Supplementary materials: Causally-distilled Deep Learning towards Explainable and Generalizable Outcomes Prediction in Critical Care

A Related works

A.1 Medical outcomes prediction

Early prediction of medical outcomes is critical for disease diagnosis and clinical intervention. To this end, medical practitioners have developed a number of scoring systems to monitor the life status of patients, such as Sequential Organ Failure Assessment (SOFA) score [87] and Multiple Organ Dysfunction Score (MODS) [50]. Recently, machine learning has rapidly advanced, prompting numerous studies on its application in predicting medical outcomes. Here, we focus on methods of using electronic health records (EHRs), which are being increasingly deployed in hospitals worldwide to store longitudinal information of patients collected in a care delivery setting [72]. Tree-based algorithms such as XGBoost [15] and Random Forest [11] are accurate and highly interpretable and have been widely used in medical outcome prediction. For example, many studies have used the XGBoost method to predict postoperative mortality [26], Acute Asthma Exacerbation [97], critical states of COVID-19 patients [61], etc. Moreover, a remarkable evolution in computational techniques has led to the development of sophisticated algorithms based on deep learning. RNN Classifier based on Multitask Gaussian Process [19] was used for early prediction of sepsis and achieved high prediction accuracy. xAI-EWS [37] is a highly explainable AI early warning score system and be used for early prediction of several acute critical illnesses. And Tomašev et al [82] proposed an RNN-based deep neural network that enables continuous prediction of future acute kidney injury. However, due to the black-box nature of neural networks, model predictions do not necessarily reflect true causality, so some works have begun to explore the use of causal learning for disease diagnosis and prediction to improve the credibility and interpretability of predictions [57, 60]. Our work differs from these studies by combining causal discovery and deep learning to provide more interpretable and generalizable predictions. Moreover, our model predicts future risks of six critical outcomes with one model.

A.2 Causality in AI

Causality plays a pivotal role in advancing artificial intelligence (AI) research, particularly in areas where understanding the underlying mechanisms of data generation is essential. Unlike traditional machine learning approaches that focus primarily on correlation and prediction, causal inference allows AI systems to go beyond mere associations by uncovering the true cause-effect relationships within data [55]. This ability to reason about interventions, counterfactuals,

and generalizations makes causality a crucial tool for building more robust, interpretable, and generalizable AI models. Integrating causal reasoning can significantly enhance AI's capacity to predict the outcomes of actions in complex environments, which is otherwise challenging for purely data-driven models [57,93]. As a result, the field of causal discovery has become a central topic of interest, bridging the gap between data-driven insights and actionable, cause-effect understanding [64].

Causal discovery, also referred to as causal structure learning, has been significantly explored in machine learning, with respect to both static environments and dynamic time-series data [64, 89]. The development of various methods in this area over the past decades has been remarkable, leading to the categorization of these techniques into several distinct approaches, each addressing causal structure learning from different perspectives.

- **Constraint-based approaches** are among the most established techniques, relying on conditional independence tests to build causal graphs. Notable algorithms like PC [75], FCI [76], and PCMCI [21,63,65] fall under this category. These methods often operate by systematically evaluating the dependencies and independencies between variables, gradually pruning possible causal connections to arrive at a directed acyclic graph (DAG). While effective in many cases, their performance can degrade in high-dimensional datasets or when dealing with complex temporal dependencies, necessitating improvements in scalability and robustness.

- **Score-based learning algorithms** offer another approach by optimizing a score function that balances model fit with penalization for model complexity. These methods aim to minimize a loss function that captures both the goodness of fit and compliance with causal structure constraints, such as the acyclicity constraint. Techniques like penalized Neural Ordinary Differential Equations (NODE) and approaches like DYNOTEARS [9, 54] introduce neural network architectures to extend these models' capacity for handling nonlinear relationships in time-series data. Additionally, these methods are particularly adept at discovering latent structures by using a continuous optimization framework, although they may require significant computational resources and careful hyperparameter tuning.

- **Additive Noise Models (ANM)** are grounded in the assumption that the effect of one variable on another can be modeled by an additive noise term. Introduced in the seminal work by [73], ANM has since been extended to accommodate nonlinear causal structures, as discussed in [29]. This category of methods is powerful for discovering causal directions, especially when the noise model aligns with the assumptions, such as non-Gaussian noise distributions. The flexibility of ANMs in handling various forms of nonlinearity has made them a popular choice for causal inference in systems where the functional relationships are not easily described by linear models, although they may suffer from identifiability issues when the noise assumptions are violated.

- **Granger-causality-based approaches** explore temporal causal relationships by examining whether the past values of one time-series improve the prediction of another, relative to predictions using only the past values of the latter. Originally introduced by [23], Granger causality has evolved beyond its linear roots to encompass nonlinear versions with deep neural networks (DNNs). Recent developments in neural Granger causality [16, 34, 44, 79, 92] have significantly broadened the applicability of this method, especially in discovering complex causal dependencies in high-dimensional, nonstationary time-series data. These deep learning approaches often leverage recurrent neural networks (RNNs) or attention mechanisms to capture intricate temporal patterns, although they introduce challenges related to interpretability and the need for large training datasets. The causal discovery module

in our cDEEP model is also based on Granger causality, which we have adapted to handle high-dimensional, non-linear, and multivariate time-series data.

- **Convergent Cross Mapping (CCM)**, proposed by [77], provides a novel perspective on causal discovery in dynamical systems. CCM reconstructs the state space from time-series data, identifying causal links by evaluating the degree to which the state of one system can predict the future states of another. This method is especially useful for systems that exhibit nonlinear and nonseparable behavior, such as ecological or economic systems. Extensions of CCM have made it applicable to more complex scenarios involving synchrony, confounding, and sparse time-series data [10,12,94]. Despite its utility, CCM's effectiveness depends heavily on the quality and length of the time-series, and it may struggle in settings with significant noise or hidden confounders.

A.3 Generalizability

Domain generalization, also referred to as the out-of-distribution generalization problem, aims to deal with the problem of stable prediction when the test distribution differs from the training distributions. Research on domain generalization has focused on three main directions:

- **Data manipulation:** The primary goal of data manipulation is to enhance the diversity of the training data. In addition to simple operations such as cropping and flipping, a popular method is called domain randomization [81, 83, 96], which increases the domain diversity by performing a series of randomized transformations on the training data. There are also several approaches that use adversarial training and generative models to generate new data distributions for enhanced generalization [40, 58, 100].

- **Domain-invariant representation learning.** Methods such as IRM [6], VREx [35], and IB-ERM [42] aim to discover stable, domain-invariant representations from source domain data by incorporating some invariant constraints. In recent years, there has been growing interest in addressing the out-of-distribution generalization problem from a causal perspective. These approaches typically construct causal associations to model the data generation process, focusing on identifying and utilizing only causal factors to achieve stable predictions [30, 31, 43, 47, 48].

- **Distributionally robust optimization.** This approach emphasizes the model's performance in worst-case scenarios and enhances generalization by minimizing risk in the worst-performing domain. Typical methods include GroupDRO and its variants [66, 67], as well as AdaRNN [18] and Diversity [45] methods, some of which specifically target generalization for time series.

- **Regularization.** Regularization in neural networks controls the trade-off between model capacity and generalization performance. Many approaches have been proposed to reduce over-fitting, such as $\mathcal{L}_1$ and $\mathcal{L}_2$ regularization. Other than that, one of the most frequently used approaches is Dropout [7, 27]. Other regularization approaches include SAE [38], batch normalization [33], and early stopping [56].

Our algorithm can be seen as a combination of domain-invariant representation learning and regularization-based generalization. We use causal discovery to identify the causal variables and thus construct a domain-invariant representation. The regularization term in the loss function also enhances the generalization ability of the model by controlling the model complexity.

A.4 Explainability

With the widespread of machine learning in real applications, the demand for developing explainable models has been growing. In field such as medicine, finance, and law, explainability can be an extremely important problem [24, 68]. Black-box models, such as deep neural networks (NNs), have the advantage of high approximation ability at the cost of decision interpretability [88]. NNs achieve very high accuracy on multiple tasks, e.g., image classification, and video reconstruction. On the other hand, the black-box nature of the complex neural networks greatly hamper the explainability of the models. This is a known trade-off between model capacity and explainability. Our question is that, is it possible to maintain the high performance of neural networks while preserving the high explainability of simple models?

One of the models that inspires us the most is LASSO, which is based on linear regression with $\mathcal{L}_1$ regularization. LASSO is highly explainable since the weights in linear models naturally reflect the contribution of each feature. The $\mathcal{L}_1$ regularizer further explains the model by selecting the most important features. However, the biggest problem of LASSO is its linear assumption, i.e., $y = \beta_{1:p}x + \beta_0 + \epsilon$. This model does not take into account the complex nonlinear and multivariate relationships buried under the massive amount of data. Recently, LASSO has been less frequently used because of the widespread of powerful neural networks which do not need linear assumptions and are proven to be capable of approximating any nonlinear functions [28]. The weights in NNs do not naturally mean feature importance [71, 74], and feature selection with weights regularization is not trivially possible. The feature importance can instead be calculated, with Shapley Values [36, 46], Integrated Gradients [78], or Granger causality [70].

Most existing methods for explaining neural networks can be divided into two categories: inside-out explanation and outside-in explanation. Almost all of these methods calculate the importance/contribution measurements of input features, only with very different paths.

- **Inside-out explanation.** This kind of approach analyzes the model decision by opening the black box, i.e., layer by layer. This category can be further sub-categorized into (i) Gradient-based methods such as Integrated Gradients [78], Grad-CAM [71], Bort [98], and Deep Taylor Decomposition [52]. This subcategory explains the model by calculating gradients on the input features. (ii) Non-gradient propagation-based methods such as DeepLIFT [74] and CAM [99] that back-propagate the contributions without direct usage of gradients. (iii) Explaining with specific network structure such as self-explaining model [4].

- **Outside-in explanation.** Approaches of this class, instead, explain the model with sensitivity-analysis-like experiments from outside the model. This category can be further sub-categorized into (i) Removing-based methods [17] such as CXPlain [70], DeepSHAP [46], and prediction difference analysis [101]. (ii) Approximate/mimic model methods such as LIME [59], AIM [88] and interpretable mimic learning [13]. (iii) Feature selection methods such as L2X [14] and INVASE [95] that explain models by identifying the most contributing features.

Our cDEEP, on the other hand, also falls in the category of outsize-in explanation. We use Granger causality to identify the causal variables and the causal variables are then used to explain the model prediction, which is a novel way to combine the explainability and generalizability of the model.

B Additional results

B.1 Population statistics

We show the presence and missing rate of the six interested outcomes in Tab. S1. "Positive" indicates that the outcome is labeled as positive in the prediction window (within the next 24 hours). When there is no evidence in the prediction window (within the next 24 hours) to identify the outcome, the label is considered "Missing". For example, if creatine level is not measured in the next 24 hours, the AKI label is considered missing.

It is observed that death has an extremely low presence rate, which is consistent with the fact that death is a rare event in the critical care setting. And the model may struggle to predict death due to the extremely imbalanced data. As a result, the AUPRC score of death is relatively low in Extended Data Fig. 5. On the other hand, ARDS has an extremely high missing rate, which may be the reason why a low AUROC score is observed in Fig. 3.

Table S1: Presence and missing rate of the six interested outcomes.

	AKI	ARDS	Circulatory failure	Death	Delirium	Sepsis
Positive	25.1%	4.2%	11.8%	0.9%	4.7%	3.2%
Missing	17.5%	94.4%	78.3%	0.0%	82.7%	88.4%

After being structured into discrete time-series, the variables are allocated into 2-hour periods. However, the sampling gaps for each variable vary significantly and may be far more than 2 hours. As a result, a large number of missing values exist in the dataset. We show the missing rate of each variable in Tab. S2.

Table S2: Missing rate of the structured dynamic variables.

Variable	**Missing Rate**	**Variable**	**Missing Rate**	**Variable**	**Missing Rate**	**Variable**	**Missing Rate**
pH	98.28%	PEEP	99.18%	WBC	96.68%	AST	98.77%
Cl	96.20%	Cr	96.26%	Glucose	96.28%	LDH	99.85%
Mg	97.62%	ALT	98.80%	Na	96.06%	TCO2	99.34%
ALP	98.80%	BUN	96.28%	Ca	96.48%	CK	99.70%
Basos	98.56%	Eos	98.50%	Lymphs	98.38%	Monos	98.39%
iCa	99.24%	Lactate	99.02%	Bili	98.83%	AG	96.80%
SpontRR	99.87%	HR	77.57%	ABPsys	94.42%	ABPdia	94.43%
ABPmean	94.40%	CVP	96.23%	RR	79.47%	SpO2	84.56%
TempF	95.44%	NIBPsys	82.01%	NIBPdia	82.03%	NIBPmean	81.93%

However, the missing rate is only slightly associated with the causal probabilities. We show the correlation between the missing rate and the causal probabilities in Fig. S1. Almost all correlations are weak, indicating that our causal discovery method is not significantly biased by the missing rates of the variables.

B.2 Calibration

We show the calibration curve of cDEEP in Fig. S2 as gray. The calibration curve is a plot of the predicted probability (x-axis) against the true probability (y-axis). The diagonal line

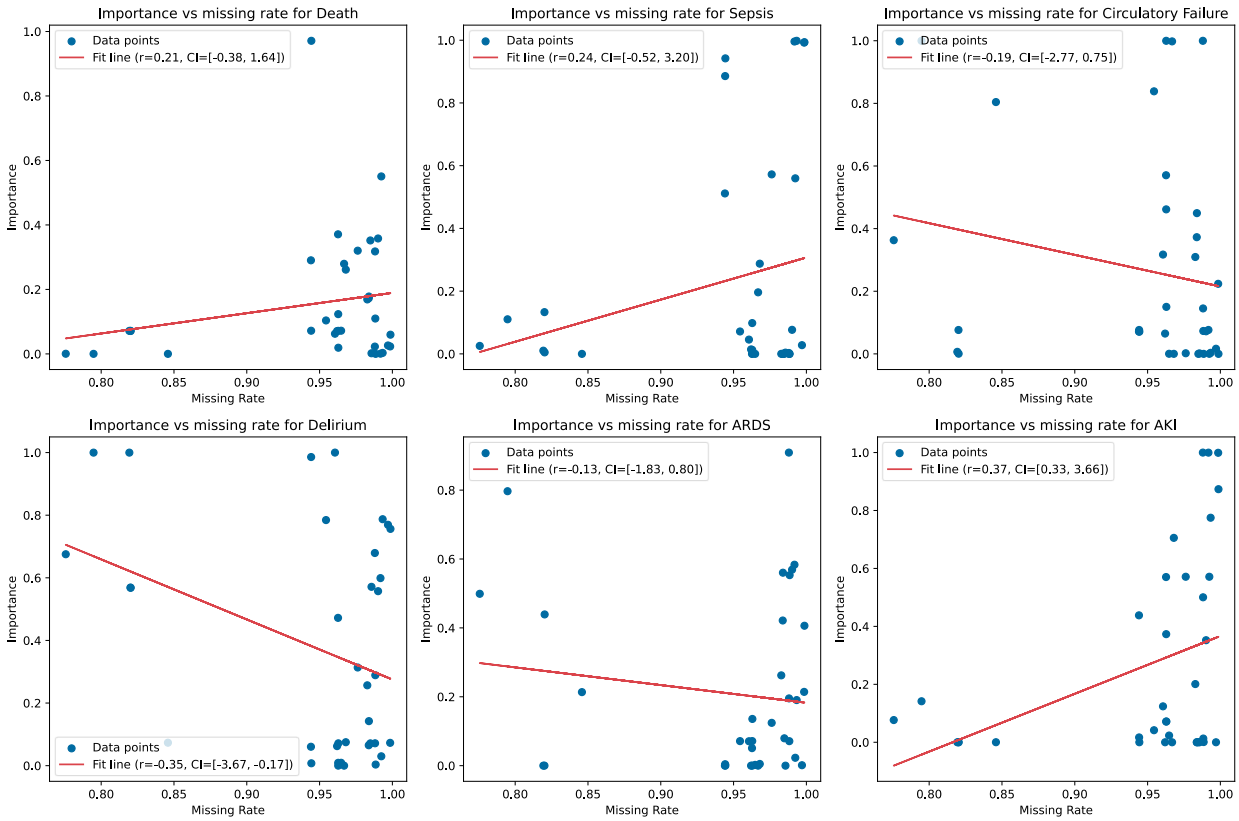

Figure S1: Correlation between the missing rate and the causal probabilities.

represents a perfectly calibrated model, where the predicted probabilities are equal to the true probabilities. The calibration curve of cDEEP is close to the diagonal line, indicating that cDEEP is well-calibrated.

We performed isotonic calibration on cDEEP on the standalone calibration dataset (which is randomly split from the in-distribution test dataset) and show the calibration curve in Fig. S2 as red. The calibration curve of isotonic calibration is closer to the diagonal line, indicating that isotonic calibration improves the calibration of cDEEP. However, after isotonic regression, the Brier scores only decrease very slightly, this is attributed to the fact that the predicted risks are highly imbalanced, shown in Fig. S3. Brier scores are calculated on the entire dataset, which may weigh much more on samples with low predicted risks. And those samples are already well-calibrated, so the Brier scores are not significantly improved after isotonic regression.

B.3 Experiments on generalizability

We show the generalizability of cDEEP in comparison to other generalizable AI methods on the out-of-distribution testing dataset. We evaluate the generalizability of cDEEP by comparing its performance with other generalizable AI methods, including Dropout, GroupDRO, IRM, and VREx. We first split the out-of-distribution testing dataset by age (i.e. training on patients aged ≤ 75 and testing on patients aged ≥ 76) and report the AUROC and AUPRC of cDEEP and other methods on the out-of-distribution testing dataset in Tab. S3 and Tab. S4, respectively.

We also split the out-of-distribution testing dataset by admission time (i.e. training on patients admitted before or in 2014 and testing on patients admitted after 2014) and report the AUROC and AUPRC of cDEEP and other methods in Tab. S5 and Tab. S6, with receiver

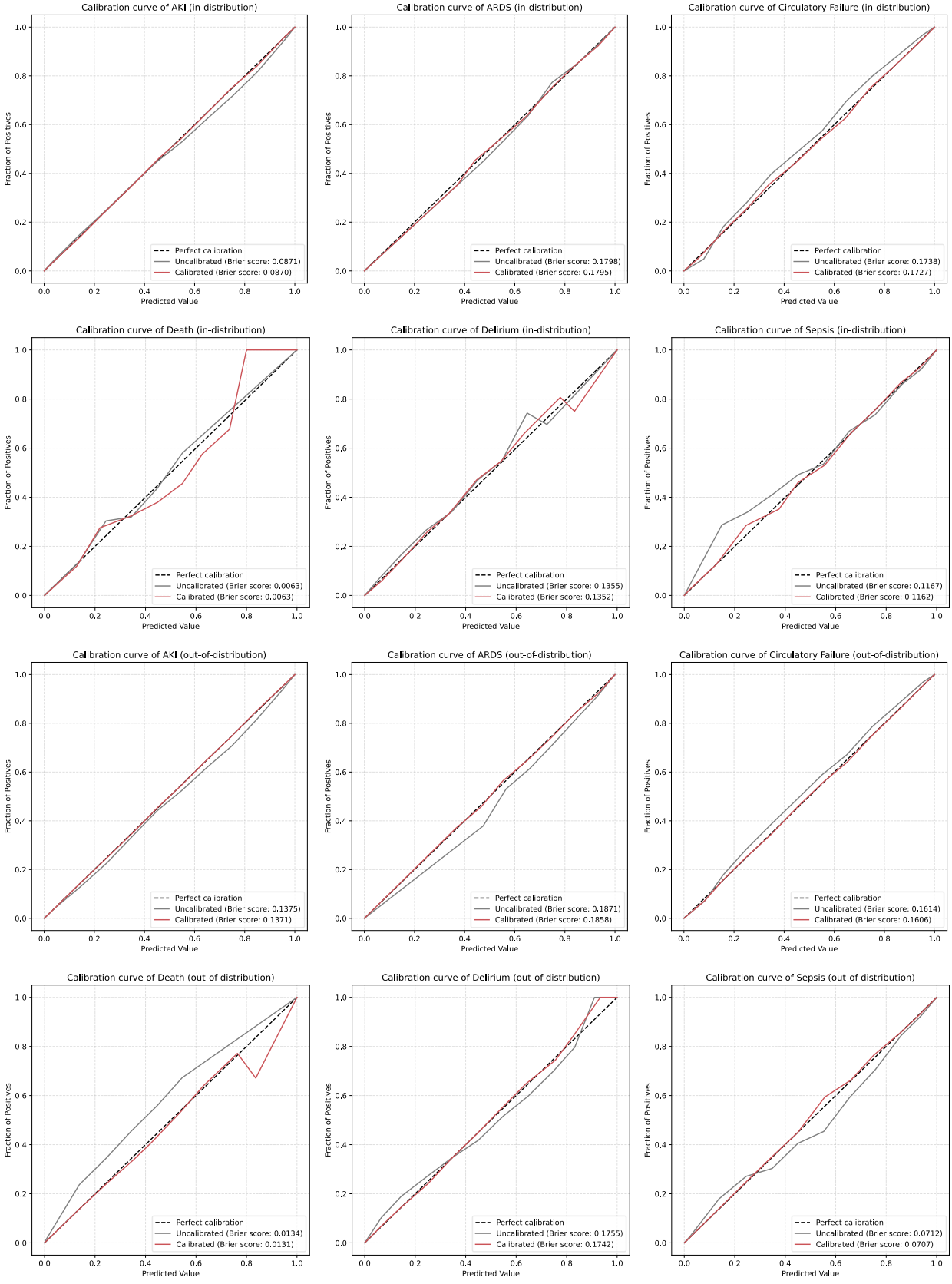

Figure S2: Calibration curves.

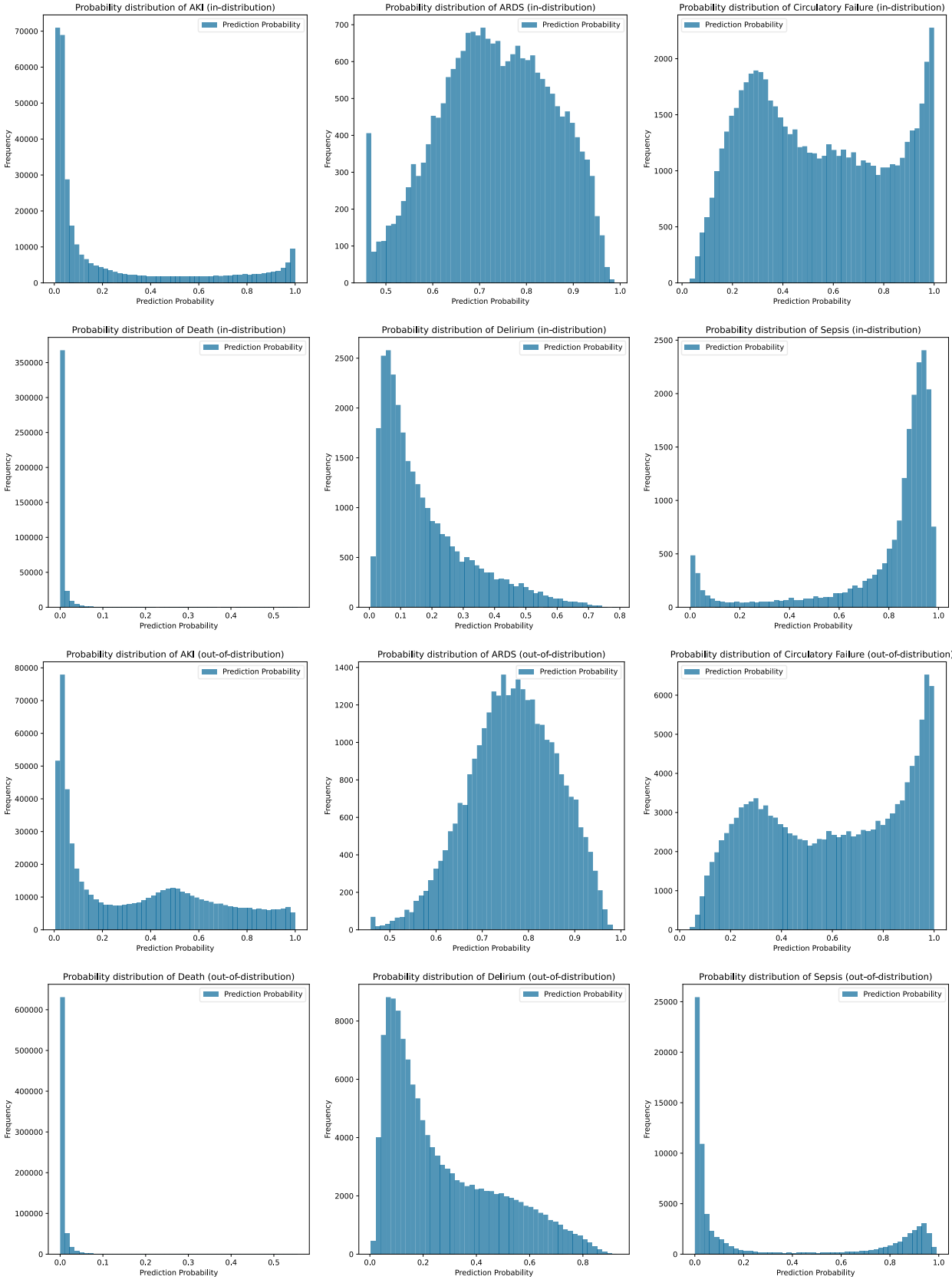

Figure S3: Histograms of all predicted risks.

operating characteristic (ROC) curves and precision-recall curves shown in Fig. S4. The results show that with significantly lower input variable numbers, **cDEEP** achieves comparable or even better performance than other generalizable AI methods already—when split datasets by age, cDEEP outperforms other methods in 3 out of 6 outcomes in terms of AUROC and 5 out of 6 outcomes in terms of AUPRC. When split datasets by admission time, cDEEP outperforms other methods in 4 out of 6 outcomes in terms of AUROC and 5 out of 6 outcomes in terms of AUPRC. **cDEEP-full**, on the other hand, achieves even better performance, beating other methods in most cases in terms of AUROC and AUPRC. These results demonstrate that cDEEP is a generalizable AI method that can achieve high performance on out-of-distribution testing datasets.

Table S3: Comparison to generalizable AI methods in terms of AUROC on out-of-distribution testing dataset split by age. The best and second-best results are highlighted in bold and underlined, respectively.

Split by age			
Methods	AKI	ARDS	Circulatory Failure
cDEEP-full	**0.878 [0.877-0.879]**	**0.659 [0.655-0.661]**	<u>0.832 [0.831-0.832]</u>
cDEEP	0.866 [0.865-0.868]	<u>0.651 [0.648-0.654]</u>	**0.834 [0.833-0.835]**
Dropout	0.853 [0.852-0.853]	0.598 [0.595-0.601]	0.799 [0.798-0.799]
GroupDRO	0.863 [0.862-0.863]	0.610 [0.607-0.613]	0.814 [0.814-0.815]
IRM	<u>0.873 [0.873-0.874]</u>	0.636 [0.634-0.639]	0.824 [0.823-0.825]
VREx	0.865 [0.865-0.866]	0.628 [0.625-0.631]	0.813 [0.811-0.814]
Methods	Death	Delirium	Sepsis
cDEEP-full	**0.894 [0.892-0.896]**	**0.736 [0.734-0.737]**	**0.958 [0.958-0.959]**
cDEEP	0.869 [0.866-0.871]	<u>0.734 [0.732-0.735]</u>	0.952 [0.952-0.953]
Dropout	0.862 [0.860-0.865]	0.724 [0.723-0.726]	0.951 [0.950-0.952]
GroupDRO	0.877 [0.875-0.879]	0.727 [0.725-0.728]	0.953 [0.952-0.953]
IRM	<u>0.885 [0.883-0.887]</u>	0.730 [0.729-0.731]	<u>0.955 [0.955-0.956]</u>
VREx	0.871 [0.869-0.873]	0.728 [0.726-0.729]	0.952 [0.951-0.952]

Additionally, we evaluate the generalizability of cDEEP on the out-of-distribution testing dataset with added noise. After normalizing the input variables to have a mean of 0 and a standard deviation of 1, we add noise to the input variables with a standard deviation of 0.05, 0.1, and 0.2, and report the AUROC and AUPRC of cDEEP and other generalizable AI methods in Fig. S5. The results show that cDEEP and cDEEP-full both achieve better performance than other generalizable AI methods in most cases, demonstrating the robustness of cDEEP to noise.

B.4 Acceleration effects of causally-decoupled inference

We propose to utilize the causally-decoupled inference techniques which only update a subset of the hidden layers of the neural network after perturbing a variable to speed up the CDE value calculation. We show the acceleration effects of causally-decoupled inference in Tab. S7 for a single patient. The results show that the causally-decoupled inference techniques can significantly decrease the numbers of forward propagation and the total time for CDE value calculation, decreasing the computation time by 63.1% and 90.5% for the V2O and V2V models, respectively.

Table S4: Comparison to generalizable AI methods in terms of AUPRC on out-of-distribution testing dataset split by age. The best and second-best results are highlighted in bold and underlined, respectively.

Split by age			
Methods	AKI	ARDS	Circulatory Failure
cDEEP-full	**0.786 [0.784-0.787]**	**0.836 [0.834-0.839]**	0.895 [0.894-0.896]
cDEEP	0.761 [0.759-0.762]	0.832 [0.830-0.835]	**0.897 [0.896-0.898]**
Dropout	0.736 [0.734-0.737]	0.805 [0.802-0.807]	0.865 [0.864-0.866]
GroupDRO	0.753 [0.751-0.755]	0.808 [0.805-0.810]	0.880 [0.879-0.881]
IRM	0.775 [0.773-0.776]	0.823 [0.820-0.825]	0.889 [0.888-0.890]
VREx	0.759 [0.758-0.761]	0.817 [0.815-0.820]	0.884 [0.883-0.884]
Methods	Death	Delirium	Sepsis
cDEEP-full	**0.223 [0.214-0.231]**	**0.540 [0.538-0.543]**	**0.903 [0.902-0.905]**
cDEEP	0.201 [0.194-0.207]	0.535 [0.532-0.537]	0.890 [0.888-0.892]
Dropout	0.155 [0.150-0.160]	0.523 [0.520-0.525]	0.880 [0.878-0.883]
GroupDRO	0.173 [0.167-0.179]	0.523 [0.521-0.526]	0.886 [0.884-0.888]
IRM	0.195 [0.190-0.200]	0.532 [0.530-0.534]	0.889 [0.888-0.891]
VREx	0.172 [0.166-0.178]	0.525 [0.523-0.527]	0.887 [0.885-0.889]

Table S5: Comparison to generalizable AI methods in terms of AUROC on out-of-distribution testing dataset split by admission time. The best and second-best results are highlighted in bold and underlined, respectively.

Split by admission time			
Methods	AKI	ARDS	Circulatory Failure
cDEEP-full	**0.895 [0.894-0.895]**	**0.679 [0.678-0.681]**	**0.833 [0.833-0.834]**
cDEEP	0.891 [0.891-0.892]	0.670 [0.668-0.672]	0.829 [0.828-0.829]
Dropout	0.884 [0.883-0.885]	0.626 [0.624-0.627]	0.812 [0.811-0.812]
VREx	0.743 [0.742-0.744]	0.612 [0.610-0.614]	0.662 [0.661-0.663]
IRM	0.880 [0.880-0.881]	0.639 [0.638-0.641]	0.813 [0.812-0.814]
GroupDRO	0.639 [0.638-0.640]	0.588 [0.586-0.590]	0.754 [0.753-0.755]
Methods	Death	Delirium	Sepsis
cDEEP-full	**0.915 [0.913-0.917]**	**0.772 [0.771-0.773]**	**0.948 [0.947-0.948]**
cDEEP	0.890 [0.887-0.893]	0.769 [0.768-0.769]	0.925 [0.924-0.926]
Dropout	0.892 [0.889-0.894]	0.763 [0.762-0.764]	0.937 [0.936-0.937]
VREx	0.884 [0.881-0.887]	0.698 [0.696-0.699]	0.924 [0.924-0.925]
IRM	0.893 [0.890-0.895]	0.728 [0.727-0.729]	0.937 [0.936-0.937]
GroupDRO	0.771 [0.766-0.775]	0.646 [0.645-0.647]	0.744 [0.743-0.745]

Table S6: Comparison to generalizable AI methods in terms of AUPRC on out-of-distribution testing dataset split by admission time. The best and second-best results are highlighted in bold and underlined, respectively.

Split by admission time			
Methods	AKI	ARDS	Circulatory Failure
cDEEP-full	**0.811 [0.809-0.812]**	**0.851 [0.850-0.853]**	**0.874 [0.873-0.875]**
cDEEP	0.804 [0.803-0.806]	0.846 [0.844-0.847]	0.872 [0.871-0.873]
Dropout	0.788 [0.786-0.789]	0.827 [0.825-0.829]	0.853 [0.852-0.854]
VREx	0.621 [0.619-0.623]	0.810 [0.809-0.812]	0.748 [0.746-0.749]
IRM	0.779 [0.777-0.780]	0.832 [0.831-0.834]	0.856 [0.856-0.857]
GroupDRO	0.491 [0.489-0.493]	0.801 [0.799-0.803]	0.800 [0.799-0.801]
Methods	Death	Delirium	Sepsis
cDEEP-full	**0.232 [0.223-0.241]**	0.572 [0.570-0.574]	**0.836 [0.834-0.837]**
cDEEP	0.199 [0.190-0.209]	**0.577 [0.575-0.579]**	0.716 [0.714-0.718]
Dropout	0.157 [0.151-0.164]	0.570 [0.568-0.572]	0.784 [0.781-0.785]
VREx	0.144 [0.137-0.152]	0.495 [0.493-0.497]	0.690 [0.688-0.693]
IRM	0.166 [0.158-0.173]	0.522 [0.520-0.524]	0.774 [0.772-0.776]
GroupDRO	0.065 [0.062-0.068]	0.426 [0.425-0.428]	0.472 [0.470-0.474]

Table S7: Acceleration effects of causally-decoupled inference, shown with the mean and 95% confidence interval of the computation time for CDE value calculation.

	V2O model	V2V model
Before acceleration	1.674 [1.415-1.933] sec.	1.532 [1.371-1.692] sec.
After acceleration	0.617 [0.490-0.744] sec.	0.146 [0.130-0.162] sec.

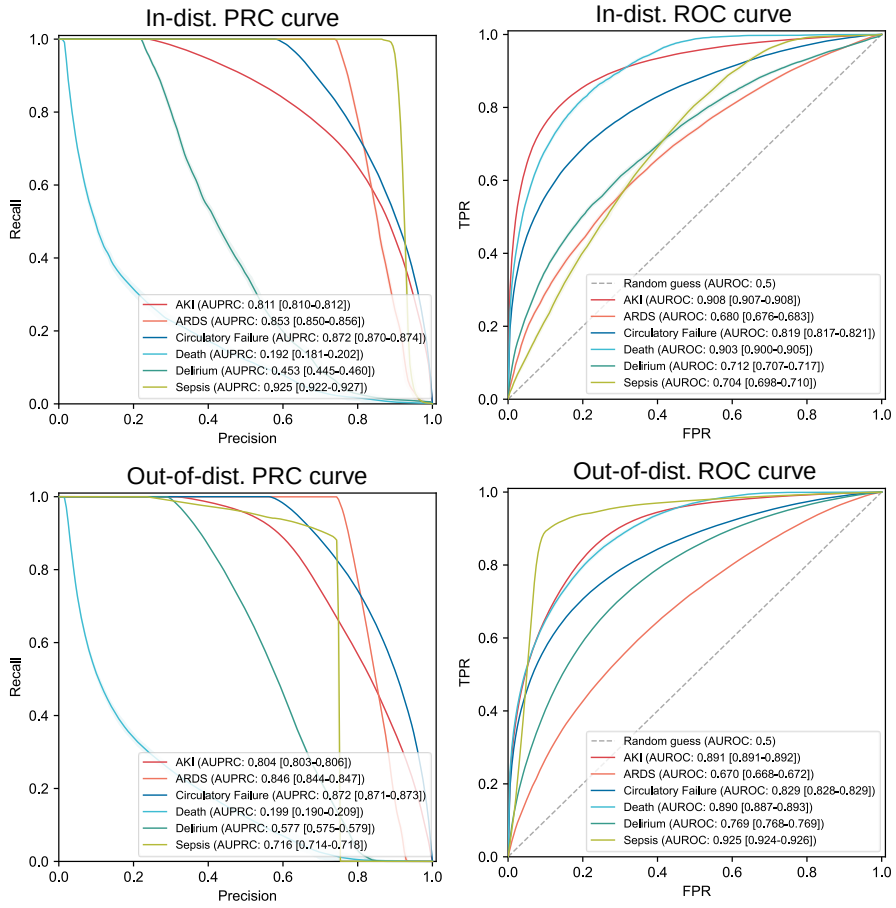

Figure S4: Receiver operating characteristic curve (ROC) and precision-recall curve (PRC) on in-distribution and out-of-distribution testing data split by admission time.

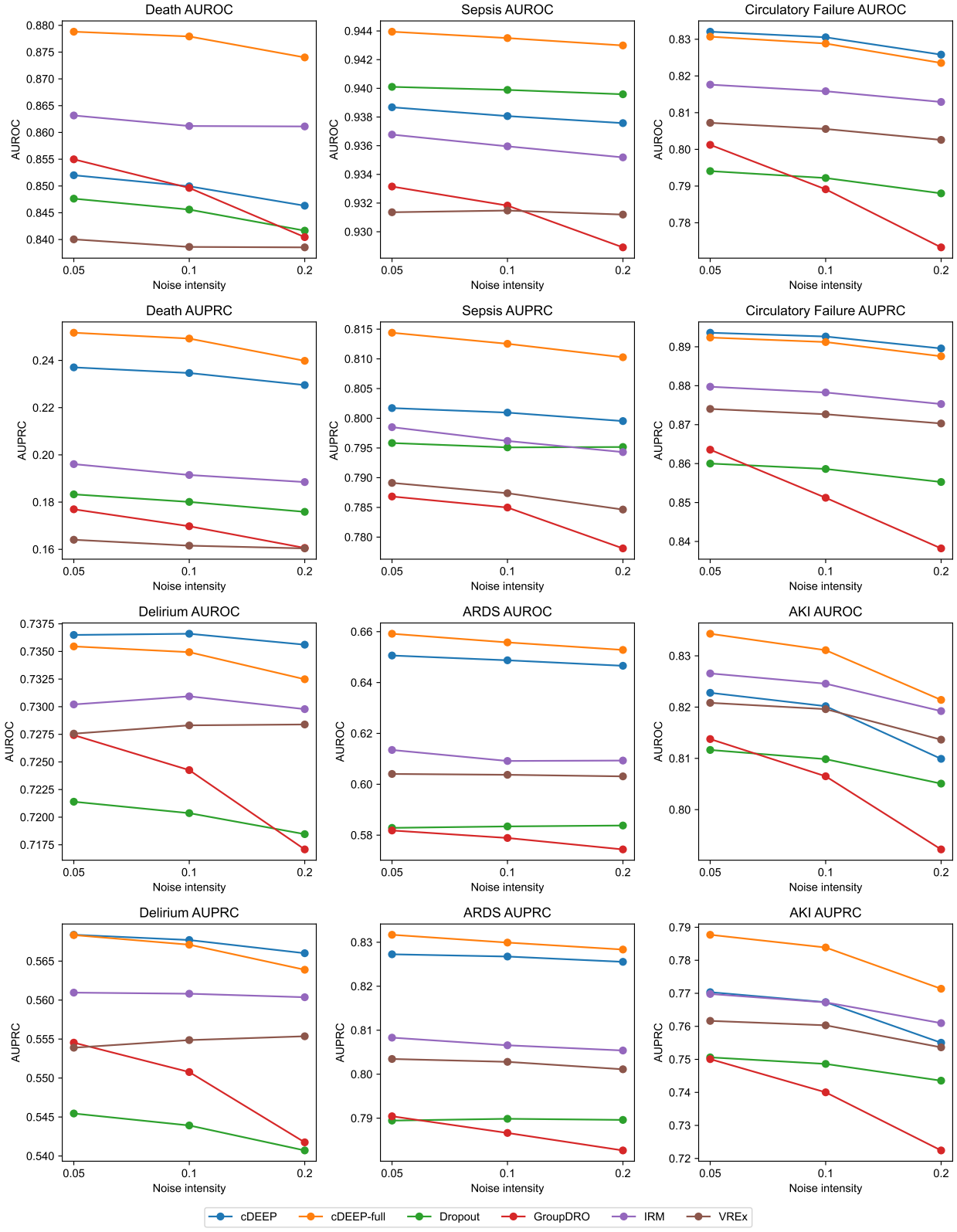

Figure S5: Comparison to generalizable AI methods in terms of AUROC and AUPRC on out-of-distribution testing dataset with added noise.

C Interpretation

C.1 Clinical insights of the interpretation results

The interpretation results in Fig. 2 highlight the clinically relevant causal relationships of the variables for each outcome, demonstrating their alignment with medical knowledge. For acute kidney injury, among the contributing variables in our interpretable approach, creatinine is integral to diagnosing and staging AKI, based on criteria like the Kidney Disease: Improving Global Outcomes (KDIGO) guidelines [1]. Elevated positive end-expiratory pressure (PEEP), a key parameter in mechanical ventilation, may compromise renal perfusion by increasing intrathoracic pressure, thereby contributing to renal dysfunction [69]. Rising lactate levels signal tissue hypoxia and anaerobic metabolism, critical processes in AKI pathophysiology. Lactate dehydrogenase release, particularly under hypoxic conditions, indicates renal tissue damage and injury to other organs [8].

In ARDS, our model has identified several key risk factors that reflect the underlying pathophysiology of the syndrome. ALT (Alanine Aminotransferase) and ALP (Alkaline Phosphatase) may indicate liver involvement in the inflammatory process, which is part of the multi-organ dysfunction associated with ARDS. Some evidence [25,51] suggests that in critically ill patients, hepatic dysfunction is recognized as a relevant clinical condition that influences the development, severity, and progression of ARDS. Heart Rate (HR) and Respiratory Rate (RR) increases are physiological responses to the stress of hypoxemia, as the body attempts to compensate for reduced oxygen delivery [84]. PEEP is a mechanical ventilation strategy used to maintain alveolar recruitment and prevent collapse, which is crucial for improving oxygenation in ARDS patients [91]. Peripheral Oxygen Saturation (Spo2) is a direct measure of oxygenation, and its decrease is a hallmark of ARDS, indicating impaired oxygenation at the tissue level. The ratio of pulse oximetric oxygen saturation to fractional inspired oxygen (SpO2/FiO2) [62,90] has been validated in clinical studies for the diagnosis and risk stratification of patients with ARDS.

In circulatory failure, decreased blood pH denotes severe metabolic acidosis and impaired oxygen utilization [53]. Conversely, a significantly reduced respiratory rate may indicate respiratory depression or central nervous system failure in the later stages of circulatory collapse [49].

Mortality is associated with immune dysfunction, organ damage, and severe metabolic derangements. Abnormal lymphocyte counts suggest diminished adaptive immunity and heightened infection risk, whereas monocyte dysregulation reflects either immune paralysis or excessive inflammation, both of which are strongly linked to adverse outcomes [2,39].

For delirium prediction, tachycardia, often induced by a stress response, may reduce cerebral blood flow, precipitating cognitive dysfunction [5,85]. An increased anion gap and heightened respiratory rate indicate metabolic acidosis and systemic inflammation, impairing cerebral metabolism. Elevated PEEP may exacerbate hypoxia by hindering cerebral venous return [22].

In sepsis, alterations in magnesium levels and the anion gap highlight electrolyte disturbances and underlying metabolic dysfunction [86]. Hypomagnesemia is associated with systemic inflammation and multi-organ dysfunction while an increased anion gap reflects metabolic acidosis secondary to tissue hypoxia and toxin accumulation [20]. In sepsis, inflammatory cytokines like F-α and IL-6 mediate generalized vasodilation and impair endothelial barrier function [3]. This results in decreased vascular tone, reduced systemic vascular resistance, and impaired perfusion to vital organs. Prolonged low diastolic blood pressure despite fluid resuscitation is a defining feature of septic shock, indicating the failure of compensatory mechanisms.

C.2 User guide for our interpretation tool

Our interpretation tool is web-based and accessible via `https://cdeep.icu/`. Upon launching the tool, users will be presented with the main interface, as shown in Fig. S6. To load prediction results, users can click the "Click here" button and select a file with a ".json" or ".crp" extension. Note that ".crp" files are compressed versions of ".json" files, offering reduced file sizes. The interface also supports drag-and-drop functionality for easy file uploads. Additionally, a sample prediction result is available for users to explore the interface without needing to upload their own files, accessible by clicking the "Use sample file" button.

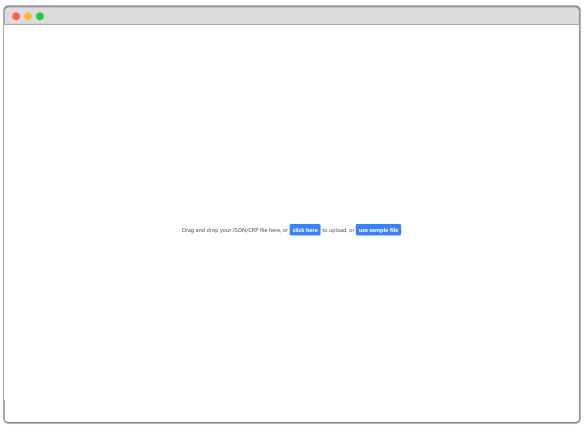

Figure S6: Initial interface.

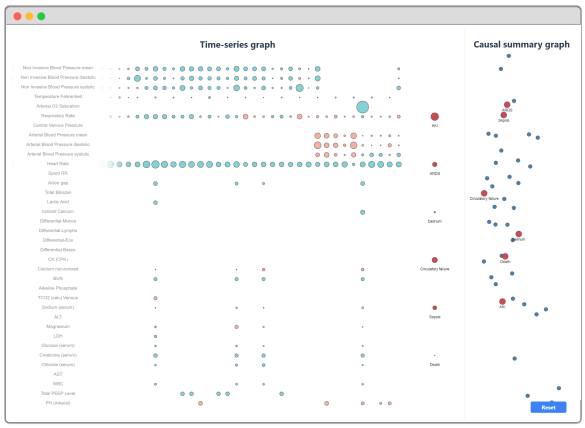

Figure S7: File loaded.

Once the prediction results are loaded, the graph displays variables as bubbles. The window is split into two sections: the left side shows the time-series graph, while the right side displays the summary graph. The color of each bubble indicates the direction of deviation from the average (red for above average and blue for below average), while the size of the bubble reflects the magnitude of the deviation. The horizontal axis represents time in 2-hour intervals, and the vertical axis corresponds to different variables. By scrolling to the right using the bottom scrollbar, users can view the prediction results for the next 24 hours, as illustrated in Fig. S7.

Users can hover over a bubble to view detailed risk predictions for any given outcome, as shown in Fig. S8. By clicking on an outcome bubble, the interface will display the potential variables that contribute to the prediction of that outcome. Causal pathways are visualized with arrows pointing from causes to effects, where the thickness of the arrows represents the CDE values. The color of the arrows indicates the direction of the causal effect (red for positive and blue for negative). Additionally, users can click on any contributing variable to explore further variables that may influence the selected one, as shown in Fig. S9. This allows users to navigate through the causal pathways leading to the prediction of the outcome of interest.

C.3 More examples of interpretation

We additionally provide more examples of interpretation results in Fig. S10, Fig. S11, Fig. S12, and Fig. S13 for patients with circulatory failure, death, sepsis, and no positive labels, respectively. Shown in the interface in Fig. S10, cDEEP based its decision of circulatory failure prediction on BUN, calcium, PEEP, etc. This indicates that these variables may be possible intervention points to prevent circulatory failure. If some of these variables are hard or impossible to intervene, users can explore further variables that may influence these variables.

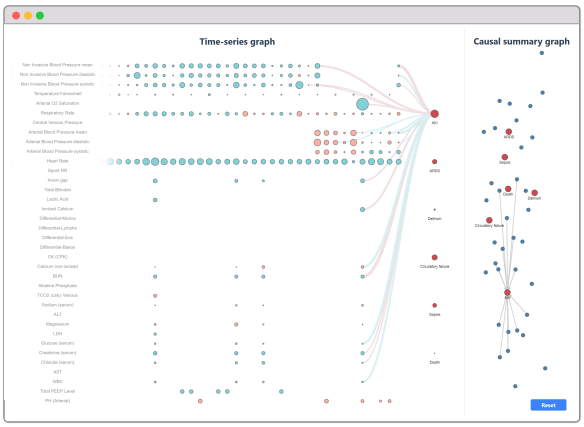

Figure S8: Clicked on one outcome.

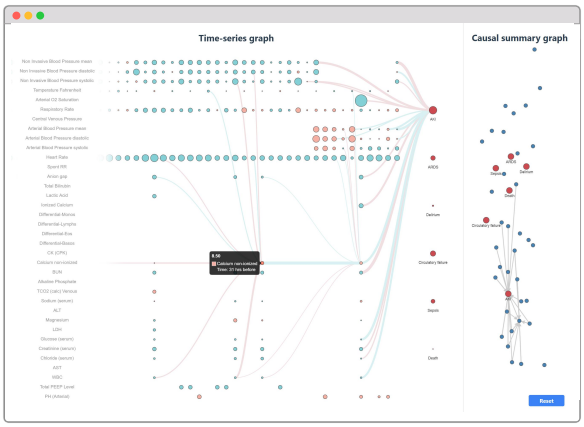

Figure S9: Clicked on one time-points of the variables.

Taking BUN as an example, users can click on BUN to explore further variables that may influence BUN, such as magnesium, glucose, and chloride.

It is imperative to note that the causal discovery results may not be strictly "causal" in the clinical sense, because of the limitation on the observational data. However, the causal discovery results can provide valuable insights into the potential relationships between variables and outcomes, guiding further investigations and interventions.

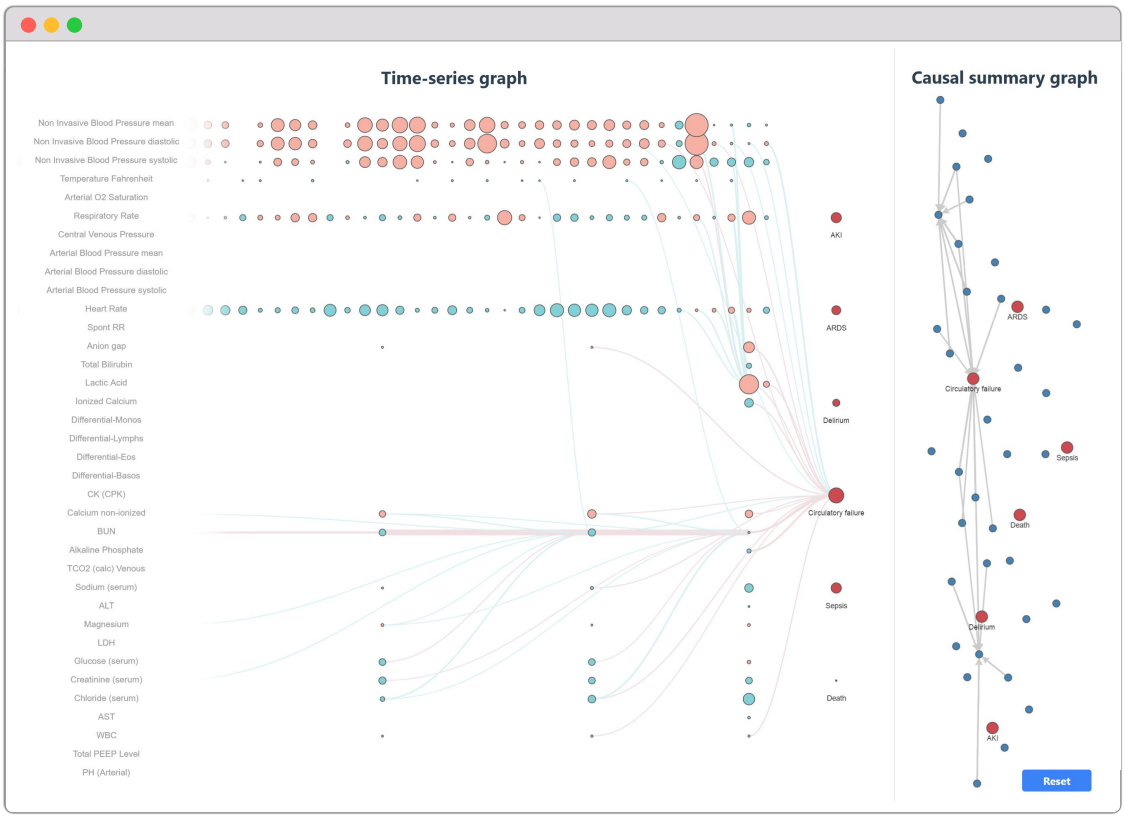

Figure S10: An example of interpretation with Circulatory failure-positive.

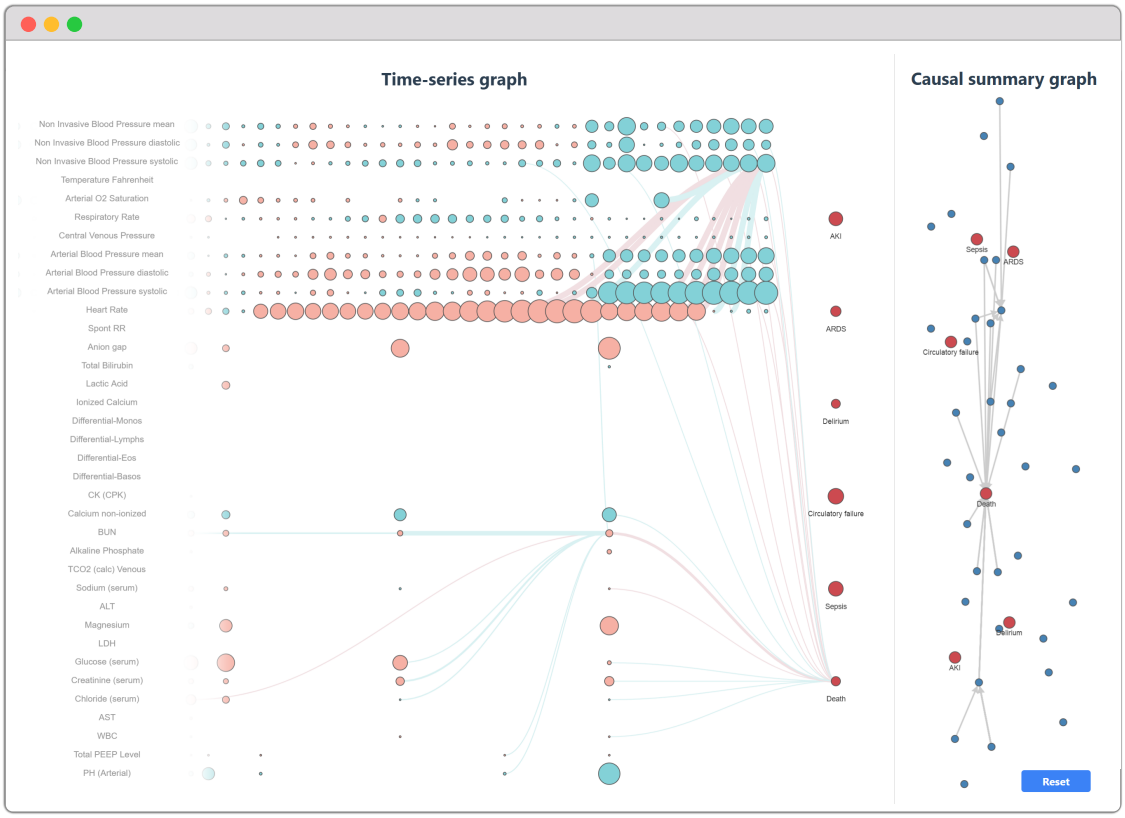

Figure S11: An example of interpretation with Death-positive.

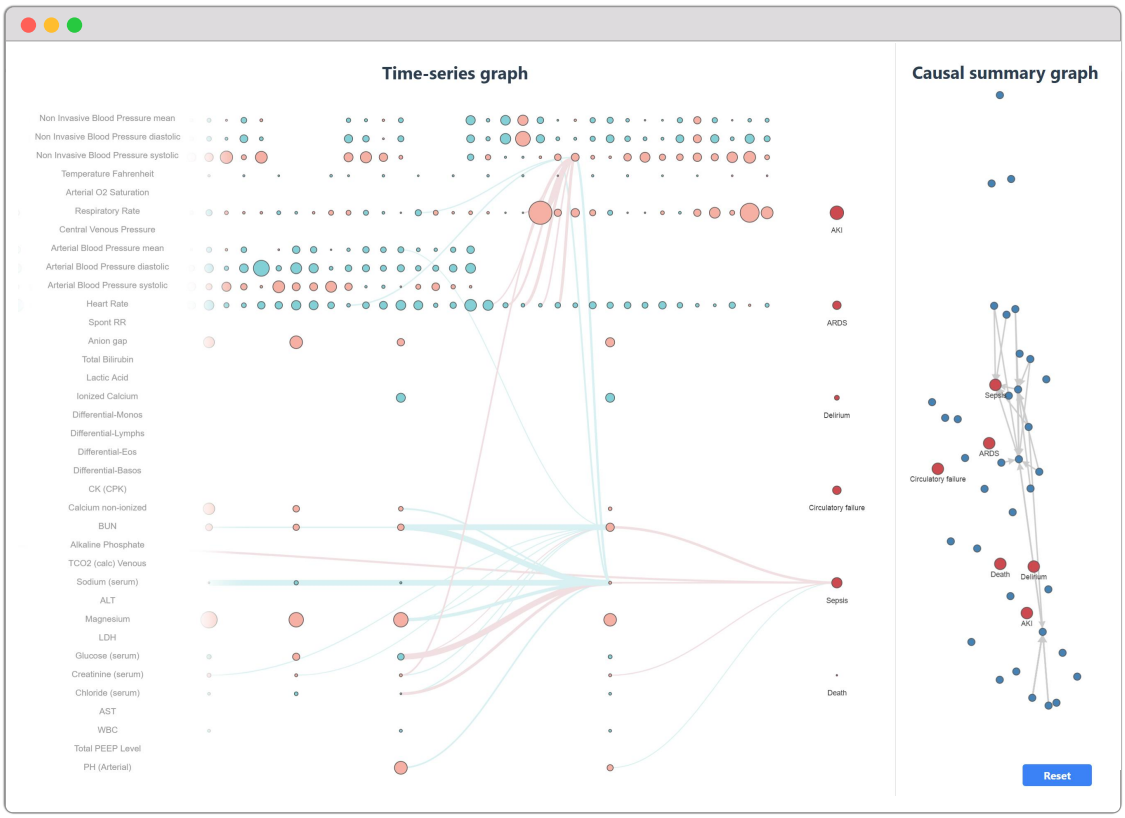

Figure S12: An example of interpretation with Sepsis-positive.

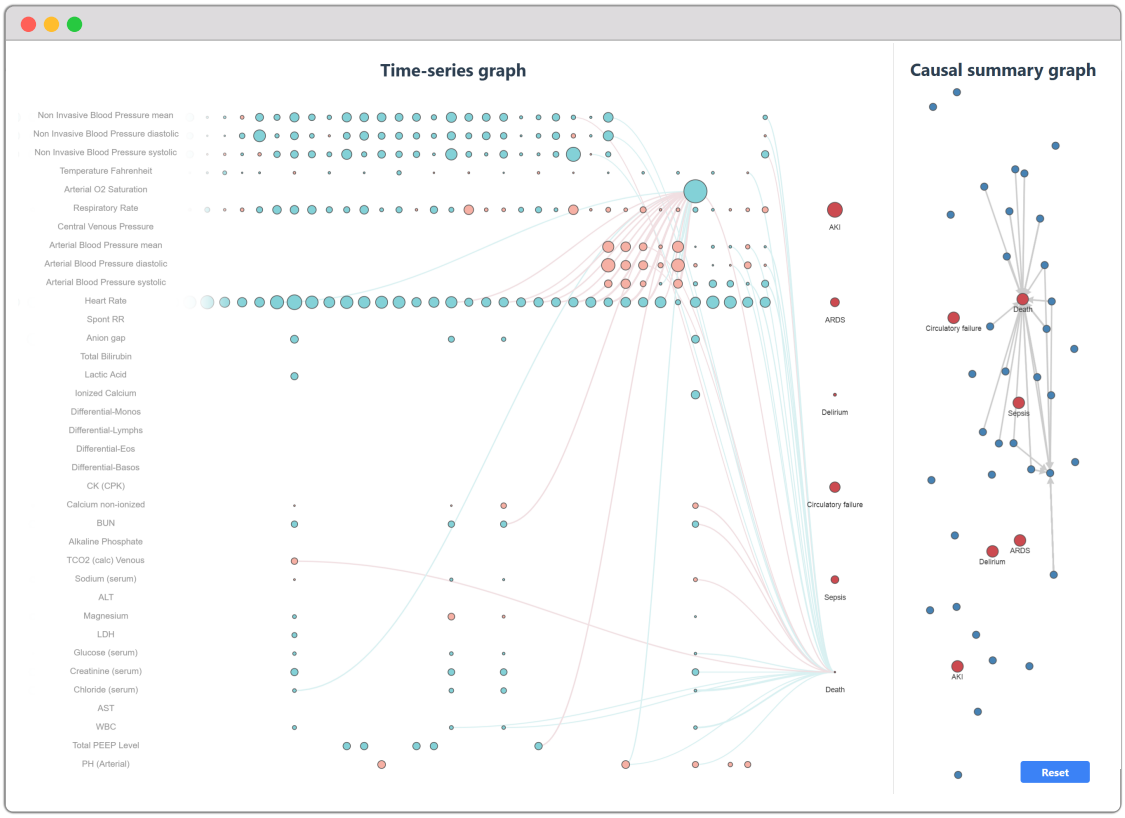

Figure S13: An example of interpretation with no positive labels.

D Implementation details

D.1 Data preprocessing

AKI/KDIGO labeling. AKI was defined based on the modified KDIGO (Kidney Disease: Improving Global Outcomes) [1] criteria. Stage I (mild) is defined as an increase in serum creatinine of $\geq$ 0.3 mg/dL or 1.5–1.9 times the baseline, or a urine output of <0.5 mL/kg/h for 6–12 hours; Stage II (moderate) is characterized by an increase in serum creatinine of 2.0–2.9 times the baseline, or urine output <0.5 mL/kg/h for $\geq$12 hours; and Stage III (severe) involves an increase in serum creatinine of $\geq$3.0 times the baseline, or reaching $\geq$4.0 mg/dL, or the initiation of renal replacement therapy, or urine output <0.3 mL/kg/h for $\geq$24 hours, or anuria for $\geq$12 hours. When prior measurements were available, we used a baseline of median annualized creatinine; in the absence of such measurements, we estimated baseline creatinine using the 'Modification of Diet in Renal Disease' formula. In total, we established 12 predictive goals, concentrating on predicting three categories of AKI (all, moderate/severe, and severe) over future intervals of 6, 12, 24, and 48 hours, aiming to enhance the evaluation of REACT's efficacy and expand its applicability.

Circulatory failure labeling. We annotated circulatory failure based on lactate and MAP [32]. Specifically, a patient was classified as currently not in circulatory failure if their MAP was greater than 65 mmHg or the patient was receiving vasopressors or inotropes, and lactate levels were $\leq$2 mmol/l. Conversely, a patient was labeled as currently in circulatory failure if their MAP fell to 65 mmHg or below and was not receiving vasopressors or not receiving vasopressors or inotropes, or if their lactate levels exceeded 2 mmol/l.

ARDS labeling. ARDS was defined based on the Berlin criteria [80]. The criteria for ARDS diagnosis include the presence of bilateral opacities on chest imaging, and a PaO2/FiO2 ratio of $\leq$300 mmHg with PEEP $\geq$5 cm H2O.

Other than the above three, the remaining predictive goals are already defined in MIMIC and eICU datasets, i.e., SOFA score for sepsis, CAM-ICU for delirium, and Death.

D.2 Loss function

In our implementation, two different loss functions are utilized for the V2V and V2O models. The V2V model employs the Masked Mean Squared Error loss function, which is defined as follows:

$$\mathcal{L}_{\text{V2V}}(\hat{x}, x) = \frac{\sum_{i=1}^{N}(\hat{x}_i - x_i)^2 \cdot (1 - m_i)}{\sum_{i=1}^{N}(1 - m_i)} \tag{1}$$

Here, $\hat{x}$ represents the model's predictions, x denotes the ground truth values, and m_i indicates the missing feature for each sample i. This formulation allows the loss to focus on the valid features by excluding those marked as missing.

For the V2O model, the Multitask focal loss function is employed, utilizing the known Focal Loss [41] to address the imbalanced data problem. The overall loss is computed as:

$$\mathcal{L}_{\text{V2O}}(\hat{y}, y) = \sum_{j=1}^{M}\sum_{i=1}^{N} \text{FocalLoss}(\hat{y}_{j,i}, y_{j,i}) \tag{2}$$

where $\hat{x}$ and x contain the predictions and corresponding labels for multiple tasks j, with N representing the number of samples for each task. Here Focal Loss is defined as:

$$\text{FocalLoss}(\hat{y}, y) = -\alpha y(1-\hat{y})^{\gamma}\log(\hat{y}) - (1-\alpha)(1-y)\hat{y}^{\gamma}\log(1-\hat{y}) \tag{3}$$

where α and γ are hyperparameters that control the balance between the positive and negative classes, as well as the focus on hard-to-classify samples, respectively. The Focal Loss function is particularly effective in scenarios where the data is imbalanced, as it helps the model to focus on the minority class and improve the overall performance.

D.3 Model training

The proposed method is implemented using a multitask prediction framework, designed to handle both dynamic and static features. The dynamic feature dimensions are automatically determined based on the selected input, with a batch size set to 256 and a total of 50 epochs for training. The model utilizes a time series attention model as the encoder, accompanied by two-layer multilayer perceptrons (MLPs) comprising 32 hidden units as decoders. It leverages the Adam optimizer in conjunction with a step-learning rate scheduler. The initial learning rate is set to 1×10^{-3}, decreasing to 1×10^{-4} over the course of training.

For the training process, a subset of 1000 batches is drawn from the entire dataset for each training epoch, ensuring variability and robustness in the training phase. After 20 epochs, the dynamic graph is frozen based on a threshold of 0.9, ensuring that only significant relationships are maintained. This approach allows for the effective integration of causal discovery within the predictive modeling framework.

Two-stage training. In practice, the learning process consists of two alternating steps: one for optimizing the neural network and the other for optimizing the causal probability matrix. Specifically, the loss function for the first step is defined as:

$$\min_{\theta_j} \sum_{p=0}^{P} \sum_{j=0}^{M} \mathcal{L}_j \left(f_{\theta_j}(\mathbf{X}_p^{\text{Pa}(y_j;\mathcal{G}^{\text{v2o}})}), \mathbf{y}_p \right) \tag{4}$$

and the loss function for the second step is defined as:

$$\min_{\mathcal{G}} \sum_{p=0}^{P} \sum_{j=0}^{M} \mathcal{L}_j \left(f_{\theta_j}(\mathbf{X}_p^{\text{Pa}(y_j;\mathcal{G}^{\text{v2o}})}), \mathbf{y}_p \right) + \mathcal{R}(\mathcal{G}^{\text{v2o}}) \tag{5}$$

The first step is optimized by the standard back-propagation algorithm where the causal probability matrix is sampled by Bernoulli distribution, while the second step is optimized with the Gumbel-Softmax trick to ensure the differentiability of the sampling process.

Fintuning on full variable input. As is discussed in the main text, we finetuned cDEEP on all available input variables, which we refer to as "cDEEP-full" to maintain fairness when compared with other methods. The model is first trained with the above two-stage training process for 20 epochs, and then finetuned with an all-one causal graph for another 30 epochs. Other parameters are kept the same as the original cDEEP model. In this way, the discovered causal associations are kept while the model is allowed to learn from information contained in non-causal variables.

Table S8: Hyper parameters for model training included in the grid search. The parameters utilized in the ultimate model are denoted in bold.

Parameter	Range
Total epoch	35, **50**, 65
Epoch when frozen	**20**, 35, 50
Freeze threshold	**0.5**, 0.7, 0.9
# of MLP layers	1, **2**, 3
# of MLP hidden nodes	16, **32**, 64
Learning rate	$10^{-2} \rightarrow 10^{-3}$, $\mathbf{10^{-3}} \rightarrow \mathbf{10^{-4}}$, $10^{-4} \rightarrow 10^{-5}$
Weight decay	0, 10^{-3}, $\mathbf{2} \times \mathbf{10^{-4}}$, 10^{-5}
Batch size	512, **768**, 1024
λ	10^{-8}, $\mathbf{10^{-7}}$, 10^{-6}, 10^{-5}
Gumbel τ	$\mathbf{1} \rightarrow \mathbf{0.1}$, $0.1 \rightarrow 0.01$

References

[1] KDIGO. *Kidney International Supplements*, 2(1):1, March 2012.

[2] Christophe Adrie, Maxime Lugosi, Romain Sonneville, Bertrand Souweine, Stéphane Ruckly, Jean-Charles Cartier, Maité Garrouste-Orgeas, Carole Schwebel, Jean-François Timsit, Jean-François Timsit, Elie Azoulay, Yves Cohen, Maïté Garrouste-Orgeas, Lilia Soufir, Jean-Ralph Zahar, Christophe Adrie, Michael Darmon, Corinne Alberti, Christophe Clec'h, Adrien Français, Aurélien Vesin, Stephane Ruckly, Frederik Lecorre, Didier Nakache, Aurélien Vannieuwenhuyze, Bernard Allaouchiche, Claire Ara-Somohano, Laurent Argault, Agnès Bonadona, Caroline Bornstain, Lila Bouadma, Alexandre Boyer, Christine Cheval, Jean-Pierre Colin, Anne-Sylvie Dumenil, Adrien Descorps-Declere, Jean-Philippe Fosse, Rebecca Hamidfar-Roy, Samir Jamali, Hatem Khallel, Christian Laplace, Alexandre Lautrette, Thierry Lazard, Eric Le Miere, Maxime Lugosi, Guillaume Marcotte, Laurent Montesino, Bruno Mourvillier, Benoît Misset, Delphine Moreau, Etienne Pigné, Stéphane Ruckly, Bertrand Souweine, Carole Schwebel, Gilles Troché, Marie Thuong, Guillaume Thierry, Dany Toledano, Eric Vantalon, Caroline Tournegros, Loïc Ferrand, Nadira Kaddour, Boris Berthe, Kaouttar Mellouk, Veronique Deiler, Kelly Tiercelet, Sophie Letrou, Igor Théodose, Julien Fournier, and On behalf of the OUTCOMEREA study group. Persistent lymphopenia is a risk factor for ICU-acquired infections and for death in ICU patients with sustained hypotension at admission. *Annals of Intensive Care*, 7(1):30, March 2017.

[3] Mansur Aliyu, Fatema Tuz Zohora, Abubakar Umar Anka, Kashif Ali, Shayan Maleknia, Mohammad Saffarioun, and Gholamreza Azizi. Interleukin-6 cytokine: An overview of the immune regulation, immune dysregulation, and therapeutic approach. *International Immunopharmacology*, 111:109130, October 2022.

[4] David Alvarez Melis and Tommi Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[5] Krisha Amit Patel, Ansh Sethi, Emad Al Azazi, Caitlin McClurg, and Tumul Chowdhury. The role of heart rate variability in predicting delirium: A systematic review and meta-analysis. *Journal of Clinical Neuroscience*, 124:122–129, June 2024.

[6] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization, March 2020.

[7] Pierre Baldi and Peter J Sadowski. Understanding Dropout. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[8] David P. Basile, Melissa D. Anderson, and Timothy A. Sutton. Pathophysiology of Acute Kidney Injury. *Comprehensive Physiology*, 2(2):1303–1353, April 2012.

[9] Alexis Bellot, Kim Branson, and Mihaela van der Schaar. Neural graphical modelling in continuous-time: Consistency guarantees and algorithms. In *International Conference on Learning Representations*, February 2022.

[10] Zsigmond Benkő, Ádám Zlatniczki, Marcell Stippinger, Dániel Fabó, András Sólyom, Loránd Erőss, András Telcs, and Zoltán Somogyvári. Complete Inference of Causal Relations between Dynamical Systems, February 2020.

[11] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[12] Edward De Brouwer, Adam Arany, Jaak Simm, and Yves Moreau. Latent Convergent Cross Mapping. In *International Conference on Learning Representations*, March 2021.

[13] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable Deep Models for ICU Outcome Prediction. *AMIA Annual Symposium Proceedings*, 2016:371–380, February 2017.

[14] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 883–892. PMLR, July 2018.

[15] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, August 2016. Association for Computing Machinery.

[16] Yuxiao Cheng, Runzhao Yang, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. CUTS: Neural Causal Discovery from Irregular Time-Series Data. In *The Eleventh International Conference on Learning Representations*, February 2023.

[17] Ian C. Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *The Journal of Machine Learning Research*, 22(1):9477–9566, 2021.

[18] Yuntao Du, Jindong Wang, Wenjie Feng, Sinno Pan, Tao Qin, Renjun Xu, and Chongjun Wang. AdaRNN: Adaptive Learning and Forecasting of Time Series, August 2021.

[19] Joseph Futoma, Sanjay Hariharan, and Katherine Heller. Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1174–1182. PMLR, July 2017.

[20] Kartik Ganesh, R. N. Sharma, Jaicob Varghese, and M. G. K. Pillai. A profile of metabolic acidosis in patients with sepsis in an Intensive Care Unit setting. *International Journal of Critical Illness and Injury Science*, 6(4):178, 2016-10/2016-12.

[21] Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. In *Advances in Neural Information Processing Systems*, volume 33, pages 12615–12625. Curran Associates, Inc., 2020.

[22] Elisa Gouvea Bogossian, Joaquin Cantos, Anita Farinella, Leda Nobile, Hassane Njimi, Giacomo Coppalini, Alberto Diosdado, Michele Salvagno, Fernando Oliveira Gomes, Sophie Schuind, Marco Anderloni, Chiara Robba, and Fabio Silvio Taccone. The effect of increased positive end expiratory pressure on brain tissue oxygenation and intracranial pressure in acute brain injury patients. *Scientific Reports*, 13(1):16657, October 2023.

[23] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[24] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120, December 2019.

[25] Raquel Herrero, Gema Sánchez, Iris Asensio, Eva López, Antonio Ferruelo, Javier Vaquero, Laura Moreno, Alba de Lorenzo, Rafael Bañares, and José A. Lorente. Liver–lung interactions in acute respiratory distress syndrome. *Intensive Care Medicine Experimental*, 8(1):48, December 2020.

[26] Brian L. Hill, Robert Brown, Eilon Gabel, Nadav Rakocz, Christine Lee, Maxime Cannesson, Pierre Baldi, Loes Olde Loohuis, Ruth Johnson, Brandon Jew, Uri Maoz, Aman Mahajan, Sriram Sankararaman, Ira Hofer, and Eran Halperin. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. *British Journal of Anaesthesia*, 123(6):877–886, December 2019.

[27] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, July 2012.

[28] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[29] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[30] Yeping Hu, Xiaogang Jia, Masayoshi Tomizuka, and Wei Zhan. Causal-based Time Series Domain Generalization for Vehicle Intention Prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7806–7813, May 2022.

[31] Zenan Huang, Haobo Wang, Junbo Zhao, and Nenggan Zheng. iDAG: Invariant DAG Searching for Domain Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19169–19179, 2023.

[32] Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3):364–373, March 2020.

[33] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456. PMLR, June 2015.

[34] Saurabh Khanna and Vincent Y. F. Tan. Economy statistical recurrent units for inferring nonlinear granger causality. In *International Conference on Learning Representations*, March 2020.

[35] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, pages 5815–5826. PMLR, July 2021.

[36] Harold William Kuhn and Albert William Tucker. *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press, March 2016.

[37] Simon Meyer Lauritsen, Mads Kristensen, Mathias Vassard Olsen, Morten Skaarup Larsen, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, and Bo Thiesson. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications*, 11(1):3852, July 2020.

[38] Lei Le, Andrew Patterson, and Martha White. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[39] Dongze Li, You Chen, Hong Liu, Yu Jia, Fanghui Li, Wei Wang, Jiang Wu, Zhi Wan, Yu Cao, and Rui Zeng. Immune dysfunction leads to mortality and organ injury in patients with COVID-19 in China: Insights from ERS-COVID-19 study. *Signal Transduction and Targeted Therapy*, 5(1):1–3, May 2020.

[40] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive Domain Expansion Network for Single Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2021.

[41] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[42] Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian Invariant Risk Minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16021–16030, 2022.

[43] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning Causal Semantic Representation for Out-of-Distribution Prediction. In *Advances in Neural Information Processing Systems*, volume 34, pages 6155–6170. Curran Associates, Inc., 2021.

[44] Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pages 509–525. PMLR, June 2022.

[45] Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, Xiangyang Ji, Qiang Yang, and Xing Xie. Diversify: A General Framework for Time Series Out-of-Distribution Detection and Generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4534–4550, June 2024.

[46] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[47] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality Inspired Representation Learning for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8046–8056, 2022.

[48] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain Generalization using Causal Matching. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7313–7324. PMLR, July 2021.

[49] Paul E. Marik, Xavier Monnet, and Jean-Louis Teboul. Hemodynamic parameters to guide fluid therapy. *Annals of Intensive Care*, 1(1):1, March 2011.

[50] John C. Marshall, Deborah J. Cook, Nicolas V. Christou, Gordon R. Bernard, Charles L. Sprung, and William J. Sibbald. Multiple Organ Dysfunction Score: A reliable descriptor of a complex clinical outcome. *Critical Care Medicine*, 23(10):1638, October 1995.

[51] George M. Matuschak and Jean E. Rinaldo. Organ interactions in the adult respiratory distress syndrome during sepsis: Role of the liver in host defense. *Chest*, 94(2):400–406, 1988.

[52] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222, May 2017.

[53] C. G. Morris and J. Low. Metabolic acidosis in the critically ill: Part 2. Causes and treatment. *Anaesthesia*, 63(4):396–411, April 2008.

[54] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. DYNOTEARS: Structure learning from time-series data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR, June 2020.

[55] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, August 2009.

[56] Lutz Prechelt. Automatic early stopping using cross validation: Quantifying the criteria. *Neural Networks*, 11(4):761–767, June 1998.

[57] Mattia Prosperi, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375, July 2020.

[58] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to Learn Single Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

[59] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, August 2016. Association for Computing Machinery.

[60] Jonathan G. Richens, Ciarán M. Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11(1):3923, August 2020.

[61] Mike D. Rinderknecht and Yannick Klopfenstein. Predicting critical state after COVID-19 diagnosis: Model development using a large US electronic health record dataset. *npj Digital Medicine*, 4(1):1–14, July 2021.

[62] Elisabeth D. Riviello, Willy Kiviri, Theogene Twagirumugabe, Ariel Mueller, Valerie M. Banner-Goodspeed, Laurent Officer, Victor Novack, Marguerite Mutumwinka, Daniel S. Talmor, and Robert A. Fowler. Hospital Incidence and Outcomes of the Acute Respiratory Distress Syndrome Using the Kigali Modification of the Berlin Definition. *American Journal of Respiratory and Critical Care Medicine*, 193(1):52–59, January 2016.

[63] Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1388–1397. PMLR, August 2020.

[64] Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, July 2023.

[65] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019.

[66] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*, September 2019.

[67] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization, April 2020.

[68] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer Nature, September 2019.

[69] Roberto Santa Cruz, Fernando Villarejo, Celica Irrazabal, and Agustín Ciapponi. High versus low positive end-expiratory pressure (PEEP) levels for mechanically ventilated adult patients with acute lung injury and acute respiratory distress syndrome. *Cochrane Database of Systematic Reviews*, (3), 2021.

[70] Patrick Schwab and Walter Karlen. CXPlain: Causal Explanations for Model Interpretation under Uncertainty. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[71] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[72] Farah Shamout, Tingting Zhu, and David A. Clifton. Machine Learning for Clinical Outcome Prediction. *IEEE Reviews in Biomedical Engineering*, 14:116–126, 2021.

[73] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyv&#228, rinen, and Antti Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

[74] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3145–3153. PMLR, July 2017.

[75] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[76] Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT press, 2000.

[77] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting Causality in Complex Ecosystems. *Science*, 338(6106):496–500, October 2012.

[78] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. PMLR, July 2017.

[79] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B. Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2022.

[80] The ARDS Definition Task Force*. Acute Respiratory Distress Syndrome: The Berlin Definition. *JAMA*, 307(23):2526–2533, June 2012.

[81] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, September 2017.

[82] Nenad Tomašev, Xavier Glorot, Jack W. Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, Alistair Connell, Cían O. Hughes, Alan Karthikesalingam, Julien Cornebise, Hugh Montgomery, Geraint Rees, Chris Laing, Clifton R. Baker, Kelly Peterson, Ruth Reeves, Demis Hassabis, Dominic King, Mustafa Suleyman, Trevor Back, Christopher Nielson, Joseph R. Ledsam, and Shakir Mohamed. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, August 2019.

[83] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training Deep Networks With Synthetic Data: Bridging the Reality Gap by Domain Randomization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 969–977, 2018.

[84] C. Trillo-Alvarez, R. Cartin-Ceba, D. J. Kor, M. Kojicic, R. Kashyap, S. Thakur, L. Thakur, V. Herasevich, M. Malinchoc, and O. Gajic. Acute lung injury prediction score: Derivation and validation in a population-based sample. *European Respiratory Journal*, 37(3):604–609, March 2011.

[85] Peter Van Dyken and Baptiste Lacoste. Impact of Metabolic Syndrome on Neuroinflammation and the Blood–Brain Barrier. *Frontiers in Neuroscience*, 12, December 2018.

[86] Dimitrios Velissaris, Vassilios Karamouzos, Charalampos Pierrakos, Diamanto Aretha, and Menelaos Karanikolas. Hypomagnesemia in Critically Ill Sepsis Patients. *Journal of Clinical Medicine Research*, 7(12):911–918, December 2015.

[87] J.-L. Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, C. K. Reinhart, PeterM Suter, and Lambertius G. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure: On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine (see contributors to the project in the appendix). 1996.

[88] Vy Vo, Van Nguyen, Trung Le, Quan Hung Tran, Reza Haf, Seyit Camtepe, and Dinh Phung. An Additive Instance-Wise Approach to Multi-class Model Interpretation. In *The Eleventh International Conference on Learning Representations*, February 2023.

[89] Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? A survey on structure learning and causal discovery, March 2021.

[90] Katherine D. Wick, Michael A. Matthay, and Lorraine B. Ware. Pulse oximetry for the diagnosis and management of acute respiratory distress syndrome. *The Lancet Respiratory Medicine*, 10(11):1086–1098, November 2022.

[91] Writing Group for the Alveolar Recruitment for Acute Respiratory Distress Syndrome Trial (ART) Investigators. Effect of Lung Recruitment and Titrated Positive End-Expiratory Pressure (PEEP) vs Low PEEP on Mortality in Patients With Acute Respiratory Distress Syndrome: A Randomized Clinical Trial. *JAMA*, 318(14):1335–1345, October 2017.

[92] Alexander P. Wu, Rohit Singh, and Bonnie Berger. Granger causal inference on dags identifies genomic loci regulating transcription. In *International Conference on Learning Representations*, March 2022.

[93] Ji Q. Wu, Nanda Horeweg, Marco de Bruyn, Remi A. Nout, Ina M. Jürgenliemk-Schulz, Ludy C. H. W. Lutgens, Jan J. Jobsen, Elzbieta M. van der Steen-Banasik, Hans W. Nijman, Vincent T. H. B. M. Smit, Tjalling Bosse, Carien L. Creutzberg, and Viktor H. Koelzer. Automated causal inference in application to randomized controlled clinical trials. *Nature Machine Intelligence*, pages 1–9, April 2022.

[94] Hao Ye, Ethan R. Deyle, Luis J. Gilarranz, and George Sugihara. Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific Reports*, 5(1):14750, October 2015.

[95] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. INVASE: Instance-wise Variable Selection using Neural Networks. In *International Conference on Learning Representations*, February 2022.

[96] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization Without Accessing Target Domain Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2100–2110, 2019.

[97] Joe G. Zein, Chao-Ping Wu, Amy H. Attaway, Peng Zhang, and Aziz Nazha. Novel Machine Learning Can Predict Acute Asthma Exacerbation. *Chest*, 159(5):1747–1757, May 2021.

[98] Borui Zhang, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Bort: Towards Explainable Neural Networks with Bounded Orthogonal Constraint. In *International Conference on Learning Representations*, February 2023.

[99] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[100] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to Generate Novel Domains for Domain Generalization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 561–578, Cham, 2020. Springer International Publishing.

[101] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. In *International Conference on Learning Representations*, 2017.