Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Learning Implicit Tasks for Patient-Specific Risk Modeling in ICU

Nozomi Nori
Graduate School of Informatics,
Kyoto University

Hisashi Kashima
Graduate School of Informatics,
Kyoto University

Kazuto Yamashita
Graduate School of Medicine,
Kyoto University

Susumu Kunisawa
Graduate School of Medicine,
Kyoto University

Yuichi Imanaka
Graduate School of Medicine,
Kyoto University

Abstract

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). 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. The ICU is an especially demanding environment for clinicians because of the diversity of patients who mostly suffer from multiple diseases of various types. In this paper, we propose a mortality risk prediction method for ICU patients. The method is intended to enhance the severity assessment by considering the diversity of patients. Our method produces patient-specific risk models that reflect the collection of diseases associated with the patient. 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 multi-task learning method where latent tasks are learned based on the collection of diseases. We demonstrate the effectiveness of our proposed method using a dataset collected from a hospital. Our method achieved higher predictive performance compared with a single-task learning method, the "de facto standard," and several multi-task learning methods including a recently proposed method for ICU mortality risk prediction. Furthermore, our proposed method could be used not only for predictions but also for uncovering patient-specificity from different viewpoints.

Introduction

Accurate assessment of the severity of a patient's condition plays a fundamental role in acute hospital care such as in an intensive care unit (ICU), where clinicians intensively attend to seriously ill patients. The ICU is an especially demanding environment for clinicians due to the *diversity of the patients*: ICU clinicians are faced with patients with various different types of diseases and who are all severely ill. 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, 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 in an attempt to enhance the severity assessment of ICU patients (Tabak et al. 2014; Ghassemi et al. 2014; 2015; Cai et al. 2015; Luo et al. 2016). The underlying assumption is that mortality risk could be used as a surrogate to describe the severity of a patient's condition, and accurate prediction of this risk 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 the patients; however, this approach might fail to capture the diversity of ICU patients. For example, a kind of stomach medicine can be administered to patients who have received artificial respiration because artificial respiration frequently causes gastric ulcers, whereas the stomach medicine is also directly administered to treat severe gastric ulcers with bleeding; as a result, the corresponding prediction rule would differ depending on which type of disease a patient has. In an attempt to address situations such as this, a recent study explored disease-specific risk modeling for ICU patients by employing multi-task learning where a task corresponds to a disease (Nori et al. 2015). They assumed that each disease has a specific prediction rule that explains its mortality risk, and therefore, customizing the risk model for each disease would enhance the predictive modeling.

Yet, their approach continues to be insufficient to capture *patient-specific* aspects of mortality risk modeling for ICU patients. Specifically, patients are usually associated with *multiple diseases*, thereby further complicating ICU patients' clinical states. In such a case, one straightforward approach might be to create a task corresponding to a combination of the diseases that each patient has; however, this approach would be ineffective due to the combinatorial explosion among diseases. This problem is illustrated in Figure 1, which shows a histogram of a combination of diseases that was created from an ICU dataset. This dataset consists of about 200,000 patients from about 170 hospitals in Japan. Each patient is associated with his or her main disease and several comorbidities. A disease is identified by a three-digit ICD-10 code, and the number of comorbidities is at most four. It is observed that many combinations are identical to

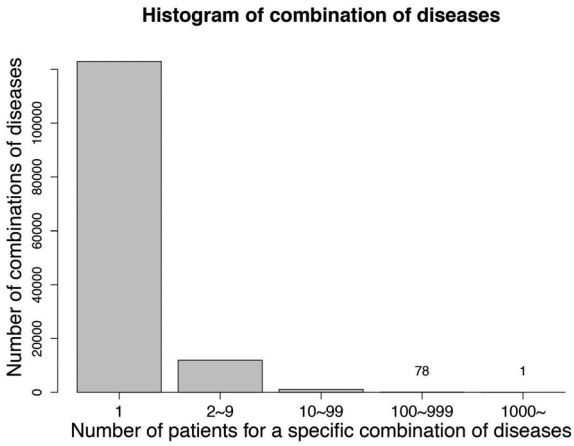

Figure 1: Histogram of combination of diseases in an ICU dataset.

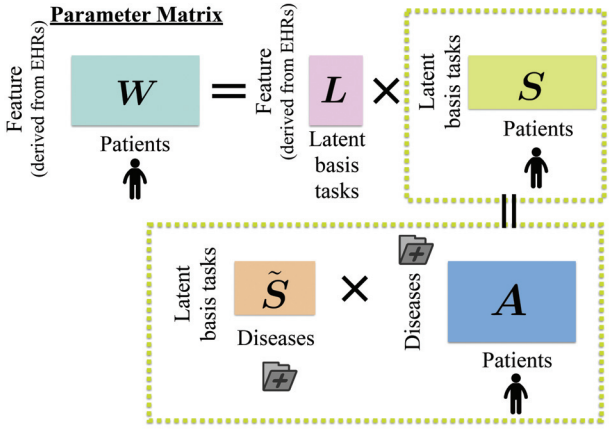

Figure 2: Our model of multi-task learning for patients with multiple diseases.

a specific patient: they only occur in one patient. As a result, a naïve approach to create tasks corresponding to combinations of diseases would be unfeasible.

In this study, we propose a multi-task learning method for ICU mortality prediction capable of producing patient-specific risk models reflecting the collection of diseases associated with the patient. We do not explicitly create tasks corresponding to the collection of diseases; instead, we assume *implicit*, or *latent* tasks and learn a latent representation of the diseases. Figure 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 composes 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 comprises 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 a multi-task learning method where latent tasks are learned based on the collection of diseases.

The contributions of our study are as follows.

- We propose a multi-task learning method for ICU mortality prediction that can produce a patient-specific risk model reflecting the collection of diseases the patient is associated with. For patient-specific modeling, one critical issue is to determine which unit we should use to model patient specificity. Our method enables us to learn the unit of the task itself based on the collection of diseases the patients are associated with, by introducing latent basis tasks.

- We demonstrate the effectiveness of our method by using a real-world dataset from a hospital. Our method is capable of constructing patient-specific models from different viewpoints.

Related Work

Patient-Specific Modeling

Addressing both the *specificity* and the *commonality* among a patient population - that is, addressing the specificity of the target patient while capturing the common structure shared among 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; on the other hand, restricting the data to those that are faithfully relevant to the target often results in an impractically small sample size. Thus, to develop clinically useful target-specific models, recent studies have focused on multi-task learning and transfer learning for clinical data. One critical issue is to determine the aspects we should use to capture the specificity of the patients, that is, how to define the tasks for patient-specific modeling. Jenna et al. investigated the effectiveness of hospital-specific (Wiens, Guttag, and Horvitz 2014) modeling via a feature-representation-transfer method. Gong et al. exploited an instance-transfer method for hospital-specific and surgery-specific risk modeling (Gong et al. 2015). Nori et al. proposed a multi-task learning method in which a task corresponds to a disease (Nori et al. 2015) and Liu and Hauskrecht developed a forecasting model that captures both the patient-specific time series pattern and population-level information for clinical time series data (Liu and Hauskrecht 2016). Our study differs from the above-mentioned research in that we learn the unit of the tasks itself 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 multi-task learning in which a task corresponds to a combination of diseases.

Mortality Risk Modeling for ICU Patients

Traditionally, mortality modeling for the ICU patients has been conducted via scoring systems such as the simplified acute physiology score (SAPS) and acute physiology

and chronic health evaluation (APACHE) both of which use fixed clinical decision rules based mainly on physiological data (Siontis, Tzoulaki, and Ioannidis 2011). However, it should be noted that these ICU scoring systems are only used in rather limited situations. Specifically, as of 2012, they were used for 10-15% of ICU patients in the US (Breslow and Badawi 2012). With the increased availability of varied data from hospital electronic health records (EHRs), the feasibility of data-driven mortality predictive models based on EHRs has been explored extensively in the clinical domain (Hug and Szolovits 2009; Joshi and Szolovits 2014; Tabak et al. 2014; Lehman et al. 2012). These studies demonstrate that EHRs can be used to generate clinically plausible mortality predictive models with superior discrimination. Many studies have attempted to adequately address the typical nature of EHRs, such as data sparsity, in developing mortality prediction methods (Caballero Barajas and Akella 2015; Ghassemi et al. 2014; 2015; Nori et al. 2015).

Nevertheless, thus far it has been implicitly assumed that one common predictive model should be developed and applied to all diseases. In a recent study (Nori et al. 2015), the authors formulated mortality prediction as multi-task learning in which a task corresponds to a disease, thereby producing disease-specific models. However, their method is unable to accommodate the collection of diseases each patient is associated with; thus, in their method each task corresponds to one disease. Yet, because patients are usually associated with multiple diseases, it would be necessary to capture patient specificity by constructing the tasks based on the collection of diseases each patient is associated with. In addition, their method does not learn the relations among diseases, whereas our method learns the relations among diseases by learning latent tasks from the collection of the diseases. Contrary to this, their 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, there should be intrinsic relations among diseases that can be learned from data and can be exploited for prediction purposes.

Multi-task learning

One of the most fundamental issues in multi-task learning is how to introduce an inductive bias in modeling task relationships. There have been numerous studies conducted in an attempt to introduce appropriate assumptions: task parameters might be assumed to lie in close proximity with each other in some geometric sense (Evgeniou and Pontil 2004), or they might be assumed to lie in a low-dimensional subspace (Ji and Ye 2009) or to share a prior in common (Yu, Tresp, and Schwaighofer 2005), to name a few. One major challenge in multi-task learning is how to avoid negative-transfer: that is, how to selectively share information among tasks in order that unrelated tasks do not affect each other. Approaches to this challenge include clustering tasks (Jacob, philippe Vert, and Bach 2009) and learning grouping of tasks with overlap (Kumar and Daumé III 2012). Our method could be considered a multi-task learning method with the assumption that tasks lie in a low-dimensional subspace, similar to several above-mentioned research (Ji and Ye 2009; Kumar and Daumé III 2012). However, our method basically differs other research in that we learn the unit of the task itself. We assume components of the tasks, namely, a collection of diseases, and assume that a combination of the components produces a task; in addition, we do not explicitly list the combination.

Learning patient-specific risk models

Problem setting

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 the EHRs, 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. They are usually several main diagnoses coded by the International Statistical Classification of Diseases and Related Health Problems, ICD, which is a widely used classification of diseases maintained by the World Health Organization. The total number of diseases 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 a mortality risk, 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. We need to address the data sparsity inherent in constructing patient-specific risk models; to that end, we learn models for all the patients jointly, i.e., 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. Under this assumption, we can write the 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 weights of linear combination 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 the matrix $\boldsymbol{L}$ and the latent representation of the patients is captured by the 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 the 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 the disease.

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

Our goal is to estimate the probability of 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, and 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 the parameter matrices $\boldsymbol{L}$ and $\tilde{\boldsymbol{S}}$ using a training dataset, we can construct a parameter matrix specialized to new patients, given a patient-disease relation matrix for encoding the collection of diseases each new patient is associated with.

Optimization problem

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

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

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

$$\Omega_1 \equiv ||\boldsymbol{L}\tilde{\boldsymbol{S}}\boldsymbol{A} - \boldsymbol{W}_0||_F^2, \quad (2)$$

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:

$$\Omega_2 \equiv ||\tilde{\boldsymbol{S}}||_F^2, \quad (3)$$

we define our regularization term as follows:

$$\Omega \equiv \lambda_1\Omega_1 + \lambda_2\Omega_2, \quad (4)$$

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

The optimization problem is defined as follows:

$$\min_{\tilde{\boldsymbol{S}}, \boldsymbol{L}} \mathcal{L}(\tilde{\boldsymbol{S}}, \boldsymbol{L}) + \Omega. \quad (5)$$

In the following, 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 $\boldsymbol{L}$ are given as follows, respectively:

$$\frac{\partial \mathcal{L}}{\partial \tilde{\boldsymbol{S}}} = \sum_{n=1}^{N}(\sigma_n - y_n)\boldsymbol{L}^{\top}\boldsymbol{x}_n\boldsymbol{A}_{*,n}{}^{\top}, \quad (6)$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{L}} = \sum_{n=1}^{N}(\sigma_n - y_n)\boldsymbol{x}_n\boldsymbol{A}_{*,n}{}^{\top}\tilde{\boldsymbol{S}}^{\top}, \quad (7)$$

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 follows, respectively:

$$\frac{\partial \Omega}{\partial \tilde{\boldsymbol{S}}} = 2\lambda_1(\boldsymbol{L}^{\top}\boldsymbol{L}\tilde{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{A}^{\top} - 2\boldsymbol{L}^{\top}\boldsymbol{W}_0\boldsymbol{A}^{\top}) + 2\lambda_2\tilde{\boldsymbol{S}}. \quad (8)$$

$$\frac{\partial \Omega}{\partial \boldsymbol{L}} = 2\lambda_1(\boldsymbol{L}\tilde{\boldsymbol{S}}\boldsymbol{A}\boldsymbol{A}^{\top}\tilde{\boldsymbol{S}}^{\top} - 2\boldsymbol{W}_0\boldsymbol{A}^{\top}\tilde{\boldsymbol{S}}^{\top}). \quad (9)$$

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

$$\frac{\partial \textbf{vec}\,(\mathcal{L}')}{\partial \textbf{vec}\left(\tilde{\boldsymbol{S}}\right)^{\top}} = \sum_{n=1}^{N}\sigma_n(1-\sigma_n)\boldsymbol{V}\boldsymbol{V}^{\top}, \quad (10)$$

$$\frac{\partial \textbf{vec}\,(\Omega_1')}{\partial \textbf{vec}\left(\tilde{\boldsymbol{S}}\right)^{\top}} = 2\boldsymbol{A}\boldsymbol{A}^{\top} \otimes \boldsymbol{L}^{\top}\boldsymbol{L}, \quad (11)$$

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, Eq.(11) produces a positive semidefinite matrix; hence, the regularization term is convex in $\tilde{\boldsymbol{S}}$ for a fixed $\boldsymbol{L}$. Similarly, the regularization term is convex in $\boldsymbol{L}$ for a fixed $\tilde{\boldsymbol{S}}$. However, they are not jointly convex. We adopt an alternating optimization procedure that converges to a local minimum. For each optimization problem, the optimal solution is found by using standard gradient-based methods. We applied the L-BFGS optimizer (Liu and Nocedal 1989) with the above-mentioned derivatives. For initializing $\boldsymbol{L}$, we adopted the following strategy: assuming an M × N matrix $\boldsymbol{W}_0$ with each column containing a parameter vector obtained from a single-task learning method, the 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.

Experiment

Setup

Dataset. We used a dataset from a hospital in Japan. [1] All the patients in the dataset underwent ICU treatment at some point during their hospital stay. For the coding of diseases,

[1] This dataset was constructed as part of the Quality Indicator/Improvement Project (Lee et al. 2011) that is administered by the Department of Healthcare Economics and Quality Management, Kyoto University.

Table 1: Comparison of averaged AUCs. For each method, AUC and standard error is shown.

Method	AUC
Proposed	0.764 ±0.001
Proposed-w/o-A	0.724 ±0.001
Proposed-w/o-pop	0.733 ±0.001
STL (separate)	0.720 ±0.001
STL (common)	0.754 ±0.001
MTL-Trace	0.738 ±0.001
MTL-Mean	0.736 ±0.000
MTL-$\ell_{2,1}$	0.738 ±0.000
MTL-DM	0.757 ±0.001

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 comorbidities the patient had at the time of admission, where the number of comorbidities is at most four. After excluding patients under 18, in this study, we obtained 296 patients. This brings the total number of the diseases to 231. For the features, we used the age, gender, main disease, and comorbidities of the patient that are coded by a four-digit ICD-10, and all the medical events for which patients were billed during their 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 otherwise 0.

Prediction settings. As a measure of mortality, we adopted in-hospital mortality; that is, if a patient died during his or her hospital stay, the patient outcome is "death" and otherwise "survival". As an evaluation measure for predictive performance, we adopted the area under the ROC curve (AUC). We randomly sampled 60% of the patients to create the training dataset and used the remaining 40% for evaluation. We used all the 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 1,062. We repeated the procedure of sampling, prediction, and evaluation and calculated the mean. The hyperparameters were tuned by three-fold cross validation in the training dataset. For our proposed method, λ_1 was tuned among $\{0, 10^{-3}, 10^{-1}\}$, and λ_2 was set to 10^{-5}. The number of latent tasks K was tuned among $\{2^2, 2^3, 2^4\}$. The matrix $\boldsymbol{W}_0$ is constructed by applying logistic regression with ℓ_2-norm regularization to all the patients. We used $\boldsymbol{W}_0$ that is learned by only the training data in each iteration. In our experiment on this dataset, increasing the number of iterative cycles in the optimization process did not improve prediction performance. Hence, we only estimated $\tilde{\boldsymbol{S}}$ using the initial $\boldsymbol{L}$ throughout the experiment.

Compared methods. We compared our proposed method with the following 8 methods. We first prepared the following two variants of our method. First, *Proposed-w/o-A* is adopted to determine the effect of the 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*; *Proposed-w/o-pop* is identical to our method with $\lambda_1 = 0$ in Eq. 4. The next two methods are single-task learning methods. The method *STL (separate)* learns separate models for each disease by using only data item relating to the particular disease, where the disease is defined based on the patient's main disease. *STL (common)* learns one common model that is applicable to all the patients by using data from all the diseases. The other four methods are multi-task learning baselines; for these methods, tasks are defined as patients' main diseases. The first method is *MTL-Trace* (Ji and Ye 2009), 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* (Evgeniou and Pontil 2004), which assumes each task parameter vector is close to the mean vector of all the tasks. The third method, *MTL-*$\ell_{2,1}$ (Argyriou, Evgeniou, and Pontil 2006), 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* (Nori et al. 2015), which integrates domain knowledge relating to the diseases and EHRs via two graph Laplacians. All the single-task and multi-task learning baselines are based on logistic regression with ℓ_2 regularization. For STL (separate) and ST(common), the ℓ_2 regularization hyperparameter was tuned among $\{10^{-3}, 10^{-1}, 10^{0}\}$. For *MTL-Trace*, *MTL-Mean*, and *MTL-*$\ell_{2,1}$, all the hyperparameters were tuned among $\{10^{-2}, 10^{-1}, 10^{0}\}$. For *MTL-DM*, the hyperparameter relating to the task similarity was tuned among $\{10^{-1}, 10^{0}\}$, the hyperparameter relating to the feature similarity was set to 10^{-4}, and the ℓ_2 regularization hyperparameter was set to 10^{-5}.

Results and Discussion

Predictive performance. Table 1 compares the AUCs of the various methods. For each method, we showed AUC and standard error. 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 the prediction performance in fact. 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 suggests the importance of capturing the diversity of ICU patients in mortality risk prediction. Lastly, the performance improvements compared to the four multi-task learning methods indicate that for patient-specific modeling, it is of importance to exploit not only one disease but also the entire collection of diseases associated with the patient.

Latent task analysis. Since the number of latent task K depends on the sample trial, we adopted one sample for the illustration purpose, where $K = 4$. We first examined the relationship among latent tasks in terms of associated diseases by using $\tilde{\boldsymbol{S}}$. Specifically, for each latent task k, we

Table 2: Example of top 10 predictive features for a latent task.

Latent task	High/Low risk	Example of category of predictive features	Example of predictive features
disease-mortality-related task ($k = 1$)	High	high-mortality diseases	C54, I31, K26
	Low	low-mortality diseases	I71, K63

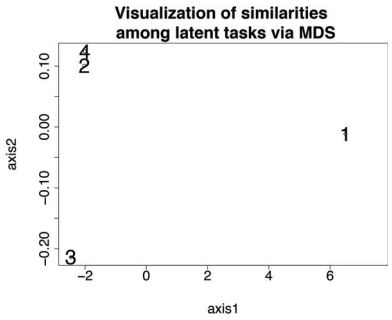

Figure 3: Visualization of similarities among latent tasks via MDS using $\tilde{\boldsymbol{S}}$.

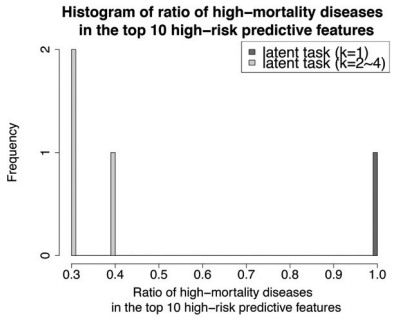

Figure 4: Histogram of ratio of high-mortality diseases in the top 10 high-risk predictive features.

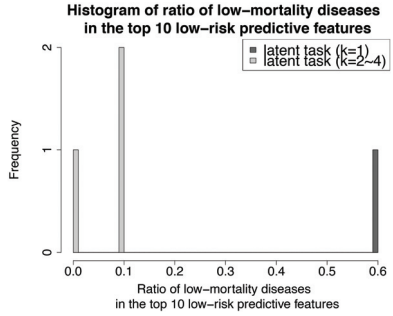

Figure 5: Histogram of ratio of low-mortality diseases in the top 10 low-risk predictive features.

associated it with its disease vector $\tilde{\boldsymbol{S}}_{k,*}$ and applied multi dimensional scaling (MDS) in an attempt to see the relationships among latent tasks. Figure 3 show the result: only one latent task ($k = 1$) was positioned as an outlier, while all the other latent tasks aligned with one dimension.

Then, we examined predictive features for each latent task by using $\boldsymbol{L}$. Specifically, we examined top 10 features with positive and negative coefficients for each latent task by calculating the following two ratios for each latent task: 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. Figure 4 and Figure 5 show histograms of them: Figure 4 shows a histogram of ratio of high-mortality diseases in the top 10 high-risk predictive features for each task, and Figure 5 shows a histogram of ratio of low-mortality diseases in the top 10 low-risk predictive features for each task. It is observed that 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 the other latent tasks.

Together with the MDS result, it is considered that one latent task ($k = 1$) plays a role to capture disease-mortality, while other latent tasks play different roles. Table 2 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 task ($k = 2 \sim 4$), 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 considered that patient-specific models are constructed from viewpoints such as whether the patient is associated with high-mortality diseases or low-mortality diseases, and whether the patient has a specific combination of diseases.

Conclusion

In this study, we considered the risk prediction problem associated with the mortality of diverse ICU patients by producing patient-specific risk models. Our proposed method could be considered a multi-task learning 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 dataset from a hospital demonstrated the effectiveness of our method by outperforming standard single-task learning methods and various multi-task learning methods in which a task corresponds to a disease. Furthermore, our method could be used for uncovering patient-specificity from different viewpoints.

Acknowledgments

Nozomi Nori was supported by Grant-in-Aid for JSPS Fellows (269329).

References

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2006. Multi-task feature learning. In *Advances in Neural Information Processing Systems 19*, 41–48.

Breslow, M. J., and Badawi, O. 2012. Severity scoring in the critically ill: Part 1-interpretation and accuracy of outcome prediction scoring systems. *CHEST Journal* 141(1):245–252.

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

Cai, X.; Perez-Concha, O.; Coiera, E.; Martin-Sanchez, F.; Day, R.; Roffe, D.; and Gallego, B. 2015. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *Journal of the American Medical Informatics Association*.

Evgeniou, T., and Pontil, M. 2004. Regularized multi-task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 109–117.

Ghassemi, M.; Naumann, T.; Doshi-Velez, F.; Brimmer, N.; Joshi, R.; Rumshisky, A.; and Szolovits, P. 2014. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 75–84.

Ghassemi, M.; Pimentel, M. A. F.; Naumann, T.; Brennan, T.; Clifton, D. A.; Szolovits, P.; and Feng, M. 2015. 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*.

Gong, J. J.; Sundt, T. M.; Rawn, J. D.; and Guttag, J. V. 2015. Instance weighting for patient-specific risk stratification models. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 369–378.

Hug, C. W., and Szolovits, P. 2009. Icu acuity: Real-time models versus daily models. In *AMIA Annual Symposium Proceedings*, volume 2009, 260–264.

Jacob, L.; philippe Vert, J.; and Bach, F. R. 2009. Clustered multi-task learning: A convex formulation. In *Advances in Neural Information Processing Systems 21*. 745–752.

Ji, S., and Ye, J. 2009. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 457–464.

Joshi, R., and Szolovits, P. 2014. Prognostic physiology: modeling patient severity in intensive care units using radial domain folding. In *AMIA Annual Symposium Proceedings*, volume 2012, 1276–1283.

Kumar, A., and Daumé III, H. 2012. Learning task grouping and overlap in multi-task learning. In *Proceedings of the 29th International Conference on Machine Learning*, 1383–1390.

Lee, J.; Imanaka, Y.; Sekimoto, M.; Nishikawa, H.; Ikai, H.; and Motohashi, T. 2011. Validation of a novel method to identify healthcare-associated infections. *Journal of Hospital Infection* 77(4):316–320.

Lehman, L.; Saeed, M.; Long, W.; Lee, J.; and Mark, R. 2012. Risk stratification of icu patients using topic models inferred from unstructured progress notes. In *AMIA Annual Symposium Proceedings*, volume 2012, 505–511.

Liu, Z., and Hauskrecht, M. 2016. Learning adaptive forecasting models from irregularly sampled multivariate clinical data. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

Liu, D. C., and Nocedal, J. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical Programming* 45(3):503–528.

Luo, Y.; Xin, Y.; Joshi, R.; Celi, L.; and Szolovits, P. 2016. 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*, 42–50.

Nori, N.; Kashima, H.; Yamashita, K.; Ikai, H.; and Imanaka, Y. 2015. Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864.

Siontis, G. C.; Tzoulaki, I.; and Ioannidis, J. P. 2011. Predicting death: an empirical evaluation of predictive tools for mortality. *Archives of Internal Medicine* 171(19):1721–1726.

Tabak, Y. P.; Sun, X.; Nunez, C. M.; and Johannes, R. S. 2014. 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* 21(3):455–463.

Wiens, J.; Guttag, J. V.; and Horvitz, E. 2014. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association* 21(4):699–706.

Yu, K.; Tresp, V.; and Schwaighofer, A. 2005. Learning gaussian processes from multiple tasks. In *Proceedings of the 22nd International Conference on Machine Learning*, 1012–1019.