.

In this thesis, we first present a multitask learning method that considers disease-specific characteristics among ICU patients in Chapter 3. To further address patient-specific characteristics, in Chapter 4, we present a multitask learning method in which the unit of the tasks itself is learned based on the collection of diseases the patients are associated with, by introducing latent basis tasks. Then, to address problem of data paucity for infrequent diseases, we study the use of inductive transfer learning for ICU patient risk modeling in Chapter 5.

2.3 Multitask and Transfer Learning for Clinical Risk Modeling

One of the most fundamental matters in recent clinical risk modeling has been to address the specificity of the target patients while capturing the common structure shared by the population. Obtaining clinically meaningful models requires the model to be personalized to the target of the analysis; however, restricting the data to those that are faithfully relevant to the target often results in an impractically small sample size. Therefore, to build clinically practical target-specific models, recent studies have focused on multitask and transfer learning for clinical data.

One important issue is to determine the aspects we should use to capture patient specificity, that is, how to define the tasks. To date, hospital-specific [35, 36, 37], surgery-specific [37], and intervention-specific [38] clinical risk models have been proposed via multitask or transfer learning. For hospital-specific modeling, Gong et al. [37] proposed an instance-transfer method that weights examples from the source domain based on their similarity to the training examples in the target domain. Jenna et al. [36] explored the use of hospital-specific risk modeling by using a feature-representation transfer method that exploits auxiliary data from other hospitals. Lee et al. [35] investigated the effectiveness of transfer learning for adapting a global model

that is shared across multiple hospitals to a specific hospital; they first trained a model using data from source hospitals and then learned a model for the target hospital by regularizing the model parameters toward that of the source data. The method by Gong et al. [37] was also applied to develop surgery-specific models. For intervention-specific modeling, Gupta et al. [38] exploited hierarchical Dirichlet process to develop a specific model for each group of interventions. These studies showed promising results, indicating that some improvement in predictive performance is possible with proposed personalized risk modeling; in addition, several studies have shown that developing personalized models via multitask or transfer learning enabled them to obtain some clinical insights by analyzing the obtained models.

Yet, there have been few studies that adopt disease or the combination of diseases as the unit of the task. One exception is a work by Wang et al. [39], that defined onset risk prediction for a disease as a task and formulated the onset risk prediction for multiple diseases as a multitask learning.

In this thesis, we focus on personalized clinical risk modeling for acute hospital care, specifically in the ICU. Because one of the most salient features of the ICU is the diversity of its patient population, the ICU is an especially suitable setting for patient-specific or personalized clinical risk modeling. Potentially, there should be several promising settings other than ICU. For instance, clinicians in radiology departments are involved in various types of diagnostic imaging collected from all diagnosis and treatment departments. While there should be some specific characteristics for each department, there would also be some common structure shared among all images handled in the department of radiology, making it another potentially good setting for patient-specific or personalized risk modeling. More broadly, our work demonstrates the potential benefit of personalized clinical risk modeling by adopting the ICU as a representative suitable setting.

Chapter 3

Simultaneous Risk Modeling of Multiple Diseases

3.1 Introduction

Accurate assessment of the severity of a patient's condition is a crucial step for clinicians planning appropriate medical care and managing patient health. In the ICU, mortality risk of a patient is often used as a proxy to describe the patient's condition. ICU clinicians attend to several patients at a time and have limited time in which to make clinical decisions; therefore, the ability to accurately forecast mortality risk stands to significantly improve the quality of hospital care, which could contribute to reducing the number of *preventable deaths*. In fact, some studies have reported that a certain number of deaths of patients who were admitted to the ICU were considered potentially avoidable [40, 41, 42], suggesting the significance of early and precise recognition of patient mortality risk in planning appropriate treatment. In this context, many studies have focused on predictive modeling of mortality risk for ICU patients [10, 11, 21, 22, 43, 44].

One salient feature that makes the ICU a particularly challenging environment is the *diversity of patients*: In contrast with hospital departments in which clinical staff treat patients with specific disease types, ICU staff are faced with a heterogeneous group of patients with a wide variety of diseases.

In this setting, however, mortality risk prediction for ICU patients is typically performed by developing one common predictive model that is shared for all diseases. Yet, each disease might have a specific prediction rule that accounts for mortality risk. For example, a kind of stomach medicine could be administered to patients who have received artificial respiration for the prevention of gastric ulcer, because artificial respiration often causes gastric ulcer. However, the same stomach medicine could also be administered to remedy severe gastric ulcer with bleeding. The corresponding prediction rule would be different depending on which type of disease a patient has.

Disease-specific context is a tacit assumption in other prediction tasks, such as hospital re-admission [7, 38] and disease progression predictions [45, 46, 47, 48], in which a model specialized to a target disease is developed with the assumption that the target disease is specified beforehand. In these cases, it is assumed that the dataset consists of only patients with the specified target disease; essentially, learning models for multiple different diseases simultaneously is not assumed to be necessary. However, ICU patients consist of a heterogeneous group with a wide variety of diseases, which calls for mortality modeling that considers disease-specific contexts. Consequently, by customizing the predictive model for each disease, mortality risk modeling would improve, thereby enhancing the quality and efficiency of ICU care.

Challenges: data scarcity and data sparsity

Despite the importance of mortality risk prediction for ICU patients, significant challenges to disease-based customization are posed by: (1) *data scarcity* resulting from disease-based customization and (2) *data sparsity* associated with EMRs.

(1) Data scarcity resulting from the disease-based customization.

Attempts to build a customized model for each disease are complicated by the limited availability of sufficiently large datasets, because most of the diseases are associated with a small number of patients. To illustrate this, in Figure 3.1 we show a disease-patient distribution plot from an ICU dataset, for which data consisting of about 330,000 patients who underwent ICU treatment were collected from about 200 hospitals in Japan. The horizontal axis represents the disease indices, which are based on the 3-digit International Statistical Classification of Diseases and Related Health Problems 10 (ICD-10) codes. The vertical axis represents the corresponding number of patients. The plot shows that most of the diseases have only a few patients. Hence, naïve cus-

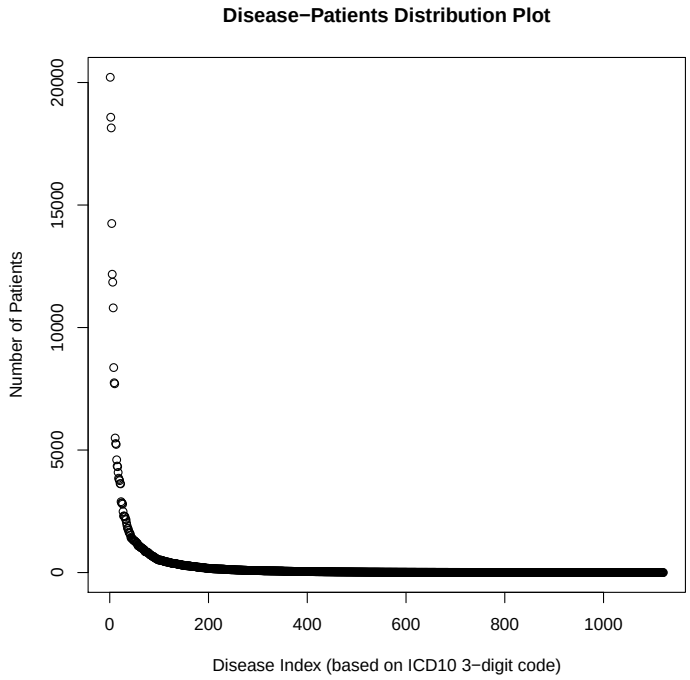

Fig. 3.1: A disease-patients distribution plot for an ICU dataset. The horizontal axis represents disease indices based on 3-digit ICD-10 codes. The vertical axis represents the number of corresponding patients. Most of the diseases only have a few patients.

tomization by disease fails due to the scarcity of data.

(2) Data sparseness associated with EMRs.

Another kind of sparsity is associated with EMRs. In mortality modeling, each patient is typically represented by an EMR that describes the patient's demographic information, clinical history, medications, and lab tests. However, raw EMRs are extremely sparse [49], a condition that also applies to our ICU cases. One reason behind this sparsity is the fact that a significant number of EMRs are subject to medical and clinical classification, which categorizes medical and clinical information from multiple viewpoints, producing highly fine-grained features. One example is the categorization of medications, which could be done in terms of drug efficacy, ingredients, form, pharmaceutical manufacturer, etc.; if two medicines are different from one point of view, such as their ingredients, then the raw codes of the two medicines will be different. However, the fact that the two medicines share all their other

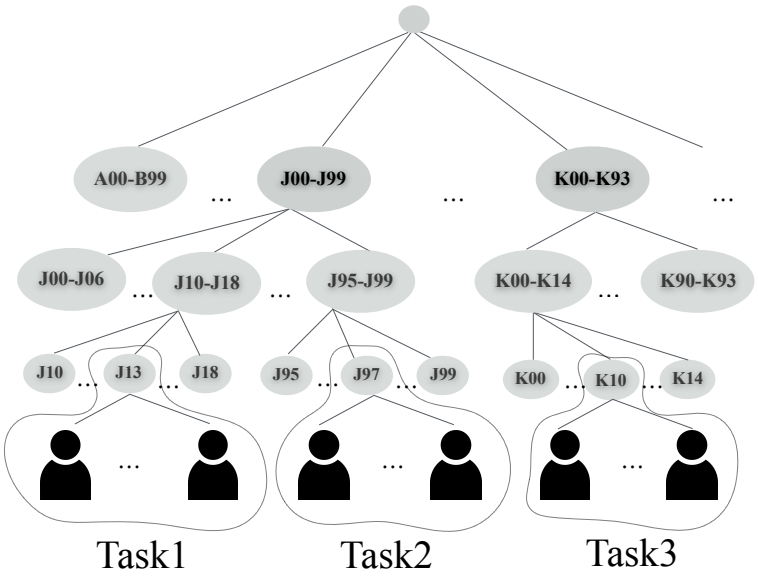

Fig. 3.2: An example of multitask learning formulation for multiple disease types. Patients are grouped into a task by their main disease as defined by the third level of the ICD-10 hierarchy. We define the similarities among diseases from this hierarchical structure to jointly learn prediction models for different diseases via multitask learning.

properties, including drug efficacy, suggests that they may have similar roles in predicting patient risk. Consequently, it would be desirable to incorporate such similarities in EMRs.

Our solution: multitask learning with medical domain knowledge

To address the challenges we have discussed here, we propose a method that effectively integrates medical domain knowledge into a data-driven approach using multitask learning (MTL) [24, 50]. In our formulation, a task corresponds to a disease and prediction tasks for different types of disease are jointly solved by sharing information across diseases. The key issue in formulating mortality prediction as an MTL problem is relating different kinds of diseases to share information among them. To appropriately model the relationships between diseases, we exploit domain knowledge of the medical classification of diseases using the ICD hierarchy. The ICD is a widely used classification of diseases that is maintained by the World Health Organization (WHO). In this hierarchical classification, diseases are categorized by cause, symptoms, morphological disparity, etc., encoding important information that

might affect mortality risk. Figure 3.2 illustrates an MTL formulation based on the ICD-10 hierarchy. In this formulation, patients with the same kind of disease, as defined by the third level of the ICD-10 hierarchy, are grouped together into a task. Then, information about the extent to which diseases are related can be obtained by exploiting the ICD-10 hierarchical structure. We integrate this medical domain knowledge by incorporating a graph Laplacian that encodes the similarities between diseases into the regularization term in the multitask formulation. Additionally, the data sparsity associated with EMRs is mitigated by integrating domain knowledge about the classification of clinical features, which is realized as another graph Laplacian, into the regularization term.

The contributions of our study in this chapter are twofold:

- We formulate in-hospital mortality risk prediction as an MTL problem, where a task corresponds to a disease, thereby incorporating the disease-specific context into mortality modeling. To integrate medical domain knowledge, we introduce a cross-regularization into the MTL formulation; medical domain knowledge is integrated by two graph Laplacians that encode similarities between diseases and among EMRs, respectively. In our cross-regularization framework, the model parameters of diseases and features are both regularized simultaneously.
- We empirically evaluate our proposed method by employing it in an in-hospital mortality prediction problem using real ICU dataset collected from a hospital in Japan. Experimental results show that the proposed method outperforms several baseline methods including logistic regression without MTL and several MTL methods that do not incorporate domain knowledge.

Table 3.1: Notation and descriptions.

Notation	Description
T	Number of diseases
N_t	Number of patients with t-th disease, where $t \in \{1, \ldots, T\}$
M	Number of total features derived from EMRs
$\boldsymbol{\phi}_n^{(t)}$	M-dimensional feature vector for n-th patient with t-th disease, where $n \in \{1, \ldots, N_t\}$ and $t \in \{1, \ldots, T\}$
$\boldsymbol{\Phi}^{(t)}$	$N_t \times M$ design matrix: $\boldsymbol{\Phi}^{(t)} \equiv [\boldsymbol{\phi}_1^{(t)}, \ldots, \boldsymbol{\phi}_{N_t}^{(t)}]^\top$
$y_{t,n}$	Class label for the n-th patient with disease t: $y_{t,n} \in \{0, 1\}$, where 0 and 1 represent survival and death in hospital, respectively.
$\boldsymbol{y}^{(t)}$	N_t-dimensional response vector: $\boldsymbol{y}^{(t)} \equiv (y_{t,1}, y_{t,2}, \ldots, y_{t,N_t})^\top$
$\boldsymbol{w}^{(t)}$	M-dimensional parameter vector for t-th disease, where $t \in \{1, \ldots, T\}$
$\boldsymbol{W}$	$M \times T$ parameter matrix: $\boldsymbol{W} \equiv [\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(T)}]$
$\boldsymbol{S}^{\text{dz}}$	$T \times T$ similarity matrix, where $\boldsymbol{S}^{\text{dz}}(i, j)$ is similarity between i-th and j-th diseases.
$\boldsymbol{S}^{\text{feat}}$	$M \times M$ similarity matrix, where $\boldsymbol{S}^{\text{feat}}(i, j)$ is similarity between i-th and j-th features.
$\boldsymbol{\mathcal{L}}^{\text{dz}}$	$T \times T$ normalized Laplacian matrix of $\boldsymbol{S}^{\text{dz}}$
$\boldsymbol{\mathcal{L}}^{\text{feat}}$	$M \times M$ normalized Laplacian matrix of $\boldsymbol{S}^{\text{feat}}$

3.2 Simultaneous Modeling of Multiple Diseases

3.2.1 Problem Definition

Let us formally define our problem: disease-dependent mortality risk modeling in which each disease has a separate, specialized model to incorporate disease-specific contexts.

Let T denote the number of diseases. A mortality risk prediction model is constructed for each disease. The t-th disease has N_t patients and the n-th patient with the t-th disease is associated with an M-dimensional feature vector $\boldsymbol{\phi}_n^{(t)}$ derived from EMRs. We create an $N_t \times M$ design matrix for

the t-th disease: $\mathbf{\Phi}^{(t)} \equiv [\boldsymbol{\phi}_1^{(t)}, \boldsymbol{\phi}_2^{(t)}, \ldots, \boldsymbol{\phi}_{N_t}^{(t)}]^\top$. Each patient is associated with a binary class label. Let $y_{t,n} \in \{0, 1\}$ denote the class label for the n-th patient with the t-th disease where $y_{t,n} = 1$ when the patient outcome is death and $y_{t,n} = 0$ otherwise. We create an N_t-dimensional response vector for the t-th disease: $\boldsymbol{y}^{(t)} \equiv (y_{t,1}, y_{t,2}, \ldots, y_{t,N_t})^\top$. Because we are interested in the probability of mortality, we adopt logistic regression and represent the posterior probability of the outcome of patient n being death as $\Pr[y_{t,n} = 1|\boldsymbol{\phi}_n^{(t)}] = \sigma(\boldsymbol{w}^{(t)^T}\boldsymbol{\phi}_n^{(t)})$, where $\sigma(a)$ is the sigmoid function: $\sigma(a) \equiv (1 + \exp(-a))^{-1}$ and $\boldsymbol{w}^{(t)}$ is an M-dimensional model parameter vector for the t-th disease.

To address the scarcity of data resulting from disease-based customization, we learn models for all the diseases jointly, sharing information across diseases. We represent the whole model parameter as an $M \times T$ matrix $\boldsymbol{W} \equiv [\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(T)}]$ and estimate parameter matrix $\boldsymbol{W}$ by minimizing an objective function that includes the log loss and some regularization terms. The regularizations are incorporated to encourage the models to respect domain knowledge, such as the relationships between diseases.

We assume that domain knowledge about similarities among diseases and among features is available, and is encoded as symmetric similarity matrices $\boldsymbol{S}^{\text{dz}}$ and $\boldsymbol{S}^{\text{feat}}$, respectively. $\boldsymbol{S}^{\text{dz}}(i, j)$ is the similarity between the i-th and j-th diseases, and $\boldsymbol{S}^{\text{feat}}(i, j)$ is the similarity between the i-th and j-th features. Note that the similarity matrix can be any symmetric matrix whose elements are positive real numbers. Notation and their descriptions are summarized in Table 3.1.

Our goal is to estimate the probability of mortality risk of a new patient represented as an M-dimensional feature vector $\boldsymbol{\phi}_{n'}^{(t')}$ ($t' \in \{1, \ldots, T\}$), given some observed pairs of $\{(\mathbf{\Phi}^{(t)}, \boldsymbol{y}^{(t)})\}_{t=1,\ldots,T}$ and two similarity matrices $\boldsymbol{S}^{\text{dz}}$ and $\boldsymbol{S}^{\text{feat}}$.

3.2.2 Cross-regularized MTL

We define the loss function as the log loss denoted by $\mathcal{L}(\boldsymbol{W})$:

$$\mathcal{L}(\boldsymbol{W}) \equiv -\sum_{t}^{T}\sum_{n}^{N_t}\{y_{t,n} \log \sigma(\boldsymbol{w}^{(t)^\top}\boldsymbol{\phi}_n^{(t)}) \tag{3.1}$$

$$+ (1 - y_{t,n}) \log(1 - \sigma({\boldsymbol{w}^{(t)}}^\top \boldsymbol{\phi}_n^{(t)}))\},$$

where $\sigma(a)$ is the sigmoid function. In addition to the loss function, we include the regularization term Ω in our objective function to incorporate domain knowledge. The optimization problem is defined as follows:

$$\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) + \Omega(\boldsymbol{W}), \tag{3.2}$$

where Ω is a regularization term that is used to avoid overfitting and to incorporate domain knowledge.

Here, we exploit our domain knowledge about the similarities among diseases and features with the following regularization terms:

$$\begin{aligned}\Omega^{\text{dz}}(\boldsymbol{W}) &\equiv \frac{1}{4}\sum_{i=1}^{T}\sum_{j=1}^{T} \boldsymbol{S}^{\text{dz}}_{i,j} \left|\left| \frac{\boldsymbol{W}_{*,i}}{\sqrt{\boldsymbol{D}^{\text{dz}}_{i,i}}} - \frac{\boldsymbol{W}_{*,j}}{\sqrt{\boldsymbol{D}^{\text{dz}}_{j,j}}} \right|\right|^2 \\ &= \frac{1}{2}\text{Tr}(\boldsymbol{W}\boldsymbol{\mathcal{L}}^{\text{dz}}\boldsymbol{W}^\top),\end{aligned} \tag{3.3}$$

$$\begin{aligned}\Omega^{\text{feat}}(\boldsymbol{W}) &\equiv \frac{1}{4}\sum_{i=1}^{M}\sum_{j=1}^{M} \boldsymbol{S}^{\text{feat}}_{i,j} \left|\left| \frac{\boldsymbol{W}_{i,*}}{\sqrt{\boldsymbol{D}^{\text{feat}}_{i,i}}} - \frac{\boldsymbol{W}_{j,*}}{\sqrt{\boldsymbol{D}^{\text{feat}}_{j,j}}} \right|\right|^2 \\ &= \frac{1}{2}\text{Tr}(\boldsymbol{W}^\top\boldsymbol{\mathcal{L}}^{\text{feat}}\boldsymbol{W}),\end{aligned} \tag{3.4}$$

where $\boldsymbol{D}^{\text{dz}}$ is a diagonal matrix, of which the (i,i)-th element is defined as $\boldsymbol{D}^{\text{dz}}_{i,i} \equiv \sum_j \boldsymbol{S}^{\text{dz}}_{i,j}$ ($\boldsymbol{D}^{\text{feat}}$ is defined similarly), $||\boldsymbol{x}||$ is the Euclidean norm of vector $\boldsymbol{x}$, and $\boldsymbol{\mathcal{L}}^{\text{dz}}$ and $\boldsymbol{\mathcal{L}}^{\text{feat}}$ are the symmetric normalized Laplacians of $\boldsymbol{S}^{\text{dz}}$ and $\boldsymbol{S}^{\text{feat}}$, respectively. The symmetric normalized Laplacian of an undirected graph with adjacency matrix $\boldsymbol{A}$ is defined as $\boldsymbol{D}^{-1/2}(\boldsymbol{D} - \boldsymbol{A})\boldsymbol{D}^{-1/2}$, where $\boldsymbol{D}$ is a diagonal matrix of which the (i,i)-th element is defined as $\boldsymbol{D}_{i,i} \equiv \sum_j \boldsymbol{A}_{i,j}$.

The regularization term $\Omega^{\text{dz}}(\boldsymbol{W})$ makes the two model parameters $\boldsymbol{W}_{*,i}$ and $\boldsymbol{W}_{*,j}$ similar if diseases i and j are similar in terms of the medical classifications given as the similarity matrix $\boldsymbol{S}^{\text{dz}}$. Similarly, the regularization

term $\Omega^{\text{feat}}(\boldsymbol{W})$ makes the two model parameters $\boldsymbol{W}_{i,*}$ and $\boldsymbol{W}_{j,*}$ similar if two features i and j are similar in terms of the clinical classifications given in similarity matrix $\boldsymbol{S}^{\text{feat}}$. Both the rows and columns of $\boldsymbol{W}$ are regularized using the two Laplacians $\boldsymbol{\mathcal{L}}^{\text{dz}}$ and $\boldsymbol{\mathcal{L}}^{\text{feat}}$, which we call cross-regularization.

Overall, we adopt the following regularization term:

$$\Omega(\boldsymbol{W}) \equiv \lambda^{\text{dz}}\Omega^{\text{dz}}(\boldsymbol{W}) + \lambda^{\text{feat}}\Omega^{\text{feat}}(\boldsymbol{W}) + \lambda^{\text{rid}}\Omega^{\text{rid}}(\boldsymbol{W}), \tag{3.5}$$

where $\Omega^{\text{rid}}(\boldsymbol{W}) \equiv \frac{1}{2}\text{Tr}(\boldsymbol{W}\boldsymbol{W}^\top)$ is the standard ridge regularization term used to avoid overfitting and $\lambda^{\text{dz}} \geq 0, \lambda^{\text{feat}} \geq 0$ and $\lambda^{\text{rid}} \geq 0$ are hyperparameters used to tune the weight of the regularization terms $\Omega^{\text{dz}}, \Omega^{\text{feat}}$, and Ω^{rid} , respectively.

Since Laplacian matrices are positive-semidefinite, the regularization term is convex as well as the loss function; hence, the optimal solution is found using standard gradient-based methods. We applied the L-BFGS optimizer [51] with the following derivatives:

$$\left[\frac{\partial \mathcal{L}(\boldsymbol{W})}{\partial \boldsymbol{W}}\right]_{*,t} = {\boldsymbol{\Phi}^{(t)}}^\top(\boldsymbol{p}^{(t)} - \boldsymbol{y}^{(t)}), \tag{3.6}$$

$$\frac{\partial \Omega^{\text{dz}}(\boldsymbol{W})}{\partial \boldsymbol{W}} = \boldsymbol{W}\boldsymbol{\mathcal{L}}^{\text{dz}}, \tag{3.7}$$

$$\frac{\partial \Omega^{\text{feat}}(\boldsymbol{W})}{\partial \boldsymbol{W}} = \boldsymbol{\mathcal{L}}^{\text{feat}}\boldsymbol{W}, \tag{3.8}$$

$$\frac{\partial \Omega^{\text{rid}}(\boldsymbol{W})}{\partial \boldsymbol{W}} = \boldsymbol{W}, \tag{3.9}$$

where $\boldsymbol{p}^{(t)} = (\sigma({\boldsymbol{w}^{(t)}}^\top \boldsymbol{\phi}_1^{(t)}), \sigma({\boldsymbol{w}^{(t)}}^\top \boldsymbol{\phi}_2^{(t)}), \ldots, \sigma({\boldsymbol{w}^{(t)}}^\top \boldsymbol{\phi}_{N_t}^{(t)}))^\top$, $\boldsymbol{y}^{(t)} = (y_{t,1}, y_{t,2}, \ldots, y_{t,N_t})^\top$, and $\left[\frac{\partial \mathcal{L}(\boldsymbol{W})}{\partial \boldsymbol{W}}\right]_{*,t}$ denotes t-th column vector of $\frac{\partial \mathcal{L}(\boldsymbol{W})}{\partial \boldsymbol{W}}$.

3.3 Empirical Study

In this section, we empirically evaluate our proposed method by employing it in an in-hospital mortality prediction problem using a real ICU dataset collected from a hospital in Japan.

3.3.1 Experimental Setup

3.3.1.1 Dataset

We used a dataset from a hospital obtained from the Quality Indicator/Improvement Project (QIP) [*1], which is administered by the Department of Healthcare Economics and Quality Management at Kyoto University [52]. Hospitals that voluntarily participated in this project provided their clinical data for research. As preprocessing, we excluded patients under 18, patients with a disease with fewer than 10 patients, and patients with a disease whose initial letter in the ICD-10 code is after N. As the unit of the task, we used the main disease that caused the patient to be admitted to the hospital, which was coded by its 3-digit ICD-10 code. We thus obtained 559 patients and 17 diseases for our study. The codes and names of the diseases that were used in our study are listed in Table 3.2 along with the number of cases (patients) and the death rate for each. As a measure of mortality, we adopted in-hospital mortality; if a patient died during his or her hospital stay, the patient outcome is "death;" otherwise, it is "survival". The age and gender of each patient were extracted and we set two binary features: "Over 65" for age and "Men" for gender. We also included the main disease and all comorbidities of the patient, which are coded into our features by their 4-digit ICD-10 codes. Additionally, we extracted all the medical events for which patients were billed during their hospital stay. These events mainly describe patient interventions such as medication, operative treatment, and laboratory tests, coded by the Diagnosis Procedure Combination (DPC) system, which is a system used for medical billing in Japan. We did not conduct any feature engineering regarding these medical events; we used all the raw codes in the DPC system as our features, and if a patient received an intervention at least once, the corresponding feature was set to 1 (otherwise, it was set to 0).

3.3.1.2 Creation of a similarity matrix

Using the information described above, we created similarity matrix $\boldsymbol{S}^{\text{feat}}$ in Equation (3.4) based on the following rule: if the intervention is medication and if two medicines have the same drug efficacy, then the similarity

[*1] http://med-econ.umin.ac.jp/QIP/

between them is set to 1; otherwise, it is 0. We detail the reason for this choice below. From the raw code of a medication, we recovered information about its efficacy, ingredients, and formulation. Of these, drug formulation was considered irrelevant to mortality. In addition, there are a certain number of medicines with common ingredients but different efficacy; we considered these medicines to potentially play different roles in mortality prediction. Conversely, medicines with same efficacy were considered to play similar roles in mortality prediction even if their ingredients are different, because medicines with different ingredients can be used to remedy the same disease by different mechanisms. Hence, we considered the following two patterns for our choice: (1) two medicines that have a common efficacy but might have different ingredients are similar, and (2) two medicines with common ingredients and efficacies are similar. However, when we choose (2), the similarity matrix in our experiment became sparse. Therefore, we considered choice (1) to better balance the sufficiency and appropriateness of the domain knowledge.

Regarding the similarities between diseases (tasks), the overall structure of our ICD hierarchy can be considered a tree. One reasonable measure of similarity between nodes in a tree is the shared number of ascendants; hence, we exploited the number of shared levels in the ICD hierarchy to define the similarity between two diseases. In our experiment, we created similarity matrix $\boldsymbol{S}^{\text{dz}}$ in Equation (3.3) as follows: the similarity between two diseases is initially set to 1; then, the number of shared levels in the ICD hierarchy is added to the similarity. We call this similarity measure *similarity with considering the root of the tree*. We also compared our proposed method with *similarity without considering the root of the tree*. In this setting, the similarity between two diseases is set to be the number of shared levels in the ICD hierarchy. For example, the similarity between task 1 (J13) and task 2 (J97) in Figure 3.2 is 2 in our proposed method and 1 in the comparison similarity measure.

3.3.1.3 Prediction setting

We randomly sampled 60% of the patients for training and reserved the remaining 40% for evaluation. All hyperparameters of both the proposed and comparison methods were tuned via 3-fold cross-validation in the training dataset. For our proposed methods, λ^{dz} and λ^{feat} were tuned from among $\{0, 10^{-1}, 10^{0}\}$ and $\{0, 10^{-4}, 10^{-3}\}$, respectively, and λ^{rid} was set to 10^{-4}.

Table 3.2: List of the diseases used in our study, along with the number of cases (patients) and the death rate for each disease.

ICD-10 3-digit disease code	Disease name	# of cases	Death rate
A41	Other septicaemia	10	0.50
C15	Malignant neoplasm of oesophagus	11	0.27
C16	Malignant neoplasm of stomach	21	0.29
C18	Malignant neoplasm of colon	10	0.00
C20	Malignant neoplasm of rectum	10	0.20
C22	Malignant neoplasm of liver and intrahepatic bile ducts	26	0.00
C34	Malignant neoplasm of bronchus and lung	28	0.11
G93	Other disorders of brain	53	0.49
I20	Angina pectoris	37	0.05
I21	Acute myocardial infarction	47	0.19
I35	Nonrheumatic aortic valve disorders	21	0.14
I50	Heart failure	42	0.26
I70	Atherosclerosis	14	0.14
I71	Aortic aneurysm and dissection	171	0.15
K56	Paralytic ileus and intestinal obstruction without hernia	22	0.32
K65	Peritonitis	26	0.27
N18	Chronic renal failure	10	0.20
	Total	559	0.20

For comparison methods, all hyperparameters were tuned from among $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$. We repeated the sampling, prediction, and evaluation procedures 20 times. The following three prediction settings were considered: (1) outcome prediction at day 1, (2) outcome prediction at day 2, and (3) outcome prediction at day 3, where the ICU admission day is defined as day 1. For the first, second, and third settings, we used all features associated with the patient that were available 1, 2, and 3 days after admittance to the ICU, respectively. The total number of features was $1,559$, $1,726$, and $1,818$ in the first, second, and third settings, respectively. The statistics of the dataset are listed in Table 3.3.

3.3.1.4 Evaluation method

We adopted four measures to evaluate predictive performance: area under the ROC curve (AUC), Recall (Sensitivity), Precision, and Specificity. AUC

Table 3.3: Statistics of the dataset.

Data	Number
Patients	559
Diseases	17
Features (day 1)	1,559
Features (day 2)	1,726
Features (day 3)	1,818

is widely used because it depends on neither the decision threshold for the probability of positives or negatives nor the class balance. For ICU mortality prediction, AUC is the most standard measure of the predictive performance of a model. If we consider triggering medical alarms to be our application scenario, our aim is to correctly identify high-risk patients. To avoid overlooking high-risk patients, recall (sensitivity), that is, the true positive rate, is crucial. In addition, false alarms inappropriately consume the limited clinical resources; therefore, precision is also important. Specificity, that is, the true negative rate, is also important. In our setting, recall (sensitivity) is the proportion of the number of true positives (patients who died during their hospital stay and were correctly predicted as "will die" by the method) to the total number of patients who died during their hospital stay. Precision is the proportion of the number of true positives to the total number of patients predicted as positive ("will die"). Specificity is the proportion of the number of true negatives (patients who survived during their hospital stay and were correctly identified as "will survive") to the total number of patients who survived their hospital stay. We illustrate the definitions of the last three measures below.

$$\text{Recall} = \frac{\#\text{ of true positives}}{\#\text{ of true positives } + \#\text{ of false negatives}}$$

$$\text{Precision} = \frac{\#\text{ of true positives}}{\#\text{ of true positives } + \#\text{ of false positives}}$$

$$\text{Specificity} = \frac{\#\text{ of true negatives}}{\#\text{ of true negatives } + \#\text{ of false positives}}$$

3.3.1.5 Comparison methods

Table 3.4 lists the comparison methods used in our study. The effect of incorporating domain knowledge of disease classification and medical features was evaluated using two variants of our proposed method: *Proposed-feat* is identical to our proposed method with $\lambda^{\text{dz}} = 0$ in Equation (3.5) and *Proposed-dz* is identical to our proposed method with $\lambda^{\text{feat}} = 0$ in Equation (3.5). We also compared our method with logistic regression without MTL. The single-task learning (STL) method *STL (separate)* learns separate models for each disease without considering the relationships between tasks or features, whereas *STL (common)* is a logistic regression with ℓ_2-norm regularization. STL (common) learns one common model applicable to all diseases, which is the standard method adopted in clinical practice. We compared L2-regularized STL (common) with L1-regularized STL (common) and found that L2-regularization often results in better predictive performance in all the methods in our preliminary experiment. Therefore, we adopted L2-regularization instead of L1-regularization in our study. For MTL baselines, we prepared the following three MTL methods: The first method, *MTL ($\ell_{2,1}$)* [27], incorporates $\ell_{2,1}$-norm regularization, which introduces group sparsity and can be considered as joint feature selection across tasks. The $\ell_{2,1}$-norm of a matrix $\boldsymbol{W}$ is defined as $||\boldsymbol{W}||_{2,1} \equiv \sum_i ||\boldsymbol{W}_{i,*}||_2$, where 2-norm of a vector $\boldsymbol{w}$ is defined as $||\boldsymbol{w}||_2 \equiv (\sum_i ||\boldsymbol{w}_i||^2)^{\frac{1}{2}}$. The second method is *MTL (Trace)* [53], which incorporates trace norm regularization with the assumption that models from different tasks share a common low-dimensional subspace. The third method, *MTL (Mean)* [25] encourages each task parameter to be close to the mean of the parameters of all tasks, incorporating the regularization term $\sum_i ||\boldsymbol{W}_{*,i} - \frac{1}{T}\sum_j \boldsymbol{W}_{*,j}||$. The three MTL baselines are all based on logistic regression, similarly to our proposed method.

3.3.2 Results and Discussion

3.3.2.1 Results in terms of each evaluation measure

Tables 3.5, 3.6, 3.7, and 3.8 shows the averaged AUC, recall (sensitivity), precision, and specificity, respectively, of the various methods in the three prediction settings (outcome predictions at days 1, 2, and 3). We conducted

Table 3.4: Comparison methods used in our experiment.

Method	Regularization	Domain Knowledge		Disease-based Customization
		Task	Feature	
Proposed	Task, Feature, ℓ_2	✓	✓	✓
Proposed-feat	Feature, ℓ_2		✓	✓
Proposed-dz	Task, ℓ_2	✓		✓
STL (separate)	ℓ_2			✓
STL (common)	ℓ_2			
MTL ($\ell_{2,1}$)	$\ell_{2,1}$, ℓ_2			✓
MTL (Trace)	Trace			✓
MTL (Mean)	Mean, ℓ_2			✓

Table 3.5: Comparison of averaged AUCs in each prediction setting. We performed a Wilcoxon signed-rank test for each pair of the method with the highest AUC and another method in each prediction setting. We confirmed that our proposed method statistically significantly ($p < 0.05$) outperformed all other methods in each setting, except for the results in bold font, if any. Regarding the boldface results, there was no statistically significant ($p < 0.05$) improvement using our proposed method over the comparison methods. There was no method that performed equally to or better than our proposed method throughout all the settings.

Method	AUC		
	Day 1	Day 2	Day 3
Proposed	**0.80**	**0.79**	**0.82**
Proposed-feat	0.72	0.74	0.75
Proposed-dz	0.79	**0.79**	**0.82**
STL (separate)	0.73	0.74	0.75
STL (common)	0.76	0.76	0.78
MTL ($\ell_{2,1}$) [27]	0.70	0.72	0.74
MTL (Trace [53])	0.70	0.73	0.76
MTL (Mean [25])	0.78	**0.78**	0.81

Wilcoxon signed-rank tests for each pair of the method with the highest performance and another method in each prediction setting. In the following, we discuss our results in terms of each evaluation measure.

AUC:

In summary, there was no method that performed equally to or better than our proposed method throughout all settings in terms of AUC. The proposed

Table 3.6: Comparison of averaged recalls in each prediction setting. We conducted a Wilcoxon signed-rank test for each pair of the method with the highest recall and another method in each prediction setting. We confirmed that our proposed method statistically significantly ($p < 0.05$) outperformed all other methods in each setting, except for the results in bold font. Regarding the boldface results, there was no statistically significant ($p < 0.05$) improvement by our proposed method over the comparison method, although there was no method that performed equally to or better than our proposed method throughout all the settings.

Method	Recall		
	Day 1	Day 2	Day 3
Proposed	**0.47**	**0.44**	**0.49**
Proposed-feat	0.24	0.30	0.31
Proposed-dz	0.42	**0.45**	**0.47**
STL (separate)	0.24	0.29	0.30
STL (common)	0.34	0.37	0.39
MTL ($\ell_{2,1}$) [27]	0.22	0.27	0.27
MTL (Trace [53])	0.19	0.27	0.32
MTL (Mean [25])	0.39	0.40	**0.46**

method outperformed all the other methods especially well in the earliest prediction setting of day 1.

(1) The proposed method performed equally to or better than all methods without domain knowledge of disease similarity throughout all settings, which suggests that the appropriate incorporation of task relatedness, i.e., how each disease relates to others, can improve predictive performance significantly, although MTL does not necessarily improve predictive performance. (2) The proposed method outperformed *Proposed-dz* in the earliest prediction setting at day 1, whereas there were no statistically significant differences between these two methods in the other prediction settings. Because the amount of data that is available for prediction at day 1 is less than that available for the other prediction settings, the above result suggests that incorporating domain knowledge related to the medical classification of EMRs can also contribute to the improvement of predictive performance when data is sparse. (3) The proposed method outperformed the most standard method in clinical domains, *STL (common)*, in all prediction settings, suggesting that we could substitute our proposed method for the conventional method.

Table 3.7: Comparison of averaged precisions in each prediction setting. We conducted a Wilcoxon signed-rank test for each pair of the method with the highest precision and another method in each prediction setting. We confirmed that our proposed method statistically significantly ($p < 0.05$) outperformed all other methods in each setting, except for the results shown in bold. Of the results in bold, there was no statistically significant ($p < 0.05$) improvement using our proposed method over the comparison method. Only *Proposed-dz* performed equally to or better than our proposed method throughout all the settings.

Method	Precision		
	Day 1	Day 2	Day 3
Proposed	**0.57**	**0.56**	**0.63**
Proposed-feat	0.43	0.48	0.52
Proposed-dz	**0.55**	**0.57**	**0.62**
STL (separate)	0.44	0.49	0.53
STL (common)	0.50	0.53	0.58
MTL ($\ell_{2,1}$) [27]	0.48	0.48	0.52
MTL (Trace [53])	0.44	**0.51**	0.56
MTL (Mean [25])	**0.54**	**0.54**	0.57

Table 3.8: Comparison of averaged specificities in each prediction setting. We conducted a Wilcoxon signed-rank test for each pair of the method with the highest specificity and another method in each prediction setting. The results in bold font outperformed all nonboldface results. There was no method that performed equally to or better than MTL ($\ell_{2,1}$) and MTL (Trace) throughout all settings in terms of specificity.

Method	Specificity		
	Day 1	Day 2	Day 3
Proposed	0.92	0.92	0.93
Proposed-feat	0.92	0.92	0.93
Proposed-dz	0.92	0.92	0.93
STL (separate)	0.92	0.93	0.93
STL (common)	0.92	0.92	0.93
MTL ($\ell_{2,1}$) [27]	**0.94**	**0.93**	**0.94**
MTL (Trace [53])	**0.94**	**0.93**	**0.94**
MTL (Mean [25])	0.92	0.92	0.92

Recall (sensitivity):

The results in terms of recall (sensitivity) were similar to those of AUC: there

was no method that performed equally to or better than our proposed method throughout all settings. In particular, the proposed method outperformed all other methods in the earliest prediction setting. The three observations noted in the AUC results also apply to the recall results.

Precision:

In terms of precision, there was no method that performed equally to or better than our proposed method throughout all settings, except for *Proposed-dz*; the proposed method and *Proposed-dz* performed equally well. Observations (1) and (3) noted in the discussion of the AUC results also apply to the precision results.

Specificity:

In terms of specificity, the variances of the different methods were relatively small in all prediction settings; all results were greater than 92 $\%$ and the maximum difference between the results of any two methods was 2 $\%$ in all settings. For comparison, the maximum difference for AUC, recall, and precision, were 10 $\%$, 28 $\%$, and 14 $\%$, respectively. Of the various methods, *MTL ($\ell_{2,1}$)* and *MTL (Trace)* could be considered the best: there was no method that performed equally to or better than these methods throughout all settings in terms of specificity.

3.3.2.2 Effect of similarity between diseases

Table 3.9 shows the comparison of the AUCs between the proposed method, which incorporates disease similarity considering the root of the tree, and the method that differs only in that it does not consider the root of the tree, in the prediction setting of day 1. We confirmed that the similarity that considered the root of the tree statistically significantly ($p < 0.05$) outperformed the other method, except in specificity, for which there was no statistically significant difference between the two methods. This result suggests that all the diseases might share some common predictive information.

3.3.2.3 Comparison of AUCs for each disease

Figure 3.3 compares the AUCs for each disease resulting from the Proposed, STL (common), and MTL (Mean) methods in the prediction setting of day 1. We chose STL (common) method as the representative STL method and MTL (Mean) as the representative MTL method, because each had the highest predictive performance among STL and MTL methods, respectively. We

Table 3.9: Effect of the similarity between diseases in the prediction setting of day 1. We conducted a Wilcoxon signed-rank test for the pair of our proposed methods that incorporate similarity with and without considering the root of the tree. The boldface results indicate better performance than the nonboldface results. We confirmed that considering the root of the tree resulted in statistically significant ($p < 0.05$) improvement over not considering it, except in specificity, for which there was no statistically significant difference between the two methods.

Similarity	AUC	Recall	Precision	Specificity
With root	**0.80**	**0.47**	**0.57**	0.92
Without root	0.77	0.34	0.49	0.91

excluded diseases C18 and C22 from the evaluation of target diseases because there was no positive example for either disease. There was no clear trend observed for the kind of diseases for which each method performs best in terms of the number of patients or the number of similar diseases. Making any such tendency clear would be important future work.

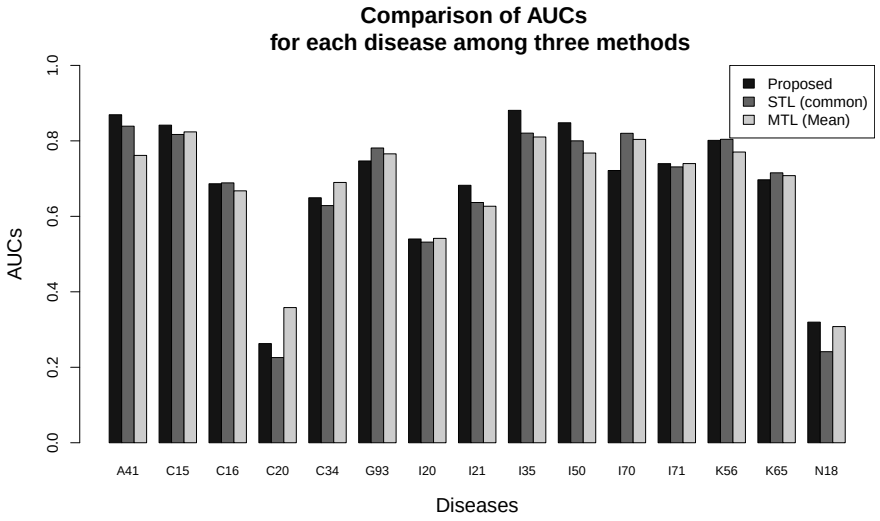

Fig. 3.3: Comparison of AUCs for each disease among the Proposed, STL (common), and MTL (Mean) methods in the prediction setting of day 1.

3.4 Related Work

3.4.1 MTL for Clinical Problems

MTL has been explored extensively for health care problems, including disease progression models [46, 47], prediction of clinical variables [54], patient outcome prediction [38], and illness severity forecasting in ICU [11]. Zhou et al. [46, 47] proposed a temporal group Lasso regularizer to encourage temporal smoothness of the predictive models across different time points. Quanz and Huan [54] applied multitask feature selection to a set of images obtained from different sources, such as MRI and PET scans, extracting a common subset of features for different tasks. Gupta et al. [38] exploited hierarchical Dirichlet process to learn a specific model for each group of interventions. Ghassemi et al. [11] examined the use of multivariate time-series modeling with a multitask Gaussian process to forecast the severity of illnesses of ICU patients.

However, none of these studies considers disease as the task unit. One exception is the recent work by Wang et al. [39] in which the authors define the onset risk prediction for a disease as a task and formulate the onset risk prediction for multiple diseases as an MTL problem. Our work differs from theirs in that our method is capable of incorporating domain knowledge into data-driven approaches. Although we exploited ICD-10 and drug efficacy information as domain knowledge in our experiments, our method can incorporate any prior knowledge encoded as similarity matrices; disease similarities and similarities between EMRs can be constructed by using prior knowledge from domain experts or by leveraging multiple other information sources. One important future direction would be to explore the construction of such disease and EMRs similarities [55]. In addition, our work differs from previous studies in that we incorporate the similarities between both the tasks (diseases) and features derived from EMRs simultaneously using graph Laplacians, which results in cross-regularization in the MTL framework. Although feature regularizer has been explored in STL to capture intrinsic relationships among features [56], it has not been explored in MTL. In addition, although cross-regularization has been proposed for semi-supervised multi-label learning [57], in which two graphs are initially constructed on instance

and category levels to incorporate instance and category similarities, it has also not been explored in MTL.

We first formulated the ICU mortality risk prediction problem as an MTL in which a task corresponds to a disease by integrating domain knowledge encoded as similarities between diseases and among EMRs, thereby incorporating disease-specific contexts into predictive mortality modeling.

3.5 Summary

Our study considered disease-specific contexts in mortality modeling in the ICU, for which we formulated mortality prediction in terms of MTL in which a task corresponds to a disease. Our method effectively incorporated domain knowledge about the medical classifications of both diseases and EMRs into a data-driven approach. Experimental results on a real hospital dataset demonstrated the effectiveness of our proposed method, which appropriately incorporated task relatedness, i.e., how each disease relates to others, leading to an improvement in predictive performance. The incorporation of domain knowledge in the form of the medical classification of EMRs was also effective. Our proposed method and results could prove effective in enhancing the understanding of disease-specific contexts, as well as improve the predictive performance of ICU mortality modeling.

Chapter 4

Learning Implicit Tasks for Patient-Specific Risk Modeling

4.1 Introduction

The ICU is an especially demanding environment for clinicians owing to the *diversity of the patients*; ICU clinicians are faced with patients with a wide variety of disease types, all of whom are severely ill. For instance, some patients are admitted to the ICU due to infectious diseases such as sepsis and pneumonia, whereas others are postoperative patients admitted after major surgery. Furthermore, patients are usually associated with multiple diseases; thus, together with the diversity of diseases, ICU patients present rather varied and complicated clinical states.

To date, numerous studies have focused on mortality risk predictive modeling under the assumption that mortality risk may be used as a surrogate for severity assessment of ICU patients [6, 9, 10, 11, 12]. Accurate risk prediction could lead to preventive actions by, for example, a medical alarm. Thus far, most studies on mortality risk prediction for ICU patients have implicitly assumed that one common risk model could be developed and applied to all patients; however, this approach might fail to capture the diversity of ICU patients. Thus, we explored disease-specific risk modeling for ICU patients by employing MTL in which a task corresponds to a disease [18]. We assumed

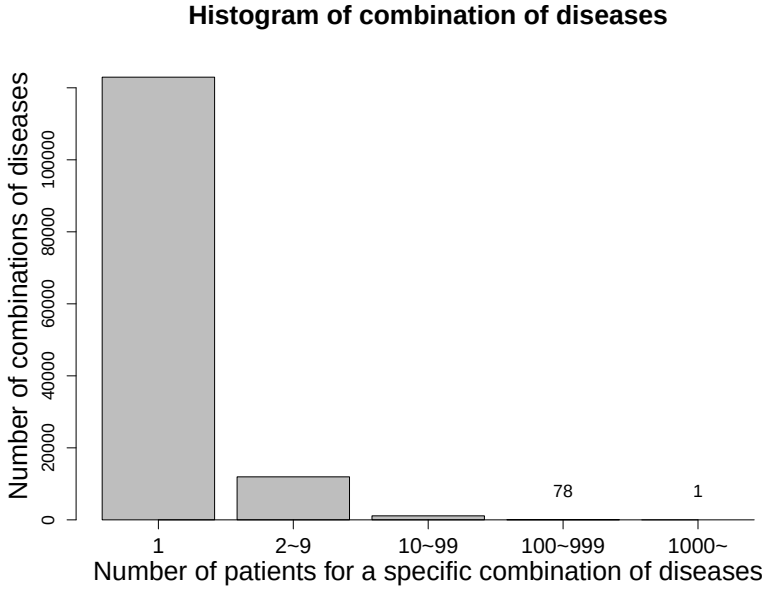

Fig. 4.1: Histogram of disease combinations in an ICU dataset. Many combinations are specific to a single patient.

that each disease has a specific prediction rule that explains its mortality risk, and therefore, that customizing the risk model for each disease would enhance predictive modeling.

Yet, this approach continues to be insufficient to capture *patient-specific* aspects of mortality risk modeling for ICU patients. Specifically, ICU patients are usually associated with *multiple diseases*, further complicating their clinical states. In such a case, one straightforward approach might be to create a task corresponding to the combination of diseases that each patient has; however, this approach would be ineffective because of combinational explosion among diseases. This problem is illustrated in Figure 4.1, which shows a histogram of disease combinations that was created from an ICU dataset consisting of about 200,000 patients from about 170 hospitals in Japan. Each patient is associated with his or her main disease, identified by a three-digit ICD-10 code, and up to 4 comorbidities. Many combinations are specific to a single patient. Thus, a naïve approach such as creating tasks corresponding to combinations of diseases, would be unfeasible.

In this study, we propose an MTL method for ICU mortality prediction that is capable of producing patient-specific risk models reflecting the collection

of diseases associated with the patient. We do not explicitly create tasks corresponding to a collection of diseases; instead, we assume *implicit*, or *latent* tasks and learn a latent representation of the diseases. Figure 4.2 illustrates our model. Specifically, we assume a small number of latent basis tasks, where each latent task is associated with its own parameter vector (which together comprise parameter matrix $\boldsymbol{L}$), and a parameter vector for a specific patient (which comprises parameter matrix $\boldsymbol{W}$) is constructed as a linear combination of these. The latent representation of a patient, namely, the coefficients of the combination (which comprise parameter matrix $\boldsymbol{S}$), is learned based on the collection of diseases the patient is associated with (which comprises the association matrix $\boldsymbol{A}$). Our method could be considered an MTL method in which latent tasks are learned based on the collection of diseases.

The contributions of our study are as follows:

- We propose an MTL method for ICU mortality prediction that can produce a patient-specific risk model reflecting the collection of diseases with which the patient is associated. For patient-specific modeling, one critical issue is to determine the unit that should be used to model patient specificity. Our method enables us to learn the task unit based on the collection of diseases the patients are associated with by introducing latent basis tasks.
- We demonstrate the effectiveness of our method using a real-world hospital dataset. Our method is capable of constructing patient-specific models from different viewpoints.

4.2 Related Work

4.2.1 Patient-Specific Modeling

Addressing both the *specificity* and *commonality* among a patient population —that is, addressing the specificity of the target patient while capturing the common structure shared by the population— has been one of the most fundamental issues at the intersection of clinical and machine-learning research. Obtaining clinically useful models requires the model to be customized to the target of the analysis; however, restricting the data to those that are faithfully relevant to the target often results in an impractically small sample size.

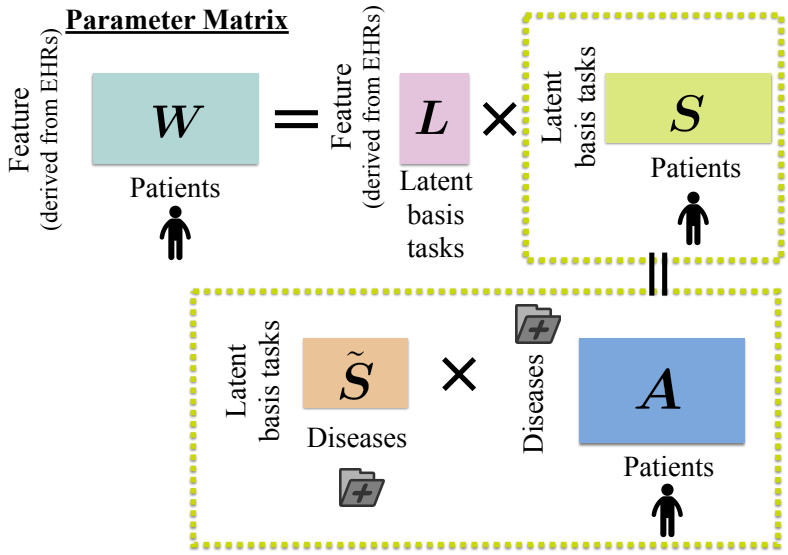

Fig. 4.2: Our MTL model for patients with multiple diseases.

Thus, to develop clinically useful target-specific models, recent studies have focused on MTL and transfer learning for clinical data. One critical issue is to determine the aspects that capture patient specificity, i.e., how to define tasks for patient-specific modeling. Jenna et al. investigated the effectiveness of hospital-specific [36] modeling via a feature-representation-transfer method. Gong et al. exploited an instance-transfer method for hospital- and surgery-specific risk modeling [37]. Liu and Hauskrecht developed a forecasting model that captures both patient-specific time-series patterns and population-level information for clinical time-series data [58]. In a previous study, we proposed an MTL method in which a task corresponds to a disease [18]. Our study differs from other research described in this section in that we learn the task unit by assuming a small number of latent basis tasks that are learned from the collection of diseases each patient is associated with. To the best of our knowledge, this is the first study in which patient risk modeling is formulated as MTL in which a task corresponds to a combination of diseases.

4.2.2 Mortality Modeling in the ICU

As noted before, thus far, it has been implicitly assumed that one common predictive model should be developed and applied to all diseases. In a pre-

vious study [18], we formulated mortality prediction as multitask learning in which a task corresponds to the patient's main disease, thereby producing disease-specific models. However, this method is unable to accommodate the collection of diseases each ICU patient is associated with; each task corresponds to one disease. However, because patients are usually associated with multiple diseases, it would be necessary to construct tasks based on the collection of diseases each patient is associated with to capture patient specificity. In addition, the previous method does not learn the relations between diseases, whereas the proposed method learns these relationships by learning latent tasks from the collection of diseases. Contrary to this, the previous method increases the similarity between the model parameters of two diseases if the diseases are similar in terms of domain knowledge via regularization. Although domain knowledge can be exploited to some extent, it is likely that intrinsic relations between diseases can be learned from data and exploited for prediction purposes.

4.2.3 Multitask Learning

One of the most fundamental issues in MTL is how to introduce an inductive bias in modeling task relationships. There have been numerous studies that have attempted to introduce appropriate assumptions: task parameters might be assumed to lie in close proximity with each other in some geometric sense [25], or they might be assumed to lie in a low-dimensional subspace [53], or to share a common prior [26]. One major challenge in MTL is avoiding negative-transfer; that is, selectively sharing information among tasks in order that unrelated tasks do not affect each other. Approaches to this challenge include clustering tasks [31] and learning groupings of tasks with overlap [32].

Our method could be considered an MTL method with the assumption that tasks lie in a low-dimensional subspace, similar to several of the studies described above [32, 53]. However, our method fundamentally differs from other research in that we learn the unit of the task. We assume components of the tasks, namely, a collection of diseases, and assume that a combination of components produces a task. We do not explicitly list the combinations.

Table 4.1: Notation and descriptions.

Notation	Description
N	Number of patients
M	Number of features derived from EMRs
$\boldsymbol{x}_n$	M-dimensional feature vector for the n-th patient
D	Number of diseases
y_n	Class label for the n-th patient: $y_n \in \{0, 1\}$, where 0 and 1 represents survival and death in hospital, respectively.
$\boldsymbol{w}_n$	M-dimensional parameter vector for the n-th patient
$\boldsymbol{W}$	$M \times N$ parameter matrix: $\boldsymbol{W} \equiv [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{N'}]$
K	Number of latent basis tasks
$\boldsymbol{L}$	$M \times K$ parameter matrix with each column representing a latent basis task
$\boldsymbol{S}$	$K \times N$ parameter matrix containing the weights of linear combination for each patient
$\boldsymbol{A}$	$D \times N$ matrix where $\boldsymbol{A}_{d,n} = 1$ if the n-th patient is associated with the d-th disease and $\boldsymbol{A}_{d,n} = 0$ otherwise.
$\tilde{\boldsymbol{S}}$	$K \times D$ parameter matrix with each column representing a latent representation of a disease.
$\boldsymbol{W}_0$	$M \times N$ matrix with each column containing a vector obtained from a single-task learning method.

4.3 Learning Patient-Specific Risk Models

4.3.1 Problem Definition

In this section, we describe our approach for learning patient-specific risk models where each patient is associated with one or more diseases. Let N denote the number of total patients. A risk model is developed for each patient. The n-th patient is represented by an M-dimensional feature vector $\boldsymbol{x}_n$ derived from EMRs, which contain a variety of information about patients, such as their demographic profile, clinical history, and medications. Each patient is also associated with one or more diseases, specified by their ICD codes, whose total number is denoted by D. Each patient is associated with a binary class label, $y_n \in \{0, 1\}$, where $y_n = 1$ when the patient died during his or her hospital stay and $y_n = 0$ otherwise. Since we are interested in the probability

of mortality, we opt for logistic regression and represent the posterior probability of the outcome of patient n being death as $\Pr[y_n = 1|\boldsymbol{x}_n] = \sigma(\boldsymbol{w}_n{}^T \boldsymbol{x}_n)$, where $\sigma(a)$ is the sigmoid function $\sigma(a) \equiv (1 + \exp(-a))^{-1}$, and $\boldsymbol{w}_n$ is an M-dimensional model parameter vector for the n-th patient.

Decomposition model. To address the data sparsity inherent in constructing patient-specific risk models, we learn the models for all patients jointly by sharing information across patients. We represent the whole parameter matrix as an $M \times N$ parameter matrix $\boldsymbol{W}$, where the n-th column vector $\boldsymbol{w}_n$ denotes a parameter vector for the n-th patient. We also assume there are K latent basis tasks and that a specific risk model for each patient can be represented as a linear combination of these latent basis tasks. We can therefore write parameter matrix $\boldsymbol{W}$ as $\boldsymbol{W} = \boldsymbol{L}\boldsymbol{S}$, where $\boldsymbol{L}$ is an $M \times K$ matrix with each column representing a latent basis task, and $\boldsymbol{S}$ is a $K \times N$ matrix containing the linear combination weights for each patient. The parameter for the n-th patient $\boldsymbol{w}_n$ is given as $\boldsymbol{L}\boldsymbol{S}_{*,n}$. The predictive structure of the latent tasks is captured by matrix $\boldsymbol{L}$ and the latent representation of the patients is captured by matrix $\boldsymbol{S}$. Next, we exploit the relations between diseases and patients. Let $\boldsymbol{A}$ denote a $D \times N$ matrix containing the relation information between diseases and patients. Specifically, $\boldsymbol{A}_{d,n} = 1$ if the n-th patient is associated with the d-th disease and $\boldsymbol{A}_{d,n} = 0$ otherwise. Using this relational information, we further rewrite matrix $\boldsymbol{S}$ as $\boldsymbol{S} = \tilde{\boldsymbol{S}}\boldsymbol{A}$, where $\tilde{\boldsymbol{S}}$ is a $K \times D$ matrix with each column representing a latent representation of a disease.

Prior knowledge. We overcome the problem of data sparsity and guide effective decomposition of the parameter matrix by assuming some prior knowledge relating to population-level information, which is some common structure shared among all patients. We assume an M $\times$ N matrix $\boldsymbol{W}_0$ with each column containing a parameter vector obtained from a single-task learning method that is learned from all patients in the training dataset. That is, $\boldsymbol{W}_0 \equiv [\boldsymbol{w}_0, \boldsymbol{w}_0, \ldots, \boldsymbol{w}_0]$, where $\boldsymbol{w}_0$ is an M-dimensional parameter vector learned by applying a machine learning method to all patients in the training dataset.

Our goal is to estimate the mortality risk of a new patient represented by an M-dimensional feature vector $\boldsymbol{x}_{n'}$, given a D-dimensional vector containing the collection of diseases the patient is associated with, some observed training dataset $\{(\boldsymbol{x}_n, y_n)\}_{n=1,\ldots,N}$, and a $D \times N$ patient-disease matrix $\boldsymbol{A}$. After learning parameter matrices $\boldsymbol{L}$ and $\tilde{\boldsymbol{S}}$ using a training dataset, we can con-

struct a parameter matrix specialized to new patients, given a patient-disease relation matrix that encodes the collection of diseases each new patient is associated with.

4.3.2 Optimization Problem

We define our loss function as the log loss denoted by $\mathcal{L}(\tilde{\boldsymbol{S}}, \boldsymbol{L})$:

$$\mathcal{L}(\tilde{\boldsymbol{S}}, \boldsymbol{L}) \equiv -\frac{1}{N}\sum_{n=1}^{N}\{y_n \log \sigma({\boldsymbol{A}_{*,n}}^{\top}\tilde{\boldsymbol{S}}^{\top}\boldsymbol{L}^{\top}\boldsymbol{x}_n) + (1-y_n)\log(1-\sigma({\boldsymbol{A}_{*,n}}^{\top}\tilde{\boldsymbol{S}}^{\top}\boldsymbol{L}^{\top}\boldsymbol{x}_n)\}. \tag{4.1}$$

We include several regularization terms to exploit our prior knowledge and to avoid overfitting. First, to utilize population-level information, we include the following regularization term Ω_1:

$$\Omega_1 \equiv ||\boldsymbol{L}\tilde{\boldsymbol{S}}\boldsymbol{A} - \boldsymbol{W}_0||_F^2, \tag{4.2}$$

where $\boldsymbol{W}_0$ is a matrix with each column containing a parameter vector obtained from a single-task learning method.

Adding the following $\ell 2$-norm regularization term

$$\Omega_2 \equiv ||\tilde{\boldsymbol{S}}||_F^2, \tag{4.3}$$

we define our regularization term as

$$\Omega \equiv \lambda_1\Omega_1 + \lambda_2\Omega_2, \tag{4.4}$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are hyperparameters for tuning the weights of the regularization terms Ω_1 and Ω_2, respectively.

The optimization problem is thus defined as follows:

$$\min_{\tilde{\boldsymbol{S}}, \boldsymbol{L}} \mathcal{L}(\tilde{\boldsymbol{S}}, \boldsymbol{L}) + \Omega. \tag{4.5}$$

We show the optimization problem is convex in $\tilde{\boldsymbol{S}}$ for a fixed $\boldsymbol{L}$ and vice versa.

The derivatives of the loss function with respect to $\tilde{\boldsymbol{S}}$ and $\boldsymbol{L}$, respectively, are given as follows:

$$\frac{\partial \mathcal{L}}{\partial \tilde{\boldsymbol{S}}} = \sum_{n=1}^{N} (\sigma_n - y_n) \boldsymbol{L}^\top \boldsymbol{x}_n {\boldsymbol{A}_{*,n}}^\top, \tag{4.6}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{L}} = \sum_{n=1}^{N} (\sigma_n - y_n) \boldsymbol{x}_n {\boldsymbol{A}_{*,n}}^\top \tilde{\boldsymbol{S}}^\top, \tag{4.7}$$

where $\sigma_n \equiv \sigma({\boldsymbol{A}_{*,n}}^\top \tilde{\boldsymbol{S}}^\top \boldsymbol{L}^\top \boldsymbol{x}_n)$.

The derivatives of the regularization term with respect to $\tilde{\boldsymbol{S}}$ and $\boldsymbol{L}$ are given as

$$\frac{\partial \Omega}{\partial \tilde{\boldsymbol{S}}} = 2\lambda_1 (\boldsymbol{L}^\top \boldsymbol{L} \tilde{\boldsymbol{S}} \boldsymbol{A} \boldsymbol{A}^\top - 2\boldsymbol{L}^\top \boldsymbol{W}_0 \boldsymbol{A}^\top) + 2\lambda_2 \tilde{\boldsymbol{S}}. \tag{4.8}$$

$$\frac{\partial \Omega}{\partial \boldsymbol{L}} = 2\lambda_1 (\boldsymbol{L} \tilde{\boldsymbol{S}} \boldsymbol{A} \boldsymbol{A}^\top \tilde{\boldsymbol{S}}^\top - 2\boldsymbol{W}_0 \boldsymbol{A}^\top \tilde{\boldsymbol{S}}^\top). \tag{4.9}$$

The second derivative of the loss function and regularization term Ω_1 with respect to $\tilde{\boldsymbol{S}}$ are given, respectively, as

$$\frac{\partial \textbf{vec}\,(\mathcal{L}')}{\partial \textbf{vec}\left(\tilde{\boldsymbol{S}}\right)^\top} = \sum_{n=1}^{N} \sigma_n (1 - \sigma_n) \boldsymbol{V} \boldsymbol{V}^\top, \tag{4.10}$$

$$\frac{\partial \textbf{vec}\,(\Omega_1')}{\partial \textbf{vec}\left(\tilde{\boldsymbol{S}}\right)^\top} = 2\boldsymbol{A}\boldsymbol{A}^\top \otimes \boldsymbol{L}^\top \boldsymbol{L}, \tag{4.11}$$

where $\boldsymbol{V} \equiv \textbf{vec}\left(\boldsymbol{L}^\top \boldsymbol{x}_n {\boldsymbol{A}_{*,n}}^\top\right)$, and $\otimes$ is Kronecker product.

The loss function is convex in $\tilde{\boldsymbol{S}}$ for a fixed $\boldsymbol{L}$ and vice versa, since the sum of positive semidefinite matrices is positive semidefinite. Since the Kronecker product of two positive semidefinite matrices is positive semidefinite, Equation (4.11) produces a positive semidefinite matrix; hence, the regular-

ization term is convex in $\tilde{\boldsymbol{S}}$ for a fixed $\boldsymbol{L}$. Similarly, the regularization term is convex in $\boldsymbol{L}$ for a fixed $\tilde{\boldsymbol{S}}$. We adopt an alternating optimization procedure; for each optimization problem, the optimal solution is found using standard gradient-based methods. We applied the L-BFGS optimizer [51] with the derivatives above. To initialize $\boldsymbol{L}$, we assume an M $\times$ N matrix $\boldsymbol{W}_0$ with each column containing a parameter vector obtained from a single-task learning method. Matrix $\boldsymbol{L}$ is then initialized to the top-K left singular vectors of $\boldsymbol{W}_0$: $\boldsymbol{W}_0 = \boldsymbol{U}\Sigma\boldsymbol{V}$. The alternating optimization procedure is terminated when some prearranged criterion is satisfied.

4.4 Empirical Study

4.4.1 Experimental Setup

4.4.1.1 Dataset

We used a dataset from a hospital in Japan, which was constructed as part of the Quality Indicator/Improvement Project [52] administered by the Department of Healthcare Economics and Quality Management at Kyoto University. All patients in the dataset were at least 18 years old and underwent ICU treatment at some point during their hospital stay. For disease coding, we adopted the three-digit ICD-10 codes and extracted the following diseases for each patient: the main disease that caused the patient's admission and up to 4 comorbidities the patient had at the time of admission. The dataset consists of patients whose outcome is either death or survival after ICU discharge, in hospital, which comprises of 312 patients associated with a total of 244 diseases. For features, we used the age, gender, main disease, and comorbidities of the patient, in addition to all the medical events for which patient was billed during the hospital stay. For the age and gender, we created two binary features: "Over 65" and "Men." The medical events mainly describe patient interventions such as medication, procedures, and laboratory tests. For patients who received an intervention once or more than once, the corresponding feature was set to 1, and 0 otherwise.

4.4.1.2 Prediction setting

To evaluate predictive performance, we measured the area under the ROC curve (AUC), recall, precision, and specificity. We randomly sampled 80%

of the patients to create the training dataset and used the remaining 20% for evaluation. We prepared two prediction settings: ICU admission and ICU discharge. For prediction at the ICU admission point, we used all features associated with the patient that were available at that time; for prediction at the ICU discharge point, we used all features associated with the patient that were available 1 day before the day he or she was discharged from the ICU. The total number of features was 1,003 in the first setting and 1,525 in the second. The statistics of the dataset are listed in Table 4.2.

Table 4.2: Statistics of the dataset.

Data	Number
Patients	312
Diseases	244
Features (ICU admission point)	1,003
Features (ICU discharge point)	1,525

We repeated the sampling, prediction, and evaluation procedures 100 times and calculated the mean of the 100 trials. Hyperparameters were tuned by 3-fold cross-validation in the training dataset. For our proposed method, λ_1 and λ_2 were set to 10^{-3} and 10^{-5}, respectively. The number of latent tasks K was tuned among $\{2^4, 2^5\}$. Matrix $\boldsymbol{W}_0$ was constructed by applying logistic regression with ℓ_2-norm regularization to all patients. $\boldsymbol{W}_0$ was learned using only the training data in each iteration. In this experiment, increasing the number of iterations in the optimization process did not improve prediction performance. Hence, we only estimated $\tilde{\boldsymbol{S}}$ using the initial $\boldsymbol{L}$ throughout the experiment.

4.4.1.3 Comparison methods

We compared our proposed method with eight others. We first prepared the following two variants of our method. First, *Proposed-w/o-A* is adopted to determine the effect of association matrix $\boldsymbol{A}$. This method learns two parameter matrices $\boldsymbol{L}$ and $\boldsymbol{S}$ without introducing $\boldsymbol{A}$; more specifically, we only used the main disease for each patient in constructing $\boldsymbol{A}$. Second, we determined the effect of regularization by preparing *Proposed-w/o-pop*, which is identical to our proposed method with $\lambda_1 = 0$ in Equation (4.4). The next two methods are single-task learning methods: *STL (separate)* learns separate

Table 4.3: Comparison of various methods used in our experiment.

Method	Unit of Tasks	Prior Knowledge	MTL
Proposed	Latent	Population model	✓
Proposed-w/o-A	Latent	Population model	✓
Proposed-w/o-pop	Latent		✓
STL (separate)	N/A		
STL (common)	N/A		
MTL (Trace [53])	Disease		✓
MTL (Mean [25])	Disease		✓
MTL ($\ell_{2,1}$) [27]	Disease		✓
MTL (DM [18])	Disease	Similarity among diseases, Similarity among features	✓

models for each disease using only data relating to the patient's main disease. *STL (common)* learns one common model that is applicable to all patients by using data from all diseases. The remaining four methods are MTL baselines; for these methods, tasks are defined as patients' main diseases. The first method is *MTL (Trace [53])*, which incorporates trace norm regularization with the assumption that models from different tasks share a common low-dimensional subspace. The second method is *MTL (Mean [25])*, which assumes that each task parameter vector is close to the mean vector of all tasks. The third method, *MTL ($\ell_{2,1}$ [27])*, incorporates $\ell_{2,1}$-norm regularization to introduce group sparsity and can be considered as joint feature selection across tasks. The last method is *MTL (DM [18])*, which integrates domain knowledge relating to the diseases and EMRs via two graph Laplacians. All single-task and multitask learning baselines are based on logistic regression with ℓ_2 regularization. For STL (separate), ST(common), *MTL (Trace)*, *MTL (Mean)*, and *MTL ($\ell_{2,1}$)*, the hyperparameter for the ℓ_2 norm was set to 10^{-4} and all the other regularization hyperparameters were tuned from among $\{10^0, 10^{-2}, 10^{-4}\}$. For *MTL (DM)*, the hyperparameter relating to task similarity was tuned from among $\{10^{-1}, 10^0\}$, and the hyperparameter relating to the feature similarity and the ℓ_2 regularization hyperparameter were both set to 10^{-4}. Table 4.3 shows a comparison of the methods used in our experiment.

Table 4.4: Comparison of averaged AUCs. We conducted a Wilcoxon signed rank-test for each pair of the method with the highest AUC and another method in each prediction setting. The results shown in bold indicate the method statistically significantly ($p < 0.05$) outperformed all the other methods in the relevant setting. For prediction at ICU admission, MTL (DM) method outperformed all other methods. For prediction at ICU discharge, our proposed method outperformed all other methods.

Method	AUC	
	Admission point	Discharge point
Proposed	0.63	**0.71**
Proposed-w/o-A	0.58	0.62
Proposed-w/o-pop	0.58	0.59
STL (separate)	0.51	0.56
STL (common)	0.64	0.68
MTL (Trace [53])	0.52	0.54
MTL (Mean [25])	0.56	0.61
MTL ($\ell_{2,1}$ [27])	0.52	0.55
MTL (DM [18])	**0.65**	0.69

Table 4.5: Comparison of averaged recalls. We conducted a Wilcoxon signed-rank test for each pair of the method with the highest recall and another method in each prediction setting. The result given in bold indicates the method statistically significantly ($p < 0.05$) outperformed all other methods in the concerned setting. In both settings, Proposed-w/o-A method outperformed all other methods.

Method	Recall	
	Admission point	Discharge point
Proposed	0.45	0.57
Proposed-w/o-A	**0.65**	**0.66**
Proposed-w/o-pop	0.30	0.32
STL (separate)	0.47	0.54
STL (common)	0.38	0.38
MTL (Trace [53])	0.47	0.52
MTL (Mean [25])	0.53	0.57
MTL ($\ell_{2,1}$ [27])	0.48	0.53
MTL (DM [18])	0.39	0.39

4.4.2 Results and Discussion

4.4.2.1 Predictive performance

Table 4.4 compares the AUCs of the various methods. For prediction at ICU admission, MTL (DM) outperformed all other methods, whereas for predic-

Table 4.6: Comparison of averaged precisions. We conducted a Wilcoxon signed-rank test for each pair of the method with the highest precision and another method in each prediction setting. The bold result, with the highest precision, indicates that the method statistically significantly ($p < 0.05$) outperformed all nonboldface methods in the concerned setting. For prediction at ICU admission, STL (common) and MTL (DM) outperformed all other methods. For prediction at ICU discharge, MTL (DM) outperformed all other methods.

Method	Precision	
	Admission point	Discharge point
Proposed	0.26	0.35
Proposed-w/o-A	0.20	0.21
Proposed-w/o-pop	0.23	0.29
STL (separate)	0.15	0.18
STL (common)	**0.26**	0.40
MTL (Trace [53])	0.15	0.17
MTL (Mean [25])	0.17	0.19
MTL ($\ell_{2,1}$ [27])	0.15	0.17
MTL (DM [18])	**0.27**	**0.43**

tion at ICU discharge, our proposed method outperformed all other methods. Since the features available at the ICU admission point are only those about diseases, age, and gender, effective patient-specific characteristics might be relatively difficult to obtain. Conversely, MTL (DM) method exploits domain knowledge relating to similarities among diseases, which might make MTL (DM) relatively effective in situations in which data are sparse.

In terms of AUC, our proposed method outperformed all other methods for prediction at ICU discharge. The performance improvement compared with the variant of our method Proposed-w/o-A suggests that exploiting multiple diseases' information for each patient can improve prediction performance. Similarly, the performance improvement compared with *Proposed-w/o-pop* suggests that population-level information is important for effective prediction. The fact that our method outperformed the two single-task learning methods in the ICU discharge prediction setting suggests the importance of capturing the diversity of ICU patients in mortality risk prediction. In addition, the performance improvements compared to the four MTL methods indicate that for patient-specific modeling, it is important to utilize the entire

Table 4.7: Comparison of averaged specificities. We performed a Wilcoxon signed-rank test for each pair of the method with the highest specificity and another method in each prediction setting. The result in bold, with the highest specificity, indicates that the method statistically significantly ($p < 0.05$) outperformed all the other nonboldface methods in the concerned setting. For prediction at ICU admission, Proposed-w/o-pop and STL (common) outperformed all other methods. For prediction at ICU discharge, MTL (DM) method outperformed all other methods.

Method	Specificity	
	admission point	discharge point
Proposed	0.73	0.77
Proposed-w/o-A	0.45	0.48
Proposed-w/o-pop	**0.79**	0.84
STL (separate)	0.45	0.48
STL (common)	**0.78**	0.87
MTL (Trace [53])	0.47	0.48
MTL (Mean [25])	0.47	0.50
MTL ($\ell_{2,1}$ [27])	0.47	0.48
MTL (DM [18])	0.78	**0.89**

Table 4.8: Example of top-10 predictive features for a disease-mortality-related latent task ($k = 1$).

High/Low risk	Example of category of predictive features	Example of predictive features
High	High-mortality diseases	J15, I46, C22, N18, I61
Low	Low-mortality diseases	F32, T14, E86, I74, R40

collection of diseases associated with each patient.

The recall, precision, and specificity results are shown in Tables 4.5, 4.6, and 4.7, respectively. In terms of recall, Proposed-w/o-A outperformed all other methods in both settings. In terms of precision, for prediction at ICU admission, STL (common) and MTL (DM) outperformed all other methods; for prediction at ICU discharge, MTL (DM) outperformed all other methods. In terms of specificity, for prediction at ICU admission point, Proposed-w/o-pop and STL (common) outperformed all other methods; for the ICU discharge prediction, MTL (DM) outperformed all other methods.

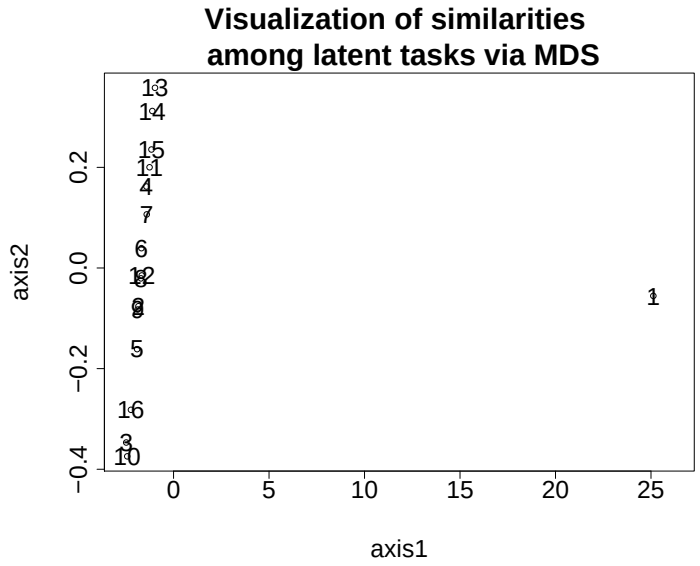

Fig. 4.3: Visualization of similarities among latent tasks via MDS using $\tilde{\boldsymbol{S}}$. Only one latent task ($k = 1$) was positioned as an outlier, while all others aligned with one dimension.

4.4.2.2 Latent task analysis

The number of latent tasks K depends on the sample trial, but the cases with $K = 16$ accounted for 80% of all cases; therefore, we adopted one sample for illustration purposes in which $K = 16$. We first examined the relationship among latent tasks in terms of associated diseases using $\tilde{\boldsymbol{S}}$. Specifically, each latent task k was associated with its disease vector $\tilde{\boldsymbol{S}}_{k,*}$ and multi-dimensional scaling (MDS) was applied to see the relationships among latent tasks. Figure 4.3 shows the result: only one latent task ($k = 1$) was positioned as an outlier, while all others aligned with one dimension.

Then, we examined predictive features for each latent task using $\boldsymbol{L}$. Specifically, we examined the top 10 features with positive and negative coefficients for each latent task by calculating the following two ratios: the ratio of high-mortality diseases in the top-10 predictive features with positive coefficients and the ratio of low-mortality diseases in the top-10 predictive features with negative coefficients, where high-mortality means mortality above the average and low-mortality means mortality below the average. Figures 4.4 and 4.5 show histograms of them: Figure 4.4 shows a histogram of the ratio of

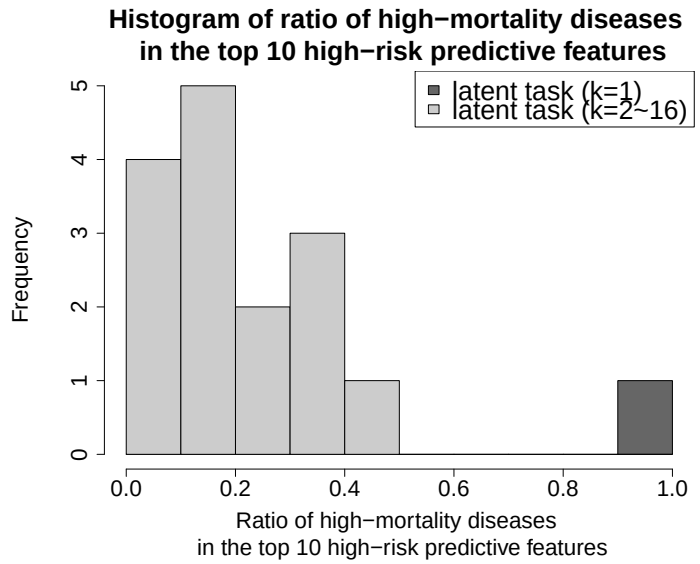

Fig. 4.4: Histogram of the ratio of high-mortality diseases in the top-10 high-risk predictive features. For a latent task ($k = 1$), the highest-risk predictive features are composed of high-mortality diseases, whereas this tendency is not observed for other latent tasks.

high-mortality diseases in the top-10 high-risk predictive features for each task and Figure 4.5 shows a histogram of the ratio of low-mortality diseases in the top-10 low-risk predictive features for each task. For a latent task ($k = 1$), the highest-risk predictive features are composed of high-mortality diseases, and the lowest-risk predictive features are composed of low-mortality diseases, whereas this tendency is not observed for other latent tasks. Together with the MDS result, one latent task ($k = 1$) plays a role in capturing disease-mortality, while others play different roles. Table 4.8 shows some examples of the top-10 high-risk features and the top-10 low-risk features for the disease-mortality-related task ($k = 1$). For the other latent tasks ($k = 2$–16), both high-risk and low-risk predictive features contained a specific combination of diseases for each task. Together with the above latent task ($k = 1$) analysis, it is possible that patient-specific models are constructed from viewpoints such as whether the patient is associated with high- or low-mortality diseases, and whether the patient has a specific combination of diseases.

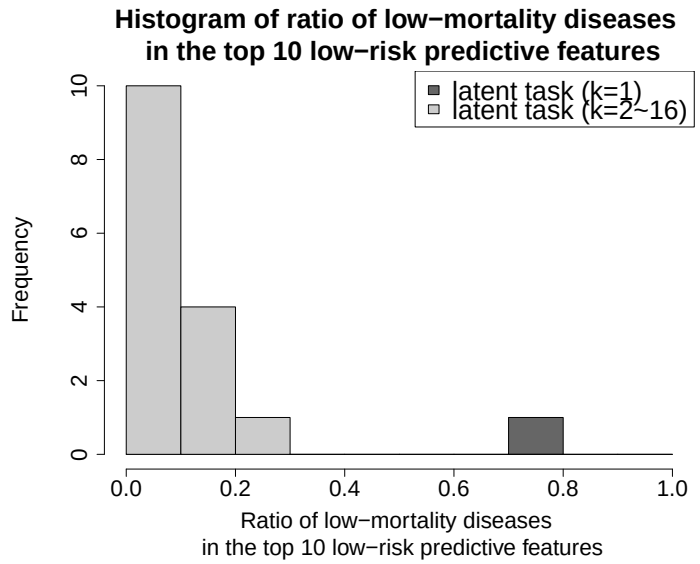

Fig. 4.5: Histogram of the ratio of low-mortality diseases in the top-10 low-risk predictive features. For a latent task ($k = 1$), the lowest-risk predictive features are composed of low-mortality diseases, whereas this tendency is not observed for other latent tasks.

4.5 Summary

In this chapter, we considered risk prediction associated with the mortality of diverse ICU patients by producing patient-specific risk models. Our proposed method could be considered an MTL method in which latent basis tasks are learned from the collection of diseases the patients are associated with. Our experimental results using a real-world hospital dataset demonstrated the effectiveness of our method by outperforming standard single-task learning methods and various MTL methods in which a task corresponds to a disease. Furthermore, our method could be used for uncovering patient-specificity from different viewpoints.

Chapter 5

Transfer Learning for Infrequent Diseases

5.1 Introduction

The vast amount of clinical data that is now stored as EMRs has provided us with an opportunity to improve the quality and efficiency of clinical care by learning from past patient data. Predictive risk modeling stands to constitute a basis for these studies, which could directly lead to clinical decision-making using, for example, medical alarms.

However, thus far, studies have not yet focused on predictive risk modeling for *infrequent* diseases, whose number of patients is relatively small over the entire population. In general, there may be only a few patients for one specific infrequent disease; however, the total number of patients with all types of infrequent diseases is nonnegligible. One example that illustrates this fact is the distribution of "rare diseases." In the US, approximately 7,000 diseases are recognized as rare diseases and the total number of patients with these diseases is estimated to be 25–30 million (about 8–12% of the population) as of 2009 [59], which equates to about 1 in 8–12 people.

Despite the importance of predictive risk modeling for infrequent diseases, the fundamental challenge in developing these models is the extremely limited amount of data available. Only a few examples of infrequent diseases might be available in the training phase, making it especially difficult to develop risk models for infrequent diseases. Moreover, medical domain knowledge is also quite limited for infrequent diseases. To address this data sparsity problem, in

this study, we explore the use of transfer learning in the context of in-hospital mortality risk modeling for ICU patients. This study is the first that examines the feasibility of transfer learning for developing predictive risk models for infrequent diseases.

This study makes the following contributions:

- We formulate in-hospital mortality risk prediction for infrequent diseases as an inductive transfer learning problem, in which all available data for all diseases are used as a source of information for learning a specific risk model for each infrequent disease.
- We evaluate the use of inductive transfer learning by applying it to in-hospital mortality risk prediction for ICU patients using a real-world dataset collected from a hospital in Japan.

5.2 Related Work

In this section, we first describe several studies of transfer learning and explain why we chose the method employed in our study. Then, we introduce studies of transfer learning for healthcare problems and describe our contributions in that context.

5.2.1 Transfer Learning

Transfer learning can be categorized into three types [34]: inductive, transductive, and unsupervised transfer learning. The method used in this chapter is categorized as inductive transfer learning, in which the source and target tasks are different, while the source and target domains may either be the same or different; further, labeled data are available in either the source or target domain.

Various methods applicable to inductive transfer learning have been proposed, including instance-transfer methods [37, 60, 61], feature-representation-transfer methods [27, 28, 62], and parameter-transfer methods [29, 63, 64]. Dai et al. [60] extended the boosting algorithm *AdaBoost* for inductive transfer learning. The resulting algorithm attempts to iteratively reweight examples in the source domain to discount the less relevant examples while encouraging more relevant source examples that contribute to the

target domain. For clinical risk modeling, an instance-weighting method was proposed [37] that weights examples in the source domain based on their similarity to the target training examples, where similarity is defined by the squared Euclidean distance in the feature space. Wu and Dietterich [61] integrated the source data into a support vector machine (SVM) framework. Argyriou et al. [27, 28] proposed methods to learn a low-dimensional representation shared across a set of multiple tasks. Similarly, a transfer learning method via dimensional reduction was proposed with the assumption that the target and source tasks are similar [62]. For parameter transfer, many studies have employed Gaussian process [29, 63, 64].

Of these various methods, we adopt the instance-weighting framework proposed by Bickel et al. [65], which can handle arbitrarily different data distributions for different tasks without making assumptions about the relationship between tasks or the data generation process. Consequently, we can learn the similarity among diseases and adopt a linear classifier that can improve the interpretability of the model, which is an important factor in clinical research.

5.2.2 Transfer Learning for Healthcare Problems

In clinical risk modeling research, Gong et al. [37] proposed an instance-transfer method to develop surgery- and hospital-specific risk models that weights examples from the source domain based on their similarity to training examples in the target domain. Jenna et al. [36] investigated the effectiveness of developing hospital-specific risk models via feature-representation transfer using auxiliary data from other hospitals; in their method, data in the source and target domains may lie in distinct but overlapping feature spaces. Lee et al. [35] explored the use of transfer learning to adapt a global model that is shared across multiple hospitals to a local hospital; they first trained a model on the source data and then learned a model for the target data by regularizing the model parameters toward those of the source data.

However, most studies have not adopted the disease as a task unit. A few exceptions are recent works on disease-specific risk modeling [18, 39]. Wang et al. explored the potential of joint risk prediction for different diseases by formulating disease onset risk prediction as a disease-based MTL [39]. In a previous study, we proposed an MTL method for mortality risk prediction for ICU patients in which a task corresponds to a disease [18].

These studies represent a promising new research avenue that may have implications for phenotyping research as well as personalized healthcare [49, 66, 67, 68]. Nonetheless, disease-specific risk modeling for infrequent diseases has not been adequately explored, although it is an important direction in which more research effort is required. Our work is the first to formulate risk prediction for infrequent diseases by transfer learning in the context of mortality modeling for ICU patients and to investigate the effectiveness of this approach via empirical study.

5.3 Disease-specific Risk Modeling via Distribution Match

We explore the potential of a transfer learning approach based on distribution matching [65] to learn disease-specific risk models for infrequent diseases. This approach, proposed by Bickel et al., can handle arbitrarily different data distributions for different tasks without making assumptions about the relationship between tasks or the data generation process, enabling us to learn the similarity between diseases and adopt a linear classifier that can improve the interpretability of the resulting model.

5.3.1 Problem Definition

Let us assume that each disease z is associated with an unknown joint distribution $p(\mathbf{x}, y|z)$ of patient features $\mathbf{x}$ and patient outcome y. The training sample $D = \langle(\mathbf{x}_1, y_1, z_1), ..., (\mathbf{x}_n, y_n, z_n)\rangle$ comprises examples from all diseases. We assume that for each example indexed with i, the input attributes $\mathbf{x}_i$, an M-dimensional feature vector representing the patient; class label $y_i \in \{0, 1\}$, which is the patient outcome, where 0 and 1 represent survival and death in the hospital, respectively; and the originating task $z_i \in \{1, \ldots, T\}$, the disease the patient has, are known. The entire sample D is governed by the mixed joint density $p(z)p(\mathbf{x}, y|z)$. Prior $p(z)$ specifies the proportion of disease z.

Our goal is to learn a hypothesis $f_z : \mathbf{x} \mapsto y$ for each disease z. Hypothesis $f_z(\mathbf{x})$ must correctly predict the true label y of unseen examples drawn from

$p(\mathbf{x}|z)$ for all z. In other words, it should minimize the expected loss

$$\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|z)}[\ell(f_z(\mathbf{x}),y)] \tag{5.1}$$

with respect to the unknown joint distribution for each disease z.

5.3.2 Approach

If we pool all available data for all diseases, we obtain a sample governed by $\sum_z p(z)p(\mathbf{x},y|z)$. Let us assume a disease-specific resampling weight $r_t(\mathbf{x},y)$ for each element of the pool of examples for target disease t. The resampling weight is introduced to match the pool to the target distribution $p(\mathbf{x},y|t)$, which can be defined as follows:

$$\begin{aligned}&\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|t)}[\ell(f(\mathbf{x},t),y)] \\ &= \mathbf{E}_{(\mathbf{x},y)\sim \sum_z p(z)p(\mathbf{x},y|z)}[r_t(\mathbf{x},y)\ell(f(\mathbf{x},t),y)].\end{aligned} \tag{5.2}$$

It has been shown that the resampling weight can be described as [65]

$$r_t(\mathbf{x},y) = \frac{p(t|\mathbf{x},y)}{p(t)}. \tag{5.3}$$

The advantage of this expression is that the resampling weight can be decided without knowing any of the disease densities $p(\mathbf{x},y|z)$, which are potentially high-dimensional. Instead, we need to model $p(t|\mathbf{x},y)$, which discriminates between the labeled instances of the target disease and labeled instances of all pooled examples of all diseases. Intuitively, it characterizes how much more likely example $(\mathbf{x},y)$ is to occur in the target disease distribution than in the mixture distribution of all diseases. We can apply any type of probabilistic classifier to this problem. We model $p(t|\mathbf{x},y)$ for all diseases simultaneously using a softmax model (multiclass generalization of a logistic model) with model parameter $\mathbf{v}$, as defined in Equation (5.4); parameter vector $\mathbf{v}$ is a concatenation of the task-specific subvectors $\mathbf{v}_z$ for each task z.

$$p(z|\mathbf{x},y,\mathbf{v}) = \frac{\exp(\mathbf{v}_z^\top \boldsymbol{\phi}(\mathbf{x},y))}{\sum_{z'}\exp(\mathbf{v}_{z'}{}^\top \boldsymbol{\phi}(\mathbf{x},y))}. \tag{5.4}$$

Equation (5.4) requires problem-specific feature mapping $\boldsymbol{\phi}(\mathbf{x},y)$. We de-

fine this mapping using the Kronecker δ as follows:

$$\boldsymbol{\phi}(\mathbf{x}, y) = \begin{bmatrix} \delta(y, +1)\boldsymbol{\phi}(\mathbf{x}) \\ \delta(y, -1)\boldsymbol{\phi}(\mathbf{x}) \end{bmatrix}. \tag{5.5}$$

With L2 regularization, we solve the following optimization problem:

$$\min_{\mathbf{v}} \lambda \mathbf{v}^\top \mathbf{v} - \sum_{(\mathbf{x}_i, y_i, z_i) \in D} \log\left(p(z_i | \mathbf{x}_i, y_i, \mathbf{v})\right), \tag{5.6}$$

where $\lambda \geq 0$ is a hyperparameter used to tune the regularization weight.

After obtaining the probabilistic model $p(t|\mathbf{x}, y)$, we can calculate the resampling weight using Equation (5.3). Then, we can minimize the expected loss in Equation (5.2) by evaluating the expected loss over the weighted training data. In this study, we adopt L2-regularized logistic regression to train an array of target models.

Our approach can therefore be summarized as follows:

Step 1. Solve the optimization problem (5.6) to train a classifier that models $p(z|\mathbf{x}, y)$ as defined in Equation (5.4) using a softmax model (multiclass generalization of a logistic model).

Step 2. Calculate the resampling weight using Equation (5.3) for each training instance in each target disease.

Step 3. Train an array of target models by L2-regularized logistic regression using the resampling weight obtained in Step 2.

5.4 Empirical Study

5.4.1 Experimental Setup

5.4.1.1 Dataset

We used a dataset from a hospital collected as part of the Quality Indicator/Improvement Project (QIP) [52] administered by the Department of Healthcare Economics and Quality Management at Kyoto University in Japan. The statistics of this dataset are summarized in Table 5.1. The patients were at least 18 years old and underwent ICU treatment at some point in their hospital stay. After excluding patients with a disease whose initial letter

in the ICD-10 code is after N, the total number of patients was 721. As the unit of the task, we adopted the main disease that caused the patient's hospital admission, as identified by a 3-digit ICD-10 code. The total number of diseases was 47. We adopted in-hospital mortality to evaluate patient outcome: if a patient died during his or her hospital stay, the patient outcome was "death"; otherwise, it was "survival".

For features describing the patients, we extracted the patient's age, gender, main disease, and all comorbidities as described by the 4-digit ICD-10 code, in addition to all medical events, such as medication, surgery, and laboratory tests recorded in the billing process during their hospital stay. For these interventions, we did not conduct any feature engineering; we used the raw codes in the DPC system as the features. If a patient received an intervention one or more times, we set the corresponding feature value to 1; otherwise, it was set to 0. Weighting features based on the number of interventions was found to deteriorate prediction performance in a preliminary experiment.

Table 5.1: Statistics of the dataset.

Data	Number
Patients	721
Diseases	47
Target diseases	22
Patients with the target diseases	631
Features	1,381

5.4.1.2 Prediction setting

We randomly sampled 60% of the patients to create the training dataset and used the remaining 40% of the patients for evaluation. We used all features associated with the patient that were available 1 day after the day he/she was admitted to the ICU, for a total of 1,381 features. Our prediction setting is *1 day after ICU admission*; that is, using only data available 1 day after ICU admission, we predict the patient's outcome.

The hyperparameter λ in Equation (5.6) was set to 10^0. For the MTL with medical domain knowledge (MTL-DM) model, regularization hyperparameters for diseases, features, and ℓ_2-norm were set to 10^0, 10^{-4}, and 10^{-4}, respectively. All the other L2-regularization hyperparameters were tuned from among $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0\}$ by 3-fold cross-validation in the train-

Table 5.2: List of diseases, along with category, based on the patient population of the disease.

# of patients	# of diseases	Disease list ICD-10 code (# of patients)
1–9	35	A41 (2), C38 (1), C80 (1), E11 (1), E85 (1), E86 (1), E87 (3), H25 (1), I05 (2), I08 (1), I11 (1), I23 (3), I25 (3), I27 (3), I30 (2), I31 (4), I33 (3), I42 (6), I46 (5), I63 (2), I65 (3), I70 (3), I73 (1), I74 (3), I80 (1), I82 (1), J15 (3), J18 (3), J98 (1), K43 (1), K56 (1), K80 (1), M16 (1), N10 (1), N18 (2)
10–99	9	I24 (17), I26 (11), I34 (13), I35 (15), I44 (14), I47 (12), I48 (10), I49 (16), I71 (80)
100–1000	3	I20 (116), I21 (173), I50 (172)

ing dataset. We set prior $p(t)$ in Equation (5.3) to $p(t) \equiv \frac{|D_t|+\gamma}{\sum_z(|D_z+\gamma|)}$, to address the zero-frequency problem; we set γ to 10^{-6} throughout our experiment.

We repeated the sampling, prediction, and evaluation procedures 100 times and calculated the mean of these trials. We then conducted a Wilcoxon signed-rank test for each pair of our transfer learning method and the comparison method to evaluate the statistical significance of any differences.

5.4.1.3 Evaluation Measure

There is a difficulty in calculating AUCs for infrequent diseases that have few positive and negative examples. Therefore, we focus on other measures in this section. We consider triggering medical alarms as our application scenario; thus, our aim is to correctly identify high-risk patients. To avoid overlooking high-risk patients, *recall* is most crucial. However, false alarms inappropriately consume limited clinical resource, and therefore, *precision* and *specificity* are also important. We therefore adopt the three measures described above for evaluation.

To evaluate the former two predictive performances for a disease, the disease has to include one or more "positive" instances. We identified 22 *target diseases* to be evaluated. Tables 5.2 and 5.3 list the diseases and target diseases, respectively, along with the disease category based on patient population.

Table 5.3: List of diseases used as target diseases, along with category, based on the patient population of the disease.

# of patients	# of diseases	Disease list ICD-10 code (# of patients)
1–9	13	A41 (2), C80 (1), E85 (1), E86 (1), I08 (1), I23 (3), I25 (3), I63 (2), I74 (3), J15 (3), J98 (1), K43 (1), K56 (1)
10–99	6	I26 (11), I35 (15), I44 (14), I47 (12), I49 (16), I71 (80)
100–1000	3	I20 (116), I21 (173), I50 (172)

5.4.1.4 Comparison methods

We compared our transfer learning method, *transfer model*, with three methods based on L2-regularized logistic regression. We compared L2-regularized logistic regression with L1-regularized logistic regression and found that L2-regularization often resulted in better predictive performance in all the methods in our preliminary experiment. Therefore, we adopted L2-regularization instead of L1-regularization in our study.

The first method is a one-size-fits-all model, or *common model*, which is one of the most standard methods for mortality prediction. This method develops a generic model that is shared among all diseases using all data.

The second method is a *separate model*, which develops a disease-specific model for a disease using only data of the target disease. In the separate model, there is no training data for diseases with only 1 patient; therefore, we consider only diseases with more than 1 patient as target diseases when comparing the results with the separate model.

The third method is an *MTL-DM model* that constructs a disease-specific MTL model with medical domain knowledge [18], which is integrated via cross-regularization by two graph Laplacians that encode similarities between diseases and among EMRs, respectively. For fine-tuning this method to create similarity among diseases, we followed the settings described in [18]. Specifically, the similarity of two diseases was defined as the number of shared levels in the ICD hierarchy and the similarity of two EMRs was defined as 1 if the EMRs describe medicine with common efficacy; otherwise, it is 0. Because several conventional MTL methods have been compared with the MTL-DM model and shown to be less effective on a dataset of the type that we have

used in this study, we adopted MTL-DM as a representative MTL method.

5.4.2 Results and Discussion

In this section, we report our experimental results. Our transfer learning method improved both recall and precision for most diseases, especially infrequent ones, when compared to several baselines, including one of the most standard methods, L2-regularized logistic regression, and multitask learning with medical domain knowledge. However, the comparison methods showed better specificity for most diseases.

5.4.2.1 Predictive performance

To evaluate the statistical significance of the mean recall, precision, and specificity, we conducted a paired Wilcoxon signed-rank test for each pair of our transfer learning method and the comparison method in each setting and identified results with statistically significant ($p < 0.05$) differences, which are summarized in Tables 5.4, 5.5, and 5.6. We provide only results with statistical significance ($p < 0.05$) here; therefore, the number of diseases is different in each setting.

Overall results.

Tables 5.4, 5.5, and 5.6 compare the recall, precision, and specificity, for each pair of our transfer model and comparison method. For most diseases, both recall and precision significantly improved with the use of our transfer method. However, comparison methods showed better specificity for most diseases. In considering triggering medical alarms as our application scenario, improving recall and precision would have significant implications, though improving specificity is important future work.

Comparison to the common model.

Tables 5.4 (a), 5.5 (a), and 5.6 (a) compare the recall, precision, and specificity, respectively, of our transfer model with the common model for each disease. Performance improved for 8 (of 10), i.e., $\frac{8}{10} \simeq 80\%$, and 7 (of 8), i.e., $\frac{7}{8} \simeq 88\%$, of diseases, in terms of recall and precision, respectively. However, the common model showed better specificity for most diseases (8 of 11), i.e., $\frac{8}{11} \simeq 73\%$ diseases.

Among the diseases with statistically significant differences, both recall and precision improved for the most infrequent diseases, with less than 10

patients (7 of 7 and 6 of 6, respectively). Of the 8 diseases whose recall improved with our model, 7 diseases, i.e., $\frac{7}{8} \simeq 88\%$, had less than 10 patients. In contrast, recall deteriorated for 2 diseases, whose numbers of patients were 172 and 173; these diseases are rather *frequent diseases*. Similarly, precision improved for 7 diseases, of which 6 diseases, i.e., $\frac{6}{7} \simeq 86\%$, had less than 10 patients. Likewise, precision deteriorated for 1 disease, whose number of patients was 173, again making it a rather *frequent* disease.

Comparison to the separate model.

Tables 5.4 (b), 5.5 (b), and 5.6 (b) compare the recall, precision, and specificity, respectively, of our transfer model with the separate model for each disease. Recall improved for 11 (of 12), i.e., $\frac{11}{12} \simeq 92\%$, of diseases and precision improved for 11 (of 14), i.e., $\frac{11}{14} \simeq 79\%$, of diseases. However, the separate model showed better specificity for most diseases (9 of 12), i.e., $\frac{9}{12} \simeq 75\%$, of diseases.

Among the diseases with statistically significant differences, both recall and precision improved for all the most infrequent diseases, which had less than 10 patients (6 of 6). Of the 11 diseases whose recall was improved with our model, 6 diseases, i.e., $\frac{6}{11} \simeq 55\%$, had less than 10 patients. In contrast, recall deteriorated for 1 disease, whose number of patients was 173; this disease is a rather *frequent disease*. Precision improved for 11 diseases, of which 6 diseases, i.e., $\frac{6}{11} \simeq 55\%$, were most infrequent diseases, having less than 10 patients. Likewise, precision deteriorated for 3 diseases, whose numbers of patients were 80, 172, and 173, again making them rather *frequent* diseases.

Comparison to the MTL-DM model.

Tables 5.4 (c), 5.5 (c), and 5.6 (c) compare the recall, precision, and specificity, respectively, of our transfer model and the MTL-DM model for each disease. For recall, performance improved for 11 (of 14), i.e., $\frac{11}{14} \simeq 79\%$, of diseases. For precision, performance improved for 11 (of 15), i.e., $\frac{11}{15} \simeq 73\%$, of diseases. However, the MTL-DM model showed better specificity for all diseases (12 of 12).

Among the diseases with statistically significant differences, both recall and precision improved for almost all of the most infrequent diseases, which had less than 10 patients (6 of 7), i.e., $\frac{6}{7} \simeq 86\%$. Of the 11 diseases for which both recall and precision improved, 6 diseases, i.e., $\frac{6}{11} \simeq 55\%$, were most infrequent diseases, with less than 10 patients. In contrast, recall deteriorated

for 3 diseases, of which 2 diseases were rather *frequent diseases*, whose numbers of patients were 172 and 173. Likewise, precision deteriorated for 4 diseases, of which 3 diseases were rather *frequent diseases*, whose numbers of patients were 80, 172 and 173.

Several observations and suggestions.

It is also notable that, in several cases, both the recall and precision of our transfer model improved when the comparison model completely failed in prediction (that is, has a predictive performance of "0"). Also note that the transfer model does not rely on medical domain knowledge, whereas the MTL-DM model exploits such knowledge, suggesting that the data-driven relationships between diseases can enhance predictive power in a manner that is complementary to existing medical domain knowledge. However, for some diseases, predictive performance, especially specificity, deteriorated when using the transfer model. Avoiding such "negative transfer" is still an open question in the field of transfer learning. In deploying our method in a real-world scenario, we need to detect and adequately manage negative transfer or we need to exploit only positive transfer; we might achieve this by combining transfer learning with a complementary classifier, such as the MTL-DM model used in our study.

Accuracy of learning before learning.

To make a transfer learning method work effectively, the source and target data should be related in some way. In our method, disease classification accuracy, which corresponds to the accuracy of the estimate in Equation (5.4) is a crucial factor for successful transfer learning. To evaluate this accuracy, Figure 5.1 shows the disease classification accuracy for each disease; the results shown are the means of the 100 trials. Since there are no training or test data for diseases whose number of patients is only 1, we excluded these diseases here. We observe that, in many diseases, classification accuracy is high compared to random prediction (an accuracy of 0.5). More specifically, about 73 % of diseases achieve accuracy of more than 0.5. It might be possible that our transfer learning approach fails in datasets with poor classification accuracy; in deploying our method in a real-world scenario, we may first evaluate classification accuracy and estimate transfer learning performance.

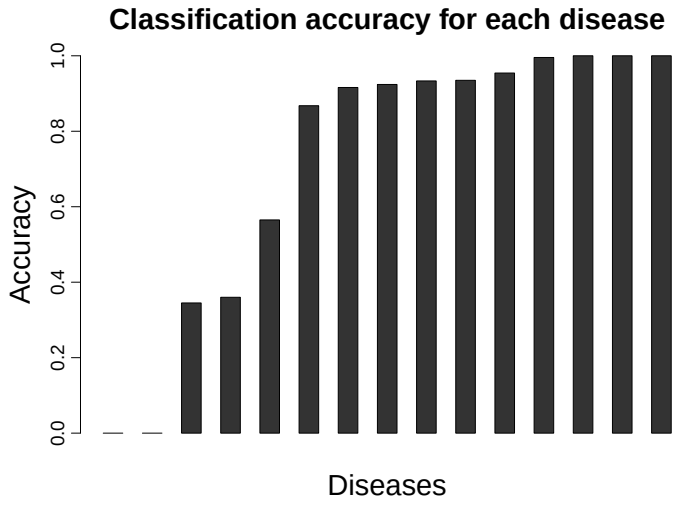

Fig. 5.1: Classification accuracy, i.e., the accuracy of the estimate of Equation (5.4), of each disease. We observe that for many diseases, classification accuracy is high compared to that of random prediction (accuracy of 0.5).

5.5 Summary

The ability to accurately predict a patient's mortality risk has an immediate practical use for ICU clinicians. The goal of this chapter was to explore the feasibility of transfer learning for disease-specific risk modeling, especially for infrequent diseases in the ICU. Our experiment demonstrated some promising results for applications of transfer learning for risk modeling of infrequent diseases. The results showed that by transferring knowledge from other diseases, both recall and precision improved for many diseases, particularly infrequent ones, when compared to several baselines, the standard method, L2-regularized logistic regression, and MTL with medical domain knowledge, although specificity deteriorated for most diseases.

Table 5.4: Comparison of recall for each pair of the transfer model and common, separate, and MTL-DM model, respectively, for each disease. For each pair of transfer and comparison model in each disease, statistically significant ($p < 0.05$) differences were confirmed via a Wilcoxon signed-rank test. For most diseases, recall significantly improved with the use of our transfer method.

(a) Recall of the transfer and common models for each disease.

	# of patients of the disease	Recall (Transfer)	Recall (Common)
Deteriorated	172	0.07	0.16
	173	0.06	0.19
Improved	1	0.42	0.29
	1	0.28	0.18
	1	0.49	0.40
	1	0.84	0.75
	1	0.20	0.13
	3	0.21	0.14
	3	0.28	0.22
	16	0.09	0.02

(b) Recall of the transfer and separate models for each disease.

	# of patients of the disease	Recall (Transfer)	Recall (Separate)
Deteriorated	173	0.06	0.30
Improved	2	0.12	0.00
	2	0.14	0.00
	3	0.07	0.00
	3	0.21	0.00
	3	0.26	0.00
	3	0.28	0.00
	11	0.15	0.00
	12	0.08	0.00
	14	0.08	0.00
	16	0.09	0.00
	116	0.05	0.00

(c) Recall of the transfer and MTL-DM models for each disease.

	# of patients of the disease	Recall (Transfer)	Recall (MTL-DM)
Deteriorated	1	0.84	1.00
	172	0.07	0.17
	173	0.06	0.35
Improved	1	0.49	0.32
	1	0.28	0.05
	2	0.12	0.04
	2	0.14	0.00
	3	0.28	0.08
	3	0.26	0.00
	11	0.15	0.01
	12	0.08	0.00
	14	0.08	0.00
	16	0.09	0.01
	116	0.05	0.01

Table 5.5: Comparison of precision for each pair of the transfer model and common, separate, and MTL-DM model, respectively, for each disease. For each pair of transfer and comparison model in each disease, statistically significant ($p < 0.05$) differences were confirmed via a Wilcoxon signed-rank test. For most diseases, precision significantly improved with the use of our transfer method.

(a) Precision of the transfer and common models for each disease.

	# of patients of the disease	Precision (Transfer)	Precision (Common)
Deteriorated	173	0.19	0.41
Improved	1	0.20	0.13
	1	0.49	0.40
	1	0.84	0.75
	1	0.28	0.18
	1	0.42	0.29
	3	0.20	0.14
	16	0.07	0.01

(b) Precision of the transfer and separate models for each disease.

	# of patients of the disease	Precision (Transfer)	Precision (Separate)
Deteriorated	80	0.42	0.62
	172	0.17	0.35
	173	0.19	0.72
Improved	2	0.12	0.00
	2	0.14	0.00
	3	0.06	0.00
	3	0.18	0.00
	3	0.19	0.00
	3	0.20	0.00
	11	0.05	0.00
	12	0.03	0.00
	14	0.02	0.00
	16	0.07	0.00
	116	0.03	0.00

(c) Precision of the transfer and MTL-DM models for each disease.

	# of patients of the disease	Precision (Transfer)	Precision (MTL-DM)
Deteriorated	1	0.84	1.00
	80	0.42	0.54
	172	0.17	0.28
	173	0.19	0.75
Improved	1	0.49	0.32
	1	0.28	0.05
	2	0.12	0.04
	2	0.14	0.00
	3	0.19	0.08
	3	0.18	0.00
	11	0.05	0.01
	12	0.03	0.00
	14	0.02	0.00
	16	0.07	0.01
	116	0.03	0.01

Table 5.6: Comparison of specificity for each pair of the transfer model and common, separate, and MTL-DM model, respectively, for each disease. For each pair of the transfer and comparison model in each disease, statistically significant ($p < 0.05$) differences were confirmed via a Wilcoxon signed-rank test. For most diseases, comparison model showed significantly better specificity than the transfer model.

(a) Specificity of the transfer and common models for each disease.

	# of patients of the disease	Specificity (Transfer)	Specificity (Common)
Deteriorated	2	0.19	0.28
	3	0.46	0.57
	3	0.70	0.79
	11	0.67	0.70
	12	0.81	0.85
	14	0.83	0.86
	15	0.88	0.93
	16	0.83	0.86
Improved	116	0.94	0.90
	172	0.95	0.89
	173	0.97	0.91

(b) Specificity of the transfer and separate models for each disease.

	# of patients of the disease	Specificity (Transfer)	Specificity (Separate)
Deteriorated	3	0.46	0.69
	11	0.67	1.00
	12	0.81	1.00
	14	0.83	1.00
	15	0.88	0.95
	16	0.83	1.00
	80	0.89	0.96
	116	0.94	1.00
	172	0.95	0.99
Improved	2	0.19	0.00
	2	0.34	0.00
	3	0.85	0.67

(c) Specificity of the transfer and MTL-DM models for each disease.

	# of patients of the disease	Specificity (Transfer)	Specificity (MTL-DM)
Deteriorated	2	0.34	0.51
	3	0.46	0.84
	3	0.75	0.98
	3	0.70	0.88
	3	0.85	1.00
	11	0.67	0.91
	12	0.81	0.96
	14	0.83	0.98
	15	0.88	0.98
	16	0.83	1.00
	80	0.89	0.95
	116	0.94	0.99

Chapter 6

Conclusion

The aim of the research completed for this thesis is to address the diversity of ICU patients for clinical risk modeling.

We first considered the disease-specific characteristics of ICU patients by formulating risk prediction as a multitask learning problem in which a task corresponds to a disease. To mitigate the problem of data paucity resulting from disease-based personalization and data sparseness associated with electronic medical records, we exploited domain knowledge relating to the similarity between diseases and among electronic medical records as an inductive bias. Our multitask learning method with medical domain knowledge outperformed several baselines for ICU mortality prediction, including the de facto single-task learning method and several multitask learning methods without domain knowledge. Our method produces disease-specific models that enables us to investigate specificity for each disease and the relationships among diseases.

Then, we presented a multitask learning method in which latent tasks are learned based on the collection of diseases associated with each patient. The proposed patient-specific risk modeling method showed higher predictive performance compared with several standard methods for mortality risk prediction of ICU patients. In addition, the proposed method is capable of uncovering patient-specificity from several viewpoints.

Lastly, we further addressed the data paucity problem for infrequent diseases, whose occurrence is relatively rare in the entire population, by inductive transfer learning. The feasibility study corroborated several promising results.

The significance of developing disease- and patient-specific models lies in their potential to enhance the understanding of disease- and patient-specificity and inherent relationships among diseases by providing disease- and patient-specific models, as well as the potential improvement of prediction performance. For example, disease-specific predictive high-risk factors or predictive latent tasks might provide us with suggestive hypotheses that could be validated by further investigation in the medical domain. In such a context, our research could also have implications for phenotyping research, whose aim is to identify patient groups or features that match some research criteria; the advancement of precision and personalized medicine would considerably depend on the success of inferring phenotypic patterns from large amount of electronic medical records. Another implication would be related to systematization of diseases. How to define, categorize, and organize diseases is a nontrivial matter; in medicine, the definition of a disease occasionally changes as new findings arise. Unveiling underlying relationships among diseases that are learned from data could provide domain experts with different viewpoints for understanding the ontology of diseases. Overall, personalized clinical risk modeling as provided in this thesis could help clinical experts enhance their understanding of diseases and patients.

We hope this study serves as a foundation for personalized clinical risk modeling for ICU patients.

Publications

Journals

(1) 則 のぞみ, 鹿島 久嗣, 山下 和人, 猪飼 宏, 今中 雄一. マルチタスク学習による集中治療室入室患者のリスクモデル構築. 電子情報通信学会論文誌和文D, Vol.J100-D, No.2, 194-204, 2017.

Referred International Conferences

(2) Nozomi Nori, Hisashi Kashima, Kazuto Yamashita, Susumu Kunisawa, Yuichi Imanaka. Learning Implicit Tasks for Patient-Specific Risk Modeling in ICU. Proceedings of the 31st National Conference on Artificial Intelligence (AAAI). 2017.

(3) Nozomi Nori, Hisashi Kashima, Kazuto Yamashita, Hiroshi Ikai, Yuichi Imanaka. Simultaneous Modeling of Multiple Diseases for Mortality Prediction in Acute Hospital Care. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 2015.

Domestic Conferences

(4) 則 のぞみ, 鹿島 久嗣, 山下 和人, 猪飼 宏, 今中 雄一. 疾病間での転移学習を用いたICU入室患者のリスク予測. 2016年度人工知能学会全国大会 (第30回), 2016.

(5) 則 のぞみ, 鹿島 久嗣, 山下 和人, 猪飼 宏, 今中 雄一. マルチタスク学習に基づく疾病コンテキストを考慮したICU入室患者の死亡リスク予測. 2015年度人工知能学会全国大会 (第29回), 2015.

References

[1] Marzyeh Ghassemi, Leo Anthony Celi, and David J Stone. State of the art review: the data revolution in critical care. *Critical Care*, Vol. 19, p. 118, 2015.

[2] Leo Anthony Celi, Marie Csete, and David Stone. Optimal data systems: the future of clinical predictions and decision support. *Current opinion in critical care*, Vol. 20, No. 5, pp. 573–580, 2017.

[3] Clemens Scott Kruse, Rishi Goswamy, Yesha Raval, and Sarah Marawi. Challenges and opportunities of big data in health care: A systematic review. *JMIR Medical Informatics*, Vol. 4, No. 4, p. e38, 2016.

[4] Charles C. Miller, Michael J. Reardon, and Hazim J. Safi. *Risk Stratification: A Practical Guide for Clinicians*. Cambridge University Press, 2001.

[5] George C.M. Siontis, Ioanna Tzoulaki, and John P.A. Ioannidis. Predicting death: an empirical evaluation of predictive tools for mortality. *Archives of Internal Medicine*, Vol. 171, No. 19, pp. 1721–1726, 2011.

[6] Xiongcai Cai, Oscar Perez-Concha, Enrico Coiera, Fernando Martin-Sanchez, Richard Day, David Roffe, and Blanca Gallego. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *Journal of the American Medical Informatics Association*, pp. 553–561, 2015.

[7] Danning He, Simon C Mathews, Anthony N Kalloo, and Susan Hutfless. Mining high-dimensional administrative claims data to predict early hospital readmissions. *Journal of the American Medical Informatics Association*, Vol. 21, No. 2, pp. 272–279, 2014.

[8] Brant W Chee, Richard Berlin, and Bruce Schatz. Predicting adverse drug events from personal health messages. In *AMIA Annual Symposium Proceedings*, pp. 217–226, 2011.

[9] Ying P Tabak, Xiaowu Sun, Carlos M Nunez, and Richard S Johannes.

Using electronic health record data to develop inpatient mortality predictive model: Acute laboratory risk of mortality score (ALaRMS). *Journal of the American Medical Informatics Association*, Vol. 21, No. 3, pp. 455–463, 2014.

[10] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 75–84, 2014.

[11] Marzyeh Ghassemi, Marco A.F. Pimentel, Tristan Naumann, Thomas Brennan, David A. Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 446–453, 2015.

[12] Yuan Luo, Yu Xin, Rohit Joshi, Leo Celi, and Peter Szolovits. Predicting ICU mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 42–50, 2016.

[13] Michael J. Breslow and Omar Badawi. Severity scoring in the critically ill: Part 1—interpretation and accuracy of outcome prediction scoring systems. *Chest*, Vol. 141, No. 1, pp. 245–252, 2012.

[14] Bekele Afessa, Mark T. Keegan, Rolf D. Hubmayr, James M. Naessens, Ognjen Gajic, Kirsten Hall Long, and Steve G. Peters. Evaluating the performance of an institution using an intensive care unit benchmark. *Mayo Clinic Proceedings*, Vol. 80, No. 2, pp. 174–180, 2005.

[15] Harlan M. Krumholz. Mathematical models and the assessment of performance in cardiology. *Circulation*, Vol. 99, No. 16, pp. 2067–2069, 1999.

[16] Joan Ivanov, Jack V. Tu, and C. David Naylor. Ready-made, recalibrated, or remodeled? *Circulation*, Vol. 99, No. 16, pp. 2098–2104, 1999.

[17] Otto Pitkänen, Minna Niskanen, Sinikka Rehnberg, Mikko Hippelinen, and Markku Hynynen. Intra-institutional prediction of outcome after cardiac surgery: comparison between a locally derived model and the EuroSCORE. *European Journal of Cardio-Thoracic Surgery*, Vol. 18,

No. 6, pp. 703–710, 2000.

[18] Nozomi Nori, Hisashi Kashima, Kazuto Yamashita, Hiroshi Ikai, and Yuichi Imanaka. Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, 2015.

[19] 則のぞみ, 鹿島久嗣, 山下和人, 猪飼宏, 今中雄一. マルチタスク学習による集中治療室入室患者のリスクモデル構築. 電子情報通信学会論文誌和文 D, Vol. J100-D, No. 2, pp. 194–204, 2017.

[20] Nozomi Nori, Hisashi Kashima, Susumu Kunisawa, Hiroshi Ikai, and Yuichi Imanaka. Learning implicit tasks for patient-specific risk modeling in ICU. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

[21] Caleb W Hug and Peter Szolovits. ICU Acuity: Real-time models versus daily models. In *AMIA Annual Symposium Proceedings*, pp. 260–264, 2009.

[22] Rohit Joshi and Peter Szolovits. Prognostic physiology: modeling patient severity in intensive care units using radial domain folding. In *AMIA Annual Symposium Proceedings*, pp. 1276–1283, 2012.

[23] Li-wei Lehman, Mohammed Saeed, William Long, Joon Lee, and Roger Mark. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. In *AMIA Annual Symposium Proceedings*, pp. 505–511, 2012.

[24] Rich Caruana. Multitask learning. *Machine Learning*, Vol. 28, No. 1, pp. 41–75, 1997.

[25] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 109–117, 2004.

[26] Kai Yu, Volker Tresp, and Anton Schwaighofer. Learning gaussian processes from multiple tasks. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 1012–1019, 2005.

[27] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pp. 41–48, 2006.

[28] Andreas Argyriou, Charles A. Micchelli, Massimiliano Pontil, and Yim-

ing Ying. A spectral regularization framework for multi-task structure learning. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 25–32, 2007.

[29] Edwin V. Bonilla, Kian Ming A. Chai, and Christopher K. I. Williams. Multi-task gaussian process prediction. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 153–160, 2007.

[30] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, Vol. 73, No. 3, pp. 243–272, 2008.

[31] Laurent Jacob, Francis Bach, and Jean-Philippe Vert. Clustered multi-task learning: A convex formulation. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pp. 745–752, 2008.

[32] Abhishek Kumar and Hal Daumé III. Learning task grouping and overlap in multi-task learning. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[33] Yu Zhang and Dit-Yan Yeung. A regularization approach to learning task relationships in multitask learning. *ACM Transactions on Knowledge Discovery from Data*, Vol. 8, No. 3, pp. 12:1–12:31, 2014.

[34] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 10, pp. 1345–1359, 2010.

[35] Gyemin Lee, Ilan Rubinfeld, and Zeeshan Syed. Adapting surgical models to individual hospitals using transfer learning. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops*, pp. 57–63, 2012.

[36] Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, Vol. 21, No. 4, pp. 699–706, 2014.

[37] Jen J. Gong, Thoralf M. Sundt, James D. Rawn, and John V. Guttag. Instance weighting for patient-specific risk stratification models. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 369–378, 2015.

[38] Sunil Kumar Gupta, Santu Rana, Dinh Q. Phung, and Svetha Venkatesh.

Keeping up with innovation: A predictive framework for modeling healthcare data with evolving clinical interventions. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 235–243, 2014.

[39] Xiang Wang, Fei Wang, Jianying Hu, and Robert Sorrentino. Exploring joint disease risk prediction. In *AMIA Annual Symposium Proceedings*, pp. 1180–1187, 2014.

[40] Helen McGloin, Sheila K Adam, and Mervyn Singer. Unexpected deaths and referrals to intensive care of patients on general wards. Are some cases potentially avoidable? *Journal of the Royal College of Physicians of London*, Vol. 33, No. 3, pp. 255–259, 1999.

[41] Hatem Ksouri, Per-Yann Balanant, Jean-Marc Tadié, Guillaume Heraud, Imad Abboud, Nicolas Lerolle, Ana Novara, Jean-Yves Fagon, and Christophe Faisy. Impact of morbidity and mortality conferences on analysis of mortality and critical events in intensive care practice. *American Journal of Respiratory and Critical Care Medicine*, Vol. 19, No. 2, pp. 135–145, 2010.

[42] Amine Ali Zeggwagh, Houda Mouad, Tarek Dendane, Khalid Abidi, Jihane Belayachi, Naoufel Madani, and Redouane Abouqal. Preventability of death in a medical intensive care unit at a university hospital in a developing country. *Indian Journal of Critical Care Medicine*, Vol. 18, No. 2, pp. 88–94, 2014.

[43] Yun Chen and Hui Yang. Heterogeneous postsurgical data analytics for predictive modeling of mortality risks in intensive care units. In *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4310–4314, 2014.

[44] Karla L. Caballero Barajas and Ram Akella. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 69–78, 2015.

[45] Blanca E. Himes, Yi Dai, , Isaac S. Kohane, Scott T. Weiss, and Marco F. Ramoni. Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. *Journal of the American Medical Informatics Association*, Vol. 16, No. 3, pp. 371–379, 2009.

[46] Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. A multi-task learning formulation for predicting disease progression. In *Proceedings of the*

17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 814–822, 2011.

[47] Jiayu Zhou, Jun Liu, Vaibhav A. Narayan, and Jieping Ye. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1095–1103, 2012.

[48] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 85–94, 2014.

[49] Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 135–144, 2014.

[50] Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, Vol. 6, pp. 615–637, 2005.

[51] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, Vol. 45, No. 3, pp. 503–528, 1989.

[52] Jason Lee, Yuichi Imanaka, Miho Sekimoto, Haruo Nishikawa, Hiroshi Ikai, and Takako Motohashi. Validation of a novel method to identify healthcare-associated infections. *Journal of Hospital Infection*, Vol. 77, No. 4, pp. 316–320, 2011.

[53] Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 457–464, 2009.

[54] Daoqiang Zhang, Dinggang Shen, and The Alzheimer's Disease Neuroimaging Initiative. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *NeuroImage*, Vol. 59, No. 2, pp. 895–907, 2012.

[55] Ping Zhang, Fei Wang, and Jianying Hu. Towards drug repositioning: A unified computational framework for integrating multiple aspects of drug similarity and disease similarity. In *AMIA Annual Symposium Proceedings*, pp. 1258–1267, 2014.

[56] Brian Quanz and Jun Huan. Aligned graph classification with regularized logistic regression. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pp. 353–364, 2009.

[57] Gang Chen, Yangqiu Song, Fei Wang, and Changshui Zhang. Semi-supervised multi-label learning by solving a sylvester equation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pp. 410–419, 2008.

[58] Zitao Liu and Milos Hauskrecht. Learning adaptive forecasting models from irregularly sampled multivariate clinical data. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 1273–1279, 2016.

[59] Robert C. Griggs, Mark Batshaw, Mary Dunkle, Rashmi Gopal-Srivastava, Edward Kaye, Jeffrey Krischer, Tan Nguyen, Kathleen Paulus, and Peter A. Merkel. Clinical research for rare disease: opportunities, challenges, and solutions. *Molecular Genetics and Metabolism*, Vol. 96, No. 1, pp. 20–26, 2009.

[60] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 193–200, 2007.

[61] Pengcheng Wu and Thomas G. Dietterich. Improving SVM accuracy by training on auxiliary data sources. In *Proceedings of the 21st International Conference on Machine Learning*, p. 110, 2004.

[62] Sinno Jialin Pan, James T. Kwok, and Qiang Yang. Transfer learning via dimensionality reduction. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, pp. 677–682, 2008.

[63] Neil D. Lawrence and John C. Platt. Learning to learn with the informative vector machine. In *Proceedings of the 21st International Conference on Machine Learning*, p. 65, 2004.

[64] Anton Schwaighofer, Volker Tresp, and Kai Yu. Learning gaussian process kernels via hierarchical bayes. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, pp. 1209–1216, 2004.

[65] Steffen Bickel, Jasmina Bogojeska, Thomas Lengauer, and Tobias Scheffer. Multi-task learning for HIV therapy screening. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 56–63, 2008.

[66] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori,

and Yan Liu. Deep computational phenotyping. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 507–516, 2015.

[67] Joyce C. Ho, Joydeep Ghosh, and Jimeng Sun. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 115–124, 2014.

[68] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 705–714, 2015.