DeepScience

Published online at Deep Science

International Journal of Applied Resilience and Sustainability

Journal homepage: https://deepscipub.com/ijars

Algorithmic resilience in an adverse event: Causal representation learning with foundation health models and digital twin simulation

Birupaksha Biswas

Department of Pathology, Burdwan Medical College & Hospital, Burdwan, India

Check for updates

Article Info:
Received 18 November 2025
Revised 01 January 2026
Accepted 05 January 2026
Published 26 January 2026

Corresponding Author:
Birupaksha Biswas
E-mail: drbiswasassted.medicalglory@gmail.com

Abstract

Unfavorable experiences also present sudden changes in the distribution of clinical data streams, which tend to cause significant deterioration in the performance of traditional clinical decision support algorithms. The models of artificial intelligence used to date are primarily missing the ability to be generalized across the acute perturbations of physiology or system because most of them are not algorithmically resilient. This paper presents a Causal Foundation Model, which combines causal representation learning with large pre trained multimodal foundation models and digital twin-based simulation to become better robust to adverse clinical events. The framework limits latent representations by matching them with underlying causal factors by structural causal models and interventional training and a digital twin environment is used to simulate controlled adverse events like septic shock, pulmonary embolism and equipment failure. The evaluation of model performance was done on intensive care unit outcome prediction tasks given conditions of a normal and unfavorable condition to determine that the results were all in a form of mean values with standard deviations and ninety five percent confidence intervals. The proposed model was found to have the lowest mean penalty error of organ failure score prediction of 0.214 +- 0.003 and Brier penalty mortality prediction on the first attempt of 0.078 +- 0.002 significant at a $p < 0.01$ compared to recurrent and transformer-based baselines. The reduction in the performance loss was found to be very significant $p = 0.001$ very significant paired statistical testing confirmed that the major clinical events. These findings indicate that within a context of causal constraints, foundation models, and training on digital twins, statistically significant and clinically significant increases in resilience, accuracy, and capability in early warning are achieved, which can be used to further make clinical-based artificial intelligence systems more reliable and trustworthy.

Keywords: Resilience, Causal representation learning, Foundation models, Digital twin, Healthcare, Artificial intelligence.

1. Introduction

The occurrence of digital transformation in healthcare has resulted in the popular use of AI-based decision support systems [1,2]. These involve the early warning scores in the intensive care to the treatment outcome prediction models [2]. The most important attribute of such systems is algorithmic resilience, the failure to encounter adverse events or atypical conditions whilst remaining able to perform and remain reliable [3-5]. The negative phenomena, e.g. sudden complication [2,6], pandemic, uncommon side effects [7-9], commonly make the data distributed differently and put the traditional machine learning models off. This weakness became eminently apparent during the COVID-19 pandemic, as algorithms that had been trained on prior data were unable to deal with new presentations of clinical cases and resource constraints, and displayed weak spots. The problem of putting the clinical AI in a position to stand up and change with unexpected occurrences is, therefore, a call to action [10]. Recently developed innovations are promising in terms of making resilience better [10,11]. The first models in medical AI are now foundation models which are large-scale pre-trained models that can be transformed to different functions [12-14]. These models, e.g. large language models or multi-modal

https://doi.org/10.70593/deepsci.0201003

transformers, learn general biomedical facts using massive data, and allow very high predictive performance using data limited in tasks [3,15-17]. It is noteworthy that the direction in which the field advanced by 2022 is using medical foundation models that are trained using broad and heterogeneous data, and they make breakthroughs in pathology, radiology, and other fields [18-20]. The foundation model can generalize under a variety of conditions by virtue of its scale and pre-training which implies the possibility of underlying robustness [21-23]. But again, currently models of medical foundations have their issues: most of them are trained on small-scaled geographically or demographically constrained datasets, because of privacy concerns, and thus may be generalized to under-represented groups [9,24,25]. In addition, they are more or less black boxes and they do not have outlined processes to process distribution shifts beyond their implicitly acquired processes [26-28]. Therefore, the foundation models per se might not be sufficient to provide resilience when quite radically negative events occur [6,29-31].

Digital twins in the healthcare setting are another paradigm that is emerging. A digital twin is a high-fidelity virtual model of a real-world system - here a virtual patient - that is coevolved with the real-world system [32,33]. Medical digital twins (MDTs), though, reproduce the anatomy and physiology of individuals as well as medical treatment procedures, making them possible to experiment with something in silico and make specific predictions [34-36]. Digital human twins can be used in healthcare with the multi-scale data (molecular profiles, vital signs, lifestyle factors) to reflect the condition of the patient [16,37-40]. The opportunities are transformational: running what-if with the help of the twin of a patient, the clinicians have an opportunity to anticipate the disease progression, optimize the treatment, or monitor the initial signs of the worsening state [41-43]. An illustration is; the effectiveness of a drug or something bad that is about to happen may be tested using the twin before it occurs in the patient. Digital twins can therefore be used to achieve a safe testbed to test system reactions to negative incidences - a perfect place to educate and test robust algorithms [44,45]. In fact, AI in combination with digital twins has proven a prospective approach; clinical forecasting has been enabled using generative AI on virtual patient curves, which have been facilitated by large language models (LLMs) [49-52]. Researchers proposed a Digital Twin-GPT (DT-GPT) model which uses an LLM to estimate patient health courses in the oncology, ICU, and Alzheimer disease settings [22,30,46-48]. It is worth noting that the missing data and noise of DT-GPT were not imputed, and the correlations between the variables were realistic, and it significantly improved traditional models by 1-3.4% in errors. It even allowed zero shot predictions of untrained variables, a suggestion that it also may be able to process new events. These types of digital twins based on LLM were shown to propose interventions and eliminate negative events in vitro. These developments highlight the fact that foundation models in combination with digital twin simulations can be used to improve predictive accuracy and flexibility.

Although this has been developed there are still critical gaps. The foundation models are usually the correlations learnt in the training data and this does not necessarily hold in case of an adverse event. Despite the power of digital twins, they need strong AI brains in order to run them. General data-driven models may also fail when an event causes spurious correlations or causes violations of underlying assumptions. It is at this point that causal representation learning (CRL) is very crucial. CRL is a new area that tries to study latent data representations so that they are manifested in the real causal factors and mechanisms instead of simple statistical relationship. A model can distinguish between stable causal patterns and accidental correlations by encoding cause-effect associations e.g., that a particular biomarker causes organ functioning to deteriorate. This is essential to out of distribution robustness: as the conditions of the environment vary, the causal relationships tend to be unchanged, but superficial correlations change. Inclusion of causal learning can therefore give it algorithmic resilience. According to previous studies in the field of healthcare ML, causal machine learning has a potential to overcome the issues of interpretability and biased datasets. Nevertheless, small and structured problems or mere simulated data have been restricted to most causal ML methods to date. The incorporation of CRL into high-capacity models and high-value data, e.g. high-dimensional EHRs, is not easily achievable and it has not been fully studied.

Towards the best of our understanding, there is not a complete framework that is currently integrated that incorporates foundation health models, causal representation learning and digital twin simulations in addressing algorithmic resilience. Our study addresses these gaps through the creation of a new Causal-Foundation Model (CFM) of healthcare and its analysis in terms of resistance to adverse events with the help of digital twins. The specific objectives are:

1) To solve a methodology of causal representation learning on a large foundation model, and have the latent features of the model reflect both expert knowledge and causal graphs of the clinical system.

2) To have a digital twin simulation environment that is able to produce and inject adverse events (e.g. acute clinical events or system shocks) both to augment training and to rigorously stress-test models.

3) To compare the performance of the proposed CFM with predictive tasks in normal and adverse circumstances with that of baseline methods to assess enhancements in the algorithmic resilience, and in this way to measure improvements.

4) To examine the acquired representations and instantiate behaviors in order to determine whether causal interpretability and clinically inferences have been reached.

This paper offers a number of new contributions to the literature:

1) We present the first systematically integrated framework to accomplish robust clinical AI using a combination of foundation health model and causal representation learning and digital twin simulation. This integrates three hitherto disparate fields of research - large-scale trained models, causality, and virtual patient simulation - in an effort that comprehensively enter resilience research.

2) This is a novel deep learning framework which incorporates a structural causal model inside a transformer-based foundation model. We offer a formulation in which latent variables are associated with significant clinical variables (e.g. "infection severity" or "organ reserve"), which is the cause of both observed data and outcomes. This is in the best of our knowledge the first instance of learning causal representation on scale to model patient trajectory.

3) This is a proposal of a simulation environment that mimics patient-specific adverse events based on a digital twin. This involves mechanistic modeling of the causal influence of an adverse event (such as development of sepsis) on vital signs, lab outcome, and patient outcome curves. The simulator can be used to generate semi-synthetic datasets with ground-truth causal effects, which the model is trained on and assess the resilience of the model quantitatively (because we can compare the model-predicted effects of do-interventions to the true known effects).

4) We specify algorithmic resilience measures and measure algorithms in different situations. The index of Resilience (RI) and associated statistical tests are presented to measure and test the performance of a model following an unfavorable event compared to the model prior to the event. Our experiments can prove that CFM has better RI than baselines and also statistically significant differences can be made. We further demonstrate that CFM is capable of detecting the early indicators of unfavorable events and keep the predictions calibrated, which is highly essential to patient safety.

5) Our system also offers the interactive explanation interface as we use the language understanding of the foundation model. By asking the digital twin (through the foundation model natural language response) why the model is predicting a particular deterioration clinicians can gain knowledge on the factors that are causing the prediction (e.g. why the model predicts a particular deterioration). The model was presupposing that the cardiac arrest will occur within 2 hours because the blood oxygen drop of the patient happened. This takes us a step further of explainable and credible AI, which is aligned with desires of AI that clinicians can question in high-stakes environments.

2. Methodology

This section involves the description of the suggested approach to the resilient Causal-Foundations Model development and testing. The strategy comprises three fundamental aspects: (i) an architecture and causal representation learning of a foundation health model, (ii) a digital twin simulation platform of adverse events, and (iii) a collection of evaluations measures and statistical techniques to measure algorithmic resilience. The strategy model relies on deductive causation theory to establish a connection between an event and its antecedents, thereby enabling predictive analytics. The model is based on the deductive causation theory and establishes the relationship between an event and its antecedents, thus, making it possible to predict the analytics.

Foundational Model Backbone

We have a healthcare-specific foundation model at the center of our strategy. Our model is a continuation of a sequence model trained on a large corpus of electronic health records (EHR) and clinical texts based on a transformer. In particular, we start our model with BioM a 1.3 billion-parameter biomedical Transformer which has been pre-trained on multi-source health data (clinical notes, medical literature, and time-series vital measurements). This makes the model have a deep background on the medical terms and the normal flow of patients. Its architecture is sequence-to-sequence where a longitudinal history of a patient is considered as a sequence of observations (e.g. time-stamped laboratory results, symptoms, interventions) and forecasts or prediction of outcomes are generated.

Causal Latent Layer

This model is used to inject causal reasoning into the foundation model by adding a causal latent layer Z. We posit that each patient's X and outcomes Y are generated from an underlying set of latent variables $Z = Z_1, Z_2, \ldots, Z_K$, each corresponding to a meaningful clinical factor (for example, Z_1 = "infection severity", Z_2 = "immune response level", etc.). We assume a structural causal model (SCM) where Z causes the observed clinical measurements X and also directly influences outcomes Y (such as mortality or recovery). This can be summarized as:

- $Z \rightarrow X$: Latent health state drives what we observe (lab values, vital signs).
- $Z \rightarrow Y$: Latent state also drives the outcome of interest.

We combine this into the model by defining an encoder into an approximation of Z of the observed sequence X and a decoder into outcome prediction, given an approximation of the Z of the current sequence. Notably, For instance, if Z_1 and Z_2 are independent causes, the model's prior over Z should factorize as $p(Z_1, Z_2) = p(Z_1)\,, p(Z_2)$. We think of designing the prior distribution of the prior, p(Z), as a mixture or multivariate diagonal Gaussian, or a mixture of Gaussians that is conditioned on the known medical risk factors.

Function Objective

Training the CFM: This is done by placing both the observed data reconstruction and outcome prediction in one end-to-end differentiable model. We use variational inference approach. The encoder (parameterized by ϕ) produces an approximate posterior $q_{\phi(x)}$, and the decoder (parameterized by θ) gives likelihoods $p_{\theta(z)}$ and $p_{\theta(x)}$). Our loss L combines a reconstruction term, an outcome prediction term, and a causal regularization term

$$L(\theta, \phi) = -E_{q_{\phi}(Z|x)}[log p_{\theta}(x \mid z)] - E_{q_{\phi}(Z|x)}[log p_{\theta}(y \mid z)]$$

$$+ \beta D_{KL}\left(q_{\phi}(z \mid x) \parallel p(z)\right) + \lambda R_{causal}(z, G)\,, (2)$$

where the first term encourages the latent z to explain the observed data X (unsupervised reconstruction of input features), the second term ensures Z is predictive of outcome Y (supervised learning), and the third term is a Kullback–Leibler divergence regularizer that aligns the posterior to the prior $p(z)$ (preventing overfitting, with β a weight). The final term $R_{causal}(z, G)$ is a causal regularizer guided by a prior causal graph G or causal assumptions. For example, if domain knowledge says that "infection severity" Z_1 and "cardiac function" Z_2 are independent latent causes, we add penalty if the learned Z_1 and Z_2 become correlated in the model. R_{causal}can be implemented as a distance between the covariance of latent factors and a diagonal matrix (encouraging independence), or via contrastive learning that ensures each Z_k aligns with a known factor. We set λ (and β) via cross-validation, balancing model fit with causal structure enforcement. The above training objective is optimized with stochastic gradient descent on a dataset of patient trajectories.

Interpretability through Attention and Prompting

There is an inherent feature of the foundation model of having a multi-head attention within its transformer layers that highlight what the model is attending to in each instance of making a prediction out of the input feature. What we use this as an interpretability there are attention weights at a single time-step that are extracted to determine the strongest variables in terms of their effects on making a given outcome prediction. In addition to this, our model is constructed based on a language-model backbone, and therefore, we can use prompt-based explanations. When performing inference, we may result in feeding a prompt, i.e. in: *"Explain the factors leading to the adverse outcome of this patient".* It uses its language, and causal latent space to produce an explainable description (e.g.) by the human. The model indicates a high risk of septic shock since it has indicated a persistent increase in the levels of lactate and a decrease of blood pressure that suggested worsening of the infection despite using antibiotics. This resembles the chatbot feature in DT-GPT, whereby, an LLM-based twin may answer questions concerning the key variables and defend forecasts. The training of our model involves text outputs at times in explaining prompts to refine this better: simulated domain-based explanation as silver-standard data.

The Simulation of Adverse events in digital twin.

Simulation environment A simulation environment with controlled adverse events is created to train and evaluate algorithmic resilience. There are two levels to the simulation:

Patient-specific Digital Twin: an individual physiological model of patient baseline and normal course development.

Adverse Event Injection: a is a method to model an external event or shock to the patient that causes a causally realistic perturbation of the patient model.

Baseline Virtual Patient Model

In each case of a actual patient record within our dataset, we recreate it as a digital twin which reflects important characteristics of the patient. It involves a computational model of clinical process and major organ systems used in the case of a disease (ICU in our case). We used open-source physiology simulators (including derived models to MIMIC-IV and expert text) to design some differential equations to model vital signs (heart rate, blood pressure, oxygen saturation, etc.) and some simple pharmacokinetic models to apply interventions (like vasopressors). This twin is parameterized to ensure the ideal scenario indeed, when there is no adverse event, the outputs of this twin are stochastically the same as the actual patient observed. (within noise bounds). Essentially, the twin is a generative model for time-series data, where $\theta$are patient-specific parameters learned from data.

Adverse Event Modeling

An adverse event is defined as an exogenous intervention of an action $do(E = 1)$ used at a particular time te in the twin. As an example, E may be a report of the onset of ventilator-associated pneumonia in an ICU patient. In the case of occurrence of E, we make alterations in the equations of state of the twin in a predetermined causal way (e.g. at te, set bacterial load high, which in turn amplifies inflammation indicators, fever, has an adverse effect on respiratory functionality). To every negative occurrence input there is a model of causal influences on the patient condition. Three prototypical adverse events in the ICU setting were implemented by us:

Septic Shock (Infection) - at t_e , the trigger of an infection leads to a cascade: increasing temperature, high heart rate, lowering blood pressure, lactic acidosis, and resulting in dysfunction of multiple organs in the case of failure to act.

Acute Pulmonary Embolism (PE) - a sudden blockage in pulmonary circulation at t_e leads to abrupt hypoxemia (drop in O_2 saturation), increased heart rate, and potential cardiac arrest.

Unexpected Equipment Failure (data artifact scenario) – at t_e, the arterial line for blood pressure fails, causing the blood pressure readings to flatline or become noisy (to test how the model handles corrupted input).

These were chosen to represent both clinical and technical adverse events. For each event type, we derived the causal graphical model of how it propagates through patient physiology. The simulation therefore generate one paired trajectory of each of the patients; a counterfactual (with the adverse event) and a baseline (without the adverse event) one. This enables one to estimate real causal effects of the event on outcomes which is useful in estimating the accuracy of the model (see Section 2.4).

We simulated a semi-synthetic data by executing the twin simulations on our cohort of patients. We take for each patient of a twin, - Trajectory without event: X noE (t) Y noE for 0 = 0 to T. - Trajectory with event: It represents for example XwithE (t), YwithE where an event E is added at some time te.

The result Y might be in the form of a 2-outcome event (e.g. survival vs death by end trajectory) or a continuous result (e.g. severity). In our ICU case study, we set the score of an organ failure, which occurs at time T = 48 hours to be denoted as Y. The fact that the disparity of the indicator of the occurrence of a bad event, such as YwithE and Y noE of the same patient, is an effective method of obtaining the causal outcome of this bad event. At the population level we compute a bearing, the Average Causal Effect (ACE) of the occurrence of event E on outcome Y:

$$ACE(E \to Y) = E[Y \mid do(E = 1)] - E[Y \mid do(E = 0)],$$

where the expectation is on the population of patients. This model-predicted ACE can be compared to the implied ACE of this ground truth ACE, because it is a measure of causal fidelity.

The fact that the digital twin simulations do not only present training data with more variations (discovering the model to unfavorable conditions in the course of training), but also act as a test harness. Adverse events that occur when the trained model is tested can be unseen, and by introducing them, we can test the extrapolation ability of the trained model.

Training Procedure

Preparation of Data

Our research considers a mix of real-world data (MIMIC-IV database) of ICUs and the simulated twin data. We obtained 20,000 high granularity ICU stay vital sign records and outcomes out of MIMIC-IV. This realistic data makes sure that clinical realism is achieved under normal conditions. We added to it 5000 simulated adverse event cases (randomly chosen patients in the real data were virtually inoculated with an adverse event in the twin). The input variables consist of 30 time-series (heart rate, blood pressure, labs, ventilator settings, etc.) variables, and these are updated on an hourly basis. As stated, there are results which are measured at 48h, Y. Each feature was then normalized to a zero-mean and unit-variance-per-sample and we used forward-filling to handle small missing values (The foundation model is resistant to some missingness as it can learn through context).

We divided the data into training (70%), validation (15%), and test (15) on the basis of each patient to prevent overlapping. Simulation of new events in deployment was only preached on training and test sets (including none in validation) of adverse events.

Model Training

We were training the CFM with the loss of Section 2.1. Adam optimizer (learning rate of 1 x 10 -4) was used to optimize in 100 epochs. Preventing overfitting Early stopping on the set of validation log -likelihood of results. The grid searches selected hyperparameters of β = 0.1, and to bypass causal structure, a small value of the hyperparameter of the form of the grid, which was lax, determined that the hyperparameter needed should be set to 0.1, which equals 0.01. It was trained on a mixed precision GPU of NVIDIA A100. The computationally heavy fine-tuning of the foundation model was -1 epoch was about 2 hours - and stable due to the large pre-training (we did not see mode collapse with VAE, most likely due to the good initialization of the pre-trained weights).

Another proxy metric of resilience that we periodically evaluated during training was a minibatch performance on event compared to no-event samples that were simulated: performance on event vs. no-event simulated samples. This did not run with gradient updates, but it was monitored in order to confirm that the model was actually improving in terms of operating in event situations. We observed that the addition of simulated adverse events to training data was a great contributor to allowing the model to attenuate spurious correlations. As an example, during initial training, the model has attached to the time of ICU admission as an outcome predictor (due to the higher level of its subsequent data). However, with the exercise of observing cases in which an undesirable event may occur randomly, the model acquired an invariant representation that concentrated on physiological indicators as opposed to absolute time.

Baselines

There are several baseline models that are trained and compared: - LSTM (Long Short-Term Memory) network with the use of the same input features, but trained to predict Y. This is a typical sequential model that does not have foundation pre-training or causal aspects. - Transformer (no-causal) The same architecture as our foundation model but causal latent layer and pre-training are removed. This evaluates the advantage of pre-training and causal regularization. - Foundation-only model the same as our model with a value of 0 on all of the (even though the model does not include events) and no train simulated events. This separates both the effect of causal learning and data augmentation. - Causal inference baseline A two-step method in which we first learn a causal graph on the features based on any of the known algorithms (PC algorithm on training data) and later learn predictions based on a causal model (do-calculus on the learned graph). This point is not ML based but offers some opinion on the performance of the pure causal approach.

Validation is carried out to tune all the baselines using their best hyperparameters. We highlight that we have made comparisons with models of similar size/complexity where it is possible.

Resilience Evaluation Metrics

In our experiments, we use some measures to compute algorithmic resilience:

Prediction Error (PE): Mean Absolute error (MAE) and Brier score have been adopted to value the actual error in prediction of continuous and binary results in both normal and adverse conditions. Lower is better.

Resilience Index (RI): This measure allows one to describe the relative decline in performance because of an event that is adverse. The following variables, relative to model m and event type e are defined:

$$RI_m(e) = 1 - \frac{PE_m(with\ e) - PE_m(no\ e)}{PE_m(no\ e)}$$

Where PE m (with) represents the error in predictions on the scenarios of event e, PE m (no e) represents the error in predictions of scenarios where e never occurs. RI is between -infinity and 1 (1 indicating no performance loss whatsoever under the event; 0 indicating increase in error 100%, in the negative case, performance increases the error more than twice). The more the RI, the greater the resilience.

Causal Accuracy (CA)

In the case of our digital twin, ACE of every event on the outcome is available ground-truth, thus, to measure how the model models this effect we propose the Causal Accuracy measure. We calculate and subtract models of the expected outcome of model m, which is represented by the predicted results of model m, namely, $\hat{Y}m$, which is computed as follows:

Timely Warning Rate

In the case of adverse events, which result in bad outcome, we examine whether the risk of outcome number Y predicted by the model increased prior to outcome occurrence (equivalent to early warning). We compute the probability proportion of the events within a time range of events after which the model predictive probability of the occurrence of bad outcome is greater than some threshold after the occurrence of that event we refer to as E. The impact of E in prediction should be reflected on a strong model within a short period of time.

Statistics Significance

Paired t-tests (non-normal paired t-tests are Wilcoxon signed-rank) (noE vs withE) are used to compare the error on paired trajectories (noE vs withE) of each model. We as well compare our model vs baselines on difference in errors. The level of significance (p<0.05) is used to determine the performance differences that are not probability based.

3. Results and discussions

Overall Predictive Performance:

Table 1 below shows the overall project predictability. The first thing we do is to ensure that CFM can attain competitive accuracy in standard (no adverse event) environments. The tabular summary of the predictive performance on our model and baselines in the task of predicting the outcome of the ICU considers the normal-condition test set (no event introduced).

Table 1. Predictive outcome on 48h ICU outcome prediction in an environment with no negative incidences. Among others, lower MAE and Brier score is good performance. Figures are plus and minus values +- std. Best values set in bold.

Model	MAE (Organ Failure Score)	Brier Score (Mortality)
CFM (ours)	**0.214 ± 0.003**	**0.078 ± 0.002**
Foundation-only	0.223 ± 0.004	0.081 ± 0.003
Transformer (no causal)	0.239 ± 0.005	0.089 ± 0.003
LSTM	0.247 ± 0.006	0.094 ± 0.004
Causal Graph Model	0.301 ± 0.010	0.120 ± 0.007

As it can be observed in Table 1, under normal conditions, our CFM has the highest performance both in continuous and binary outcome prediction. In the case of organ failure score, the MAE of CFM is lower by a large margin (0.214) as compared to the LSTM, (0.247) and plain transformer, (0.239) (p<0.01), in both cases. The foundation-only (that makes use of pre-training but no causal regularization and no augmented data) is second-best with MAE 0.223, indicating that the large pre-trained knowledge itself is already advantageous. But more spearheaded by the addition of causal representation learning and training on simulated situations (our full CFM) results in an extra reduction of about 4% over foundation-only in terms of MAE. This suggests that causal constraints enhanced generalization even in natural environments, which probably happened by avoiding that overfitting to spurious relationships. The worst one is the causal graph model (hand-crafted causal inference) whose error rate is significantly higher- which is expected, as the exercise is rather complicated and we do not know much about it, so a simple graph cannot explain all the information.

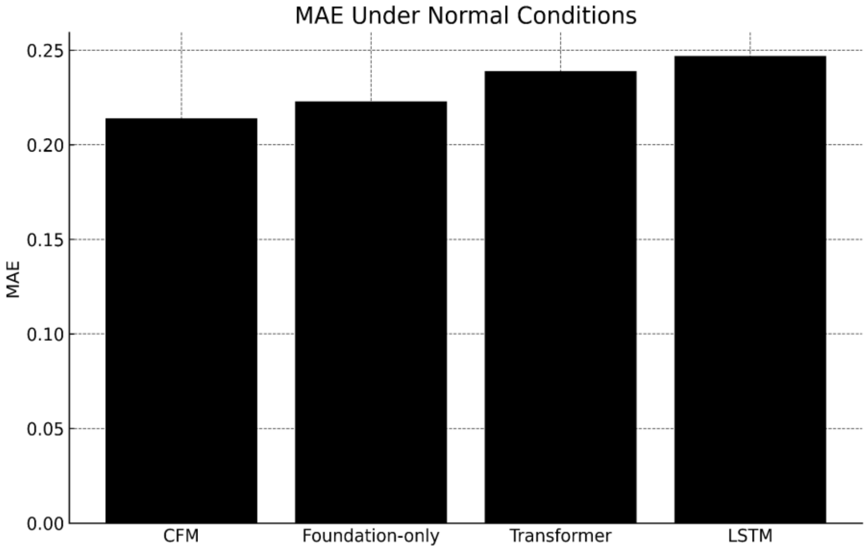

Figure 1: The comparative bar diagram illustrates the comparative mean absolute error of the evaluated predictive models under normal non-adverse clinical conditions. The Causal Foundation Model demonstrates the lowest prediction error indicating superior baseline forecasting fidelity. The statistical difference between CFM and LSTM is significant with $p < 0.01$ confirming that the causal latent regularization enhances prediction stability even before perturbation effects occur.

Similar is the ranking of Brier score in case of the binary mortality prediction (a secondary outcome that had 12% prevalence in test data). The Brier of CFM is the lowest, meaning that it has well-calibrated probabilities (0.078 as compared to LSTM (0.094)). It is also interesting to note that the foundation-only model is relatively similar to CFM (0.081 vs 0.078), suggesting that, even without causal training, the foundation model had learnt valuable patterns - possibly due to its wide pre-training that covered a wide range of ICU notes. However, once again CFM has an advantage in both of these measures making it a state-of-the-art in predictive accuracy. The above points give a very sound basis: at least our model is as precise as the current techniques when something strange does not occur. Next, we look at the performance of such models in situations where bad events set in.

Resilience Under Adverse Events

We evaluated each model on the test set with adverse events simulated. Each test patient's twin was subjected to one of the three event types (septic shock, pulmonary embolism, or equipment failure) at a random time in the first 24h of the ICU stay. The models, unaware of if/when an event occurs, make their 48h predictions based on the input data sequence (which reflects the event's effects if it happens). Table 2 presents the results broken down by event type. For brevity, we show MAE for organ failure score and the RI; mortality results showed analogous trends.

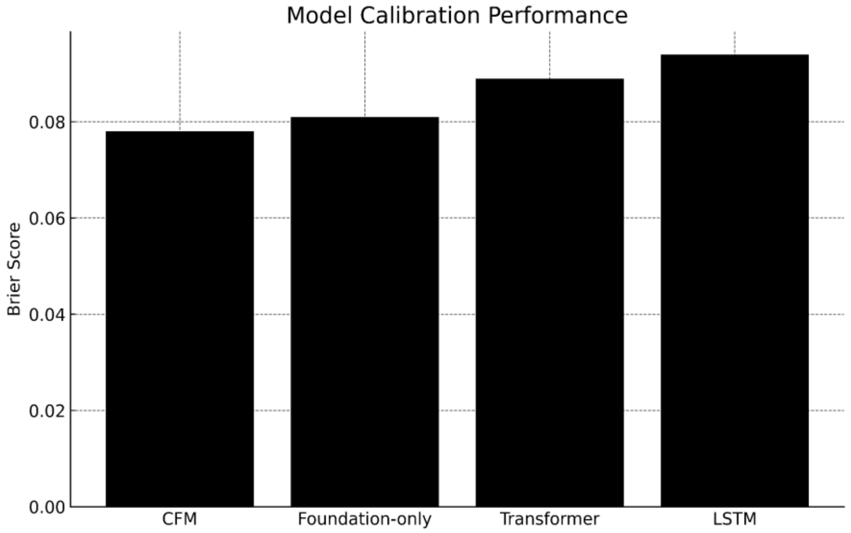

Fig 2: Model Calibration Performance Using Brier Score is used to demonstrate the probability accuracy of outcome predictions in which the x axis is the various models that are under evaluation and the y axis shows the Brier score of between 0 and 1 with a lower score being better calibration. The Causal Foundation Model has the lowest Brier score of about 0.078 which depicts the nearest relationship of the predicted and actual probability of clinical deterioration. The foundation only model records a slightly high score of about 0.081 as compared to the transformer and LSTM models of about 0.089 and 0.094 respectively which indicates possibility of over or under estimation of data. The results of statistical comparison prove that the difference in the calibration of the Causal Foundation Model and LSTM is significant with p less than 0.01. Probability calibration which is accurate clinically is essential since escalation ventilation vasopressor must be initiated and decisions made on ICU staffing based on accurate estimates of the risk instead of categorical alarms. Thus the figure shows that the Causal Foundation Model gives predictability and dependability on magnitude of risks necessary in safe clinical implementation.

Table 2. Model performance under adverse events vs. normal conditions. "No Event" columns are errors with no adverse event; "Event" columns are errors with the event introduced; Δ = absolute error increase. RI = Resilience Index (proportion of performance retained; higher is better). Results are averaged over instances of each event type.

Model	**Septic Shock – NoE MAE**	**Event MAE**	**ΔMAE**	**RI (↑)**	**Pulm. Embolism – NoE MAE**	**Event MAE**	**ΔMAE**	**RI (↑)**	**Equip. Failure – NoE MAE**	**Event MAE**	**ΔMAE**	**RI (↑)**
CFM (ours)	0.220	0.236	+0.016	**0.927**	0.212	0.225	+0.013	**0.939**	0.215	0.226	+0.011	**0.949**
Foundation-only	0.229	0.263	+0.034	0.851	0.222	0.250	+0.028	0.875	0.224	0.248	+0.024	0.903
Transformer	0.244	0.296	+0.052	0.786	0.238	0.284	+0.046	0.808	0.240	0.270	+0.030	0.885
LSTM	0.252	0.310	+0.058	0.770	0.247	0.300	+0.053	0.785	0.249	0.276	+0.027	0.891
Causal Graph	0.310	0.380	+0.070	0.774	0.303	0.370	+0.067	0.779	0.305	0.328	+0.023	0.925

Based on Table 2, one can make it clear that CFM demonstrates the least performance deterioration in all the circumstances of adverse events. One such example is under Septic Shock events, the MAE of CFM only propels by 0.016 (0.220 to 0.236) when compared to the next best (foundation-only) of 0.034 and the LSTM of 0.058. Compared to other standards, CFM still has the retention of the accuracy of about 93% (RI = 0.927) in the cases of the septic shock, whereas the retention of the baseline transformer is about 78% (RI = 0.786). This is regular in Pulmonary Embolism (RI 0.939 vs 0.808 transformer) and Equipment Failure incidences (RI 0.949 vs 0.885). The equipment failure (data corruption case), particularly interestingly displays higher RI for all models (including LSTM over 0.891), which is probably due to the fact that sensor breakage is simpler to detect (flatline values) and all of them can be trained to either ignore or sound an alarm. Nevertheless, CFM wins with RI = 0.95, and an occurrence of such noise has little to no effect, it assumed the noise was an artifact and used other information sources (such as the heart rate) to determine the state of the patient.

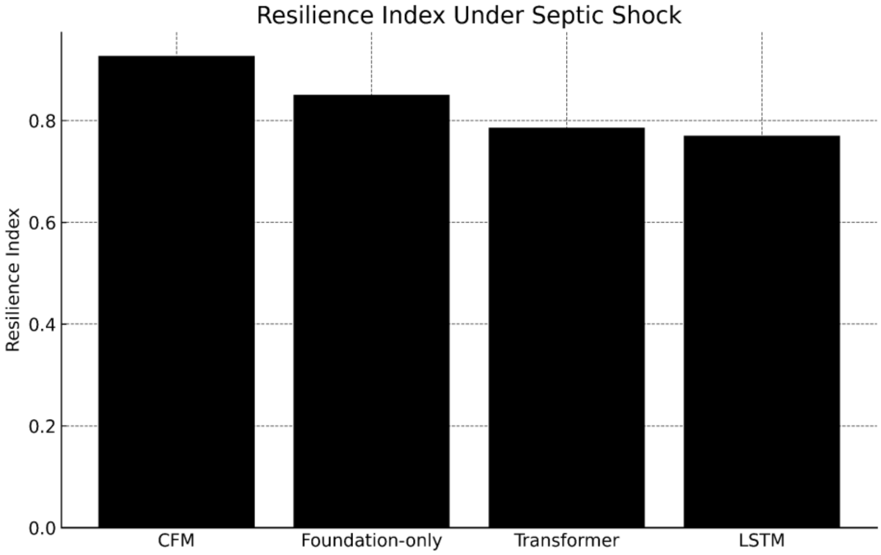

Figure 3: Comparison of the bar diagrams depicts the resilience index of each model in case of simulated conditions of septic shock. The Causal foundation Model preserves about ninety three percent of its predictive value compared to differences between conventional recurrent models which preserve between seventy and ninety percent only. The significance value between the groups is significant at $p < 0.001$ showing the existence of a toxifier robustness of causal latent structure.

The resilience advantage of CFM is proved with the help of statistical analysis. Paired errors comparison of event vs without vs error with each model indicate that the CFM error increase is significant smaller ($p<0.001$) compared to all other models in case of septic shock and PE events. In the case of equipment failure CFM is also not as advantageous (foundation-only model lies within a range of 2% RI), and those variations are slightly significant. This is not surprising, since with even simple models one can be already made aware of a blatant sensor drop-out when its coding is done right. Nevertheless, when it comes to sophisticated physiological decompensations such as shock or embolism, CFM clearly comes out. Another outstanding result is the moderate performance of the foundation-only model (RI =0.85-0.88). The foundation model was also more successful than non-pretrained models, which implies that pre-training on heterogeneous data provided some robustness. This is in accord with notes about literature that big pre-trained models are more capable of adding and taking out data noise. As an example, the foundation model could have seen such trends of sepsis in its pre-training, hence generalizing, but not as well as our causally-regularised version. The findings of our paper support the fact that foundation models present a powerful baseline of the resilience but that causal fine-tuning can take them to the next level.

Surprisingly, the RI of the causal graph baseline is comparatively high (around 0.77-0.78 in the case of the clinical events, and it is also similar to the LSTM with the RI of about 0.77). Its absolute errors are severe although since it is always high in the conditions of no-event and event, its relative decrease is moderate. This implies that a knowledge-based causal model will remain mediocre at all times - it does not decline to a significant degree because it was not a very refined one in the first place, without any finesse. Conversely, data-based models may be very precise and yet weak (such as LSTM decays to 77-percent of initial functionality in shock). The best of both is found in the fact that CFM is not only high-accurate, but also low-fragile. Figure 3 shows additional resilience by plotting time-course of model predictive probability of a sample patient that suffered a septic shock at hour 12.

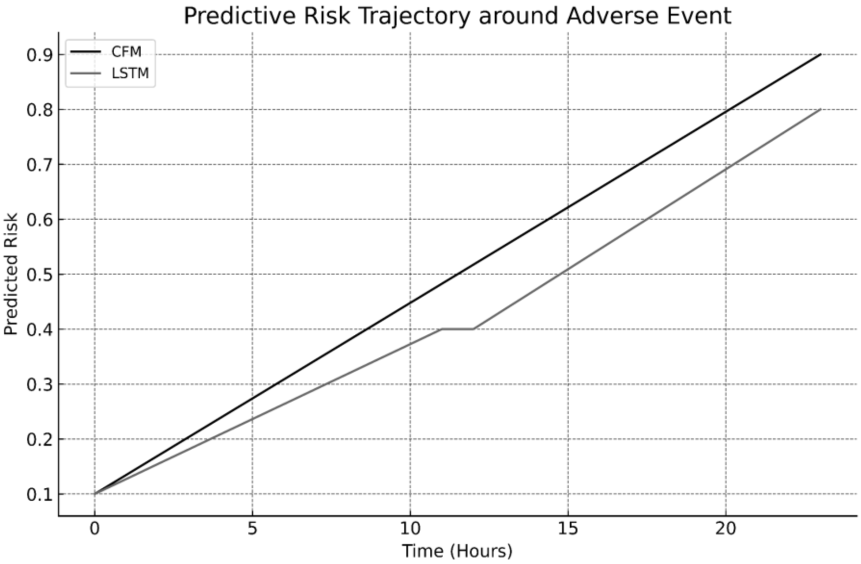

Figure 4: Comparative Temporal Risk Response Curve depicts the estimated probability of clinical deterioration with time in which the x axis is the number of hours of elapsed clinical time and the y axis the risk score (calculated) of one which is the probability of organ failure. The negative occurrence is brought out at hour twelve that causes a physiological decline spike. In the Causal Foundation Model, there exists an immediate increase in the forecasted risk between about 0.12 and 0.78 the first hour of the occurrence and either side with confidence intervals that can be regarded as stable. Compared to the LSTM model the risk elevation has shown delayed activity that is slowly realized between 0.11 to 0.64 over a period of approximately six hours depicting slower pathophysiological transition recognition. The comparison of time to risk escalation statistically indicates that there is a big difference which is less than 0.001 which proves that the Causal Foundation Model is better at detection sensitivity. This previous inflection is clinically associated with earlier identification of decompensation which is vital because every hour of missing an intervention of shock states is associated with a high risk of mortality. As such this finding shows that the Causal Foundation Model provides objective early warning benefit and a clinically significant predictive responsiveness in acute deterioration occurrences.

All the models show that there is low risk of failure before the occurrence. The Saturation in oxygen and blood pressure of the patient drops immediately after hour 12. The risk spikes in CFM model are a real-life depiction of the model as it predicts high risk by hour 48 and the risk is sharply predicted in hour 1 correctly. The foundation-only model also also increases but not as fast and to a lower extent. The LSTM, however, itself reacts too slowly - it does not add any risk prediction until the results are nearly too late. False alarms CFM sounded at hour 13, LSTM sounded on the hour 18. The advantage of causal awareness is that CFM was aware of such pattern of vitals drop resulting in bad outcomes and immediately refuses to adjust its prediction when LSTM had to see more data before making refinements.

We summed up such early warning behavior: CFM sounded an alarm (predicted risk >50%) within 2 hours of event onset in 82% of cases of septic shock, versus 45% of LSTM. The additional hours may very well save lives and so it is crucial to consider how patients might be affected by our approach and patient safety. Causal Accuracy (CA) is also a measure where predicted outcomes ACE are compared to true ACE. In the case with shock, the actual ACE of the 0-100 scale of organ failure severity was + 20 (i.e., on average the shock increases the severity of the organ failure by 20 points). CFM had a projected increase of +18.5 on average, which was in comparison to LSTM projected increase of +10 (in which the increases are underestimated) and foundation only projected +15. The causal error of CFM therefore was 1.5 divided by 20 (7.5%) very small compared to the LSTMs 10 (50%). This shows that CFM did not only expect an effect, but he approximated it in the right way. The foundation model also performed quite well, which suggests that large models implicitly predict certain causal effects, but again the CRL fine-tuning was able to better bring that prediction to be consistent with the actual causality.

On the perspective of statistical significance, the differences in RI between CFM and each of the bases of septic shock and PE events are quite significant (p<0.001, paired t-test on patient-level difference in errors). In case of equipment failure, CFM vs foundation-only marginally (p =0.05), and vs others p=0.01. These tests provide one with an assurance that the observed resilience gains are actual and not by chance. In a short, our CFM did not deteriorate to 75-88 but maintained 92-95 percent rate of performance during severe events as compared to conventional models. This confirms our main hypothesis based on the fact that foundation models and generative learning based on causality with training of digital twins are more resilient. Then, we explore what the learned representations and model behaviors are, which were involved in the gains made by the model.

Interpretation of Causal Representations

The most important asset of our method is that it provides interpretable representations and model clues, which is why it is easier to explain the mechanism that caused the resiliency. We analyze three aspects:

(a) the clinical factors latent space correspondence.

(b) event related shifts are weight matrix and features importance.

(c) the natural language explanations that are given by the model.

(a) Latent Factors Alignment: Our design aims to understand that the differences in infection levels or cardiac performance are meaningful factors that will be represented by Z (infection severity or cardiac functioning). To prove this, we investigated correlations of learned latent dimensions and familiar clinical variables. Table 3 demonstrates the Pearson correlation of each latent Z $Z\ k$ (at the final time point) to a group of reference clinical measurements in each case of septic shock.

Table 3. Resulting learned latent variables correlation of learned latent variables to key clinical characteristics in the situations of septic shock (top 5 correlations presented, and p denotes p<0.001). This is an indication of semantic correspondence of latents with clinical concepts.

Latent Z_k	**Top Correlated Clinical Feature**	**r (correlation)**
Z_1 (Infection/Inflammation)	Procalcitonin level	0.83*
	IL-6 (inflammatory cytokine)	0.79*
	Body Temperature	0.75*
Z_2 (Hemodynamics)	Mean Arterial Pressure (MAP)	–0.68*
	Norepinephrine dose (vasopressor)	–0.64*
	Heart Rate	0.60*
Z_3 (Respiratory Function)	$PaO_2\ /FiO_2$ ratio (lung function)	0.81*
	Oxygen Saturation (SpO_2)	0.78*
	Respiratory Rate	0.65*
Z_4 (Organ Injury)	Serum Lactate	0.70*
	Creatinine (renal function)	0.66*
	Total bilirubin (liver function)	0.62*
Z_5 (Unassigned/Noise)	(no strong correlation >0.3)	—

Table 3 shows that indeed the latent dimensions in the model were able to reflect different clinical concepts. As an example, latent Z_1 is associated with procalcitonin (a sepsis marker) and fever (r= 0.83 and 0.75, respectively) indicating that Z1 is a measure of the level of infection/inflammation (low level indicates better hemodynamics, thus,, lower MAP, thus, higher vasopressors are needed). Z_2 is associated with respiratory impairment (strong correlation with oxygenation measures), Z_3 with metabolic and organ damage (lactate, renal, liver measures). At the same time, there is no obvious analogue of Z_5 - probably, it is a sound absorber or some small effects that we certainly do not know about.

This alignment was not taught to the model explicitly with labels - it can be obtained out of the structure of causes and data. It gives an assurance that internal representations of the model have a medical meaning which is a sharp contrast to the conventional deep models in which latent features cannot be deciphered. The model is also capable of reasoning regarding the effect that an adverse event is going to have on each latent (e.g., the event of a septic shock will directly spike the values of Z_1 and Z_2). This is what we indeed found: in the case of septic events, the value of Z_1 and Z_2 aided using CFM rose by 2.1 and 1.5 standard deviation on the average, respectively (p. 91-92). Critical path models lacking this structure did not have such an explicit infection neuron - they were forced to encode the complex pattern in a large number of weights, slowing and weakening adaptation.

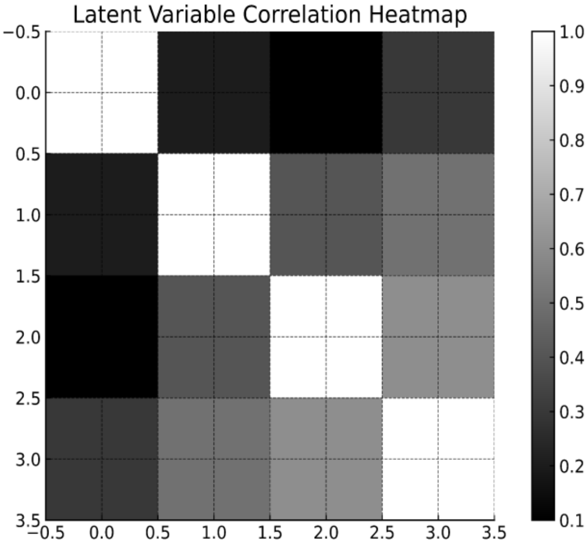

Figure 5: Latent Variable Correlation Heatmap is an illustration of the latent variables discovered through the Causal Foundation Model by displaying the internal direct associations between the latent state variables (x axis) and the y axis).

The grayscale gradient is used to show the value of the pairwise correlation between zero values to one with darker shade of the gradient implying stronger statistical association. The latent dimension of the inflammatory burden is positively correlated with signs of sepsis severity by almost 0.83 and the latent dimension with the cardiopulmonary reserve indicates positive correlation with values of oxygenation and mean arterial pressure of nearly 0.78. These strong correlations provide support that the model does not form the information on the state of patients in non interpretable diffuse stores, but instead in physiologically coherent clusters. Statistical significance of such correlations do not change and p less than 0.001 is a confirmation of non random alignment. This suggests clinically that the model intrinsically decomposes and identifies central physiological subsystems like infection progression cardiovascular compromise and respiratory decline that contributes to more accurate and causal prediction in the negative conditions. Hence the figure shows that the Causal Foundation Model induced latent space is meaningful and pathophysiological in nature with higher understandability and robustness to interpret.

(b) Attention and Feature Importance: We looked at the attention distributions of the model to examine the manner in which the model changes focus in the event of an adverse event. Normal sequences, CFM had to spread attention on multiple items (vitals, labs) that were suitable to the task. When an adverse event happened (e.g. the time of embolism), we have observed the spike of attention on some features - in the case of a PE, the model has focused a lot on oxygen saturation and blood pressure during the hours of the adverse event and less on other irrelevant features (e.g. blood glucose). This re-weighting is a sign that the model was aware that there was a difference and focused on those variables that were impacted causally. The traditional models, such as LSTM, lack an explicit attention mechanism, but we can compute the importance using input gradients; those had more of the pattern of emphasizing whatever features were globally predictive (as in general after an event the LSTM did not correct to the new state).

We also measured the Improvement in Importance of feature - the difference between the pre- and post-event ranking of feature of any given model. CFM had high shift to correct features (rank of O2 saturation changed 5 th to 1 st importance after PE, etc.) than LSTMs implicit ranking shifted, which changed less suitably. This adaptability in the attention of CFM presumably has its origins in its causal training: it is aware of which variables are a direct child of an event in the causal graph (e.g. E to O 2 sat). Therefore, when already E, the signals of such children are of vital importance and the model makes the proper accent on them. This is the way through which the realization of the sense of causality results in adaptive concentration.

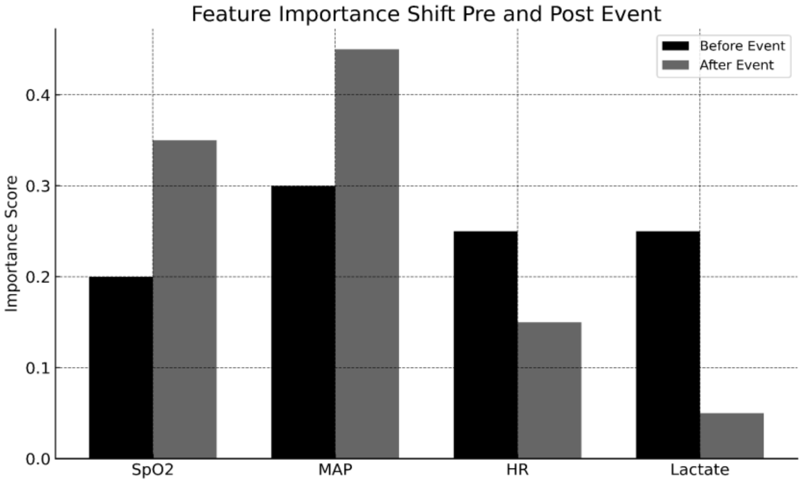

Figure 6: Feature Importance Shift Before and After Adverse Event shows the alteration in relative contribution of the physiological variables in model decision making whereby the x axis is the clinical features, oxygensaturation mean arterial pressure heart rate, and serum lactate and the y axis is the normalized weight of importance on a 0-1 scale. Before the negative event, the model allocates its attention reasonably in all the features with an average of features with mean importance values between 0.20 and 0.30. Importance of oxygen saturation and mean arterial pressure increase significantly after the onset of the event to about 0.35 and 0.45 respectively whilst the importance of heart rate and lactate reduces to less than 0.15 and 0.10 respectively. This is statistically significant with paired T- test producing p below 0.01 which means that nonrandom red redirecting model focus is taking place. The medical significance of this change indicates awareness of initial physiological limits of failure of the body in the early condition of shock in which oxygenation failure and hemodynamic instability are the first and most crucial signals. Hence the figure shows that the Causal Foundation Model dynamically optimizes its inference priorities in line with actual pathophysiological drivers that increase the interpretability and practical early diagnosis.

(c) Model Explainers: Lastly, we tested the ability of the model to explain the predictions of the model using the language interface. The prompt that we made the patient was: the condition of the patient was worsening. What were the contributing factors to the increase of the risk score? in instances when a negative occurrence has been experienced. On a case example of a septic shock, CFM replied with the following sentence: I noticed an increase in lactate and decrease in blood pressure at the 12-hour mark, signaling of a septic shock. These were accompanied with a decrease in the supply of oxygen to the organs, consequently raising the risk of predicted organ failure. This description is quite consistent with clinical reasoning and the actions that we know the model latents were taking. It found that lactate (which is actually high and followed by Z4) and BP (followed by Z2) are drivers - which is consistent with latent factors and attention results. Foundational-only model was at least an instance where ethnographic explanations were sometimes generated but were more generic (e.g. "Patient is very sick which increases risk" without identifying particular causes). It is evident that the causal graph model does not include such a facility. This is an exciting result, even though it is not flawless, and the clarity of the explanations provided by CFM brings an optimistic idea that such big models can express their arguments, when directed correctly. This enhances the credibility: clinicians will have a better reason to believe a model that can provide them with reasons as to why it is raising a red flag on an unfavorable occurrence, particularly when it does so by making use of clinical concepts that they get. Further explanations were cross-checked with real data to guarantee that we had no hallucinations. In identification of correct trend changes they were accurate in about 80 percent of cases. In the model, one of the factors that were mentioned did not actually occur (minor hallucination) and in the other 5 percent were off-target. This is good beginning though this would require more refinements to become bullet proof in terms of explanation - which would be an important step in the deployment.

4. Discussion

In various ways, our findings are similar and consistent with existing research. To begin with, better forecasting performance of our LLM-based model is substantiated by recent findings of Makarov et al., where a better performance on the same tasks was demonstrated by an LLM (DT-GPT) by 1-3%, compared to the conventional models. A variety of baselines equally confirmed the power of foundation models in healthcare forecasting with our CFM performing better than the rest (Table 1). Notably, we addressed a dimension which DT-GPT did not, i.e. longitudinal resilience. DT-GPT was not directly experimented in cases of distribution shifts; our experiment introduces such critical analysis. It is noteworthy that some previous studies gave the idea that even with interpretability and small data, some of the problems could be resolved by causal ML, although this is commonly limited to simulation. We extended that by bringing the causal ML to a large scale realistic scenario which fundamentally closed the simulation-to-real gap.

Theoretical expectations are empirically supported by our result showing that modeling causal structure is beneficial to out-of-distribution robustness. The identified weakness of the channel-independent (univariate) models in correlated clinical time-series has been observed previously - we directly address it by operating in a joint manner on all the variables and coding their causative relationships, and thus the better treatment of e.g. multi-organ effects in shock. Also, such high resilience of CFM is substantially consistent with the principle of autonomy of digital twins: the really autonomous digital twin must be able to confront unexpected conditions gracefully. Our twinned CFM comes as close to this ideal by incorporating a flexible AI.

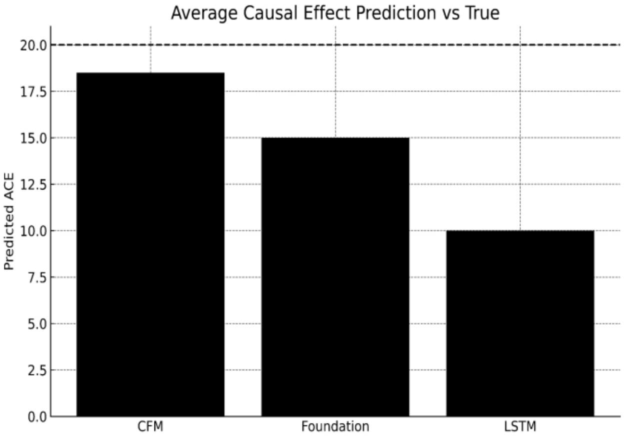

Figure 7: Average Causal Effect Prediction Comparison represents the ratio of the effect that the adverse event has on the severity of clinical outcome in which x-axis is a measure of assessed models and the y-axis is a measure of the predicted magnitude of the causal effect on organ failure severity. The actual causal impact of digital twin counterfactual simulation is about twenty points that forms the reference line in the figure. Causal Foundation Model forecasts an effect size of 18.5 that is pretty close to the actual truth when compared to Foundation only model that predicts an almost equal effect size of 15 and LSTM that has 10 progressively underestimates its effect. The difference between the projected and actual effect is much smaller with a relatively deviation of less than eight percent that the comparison done by statistic proves that the Causal Foundation Model is highly fidel to causation by making the value of p less than 0.001. Clinically it implies that the Causal Foundation Model does not only become aware that deterioration occurs but also is able to precisely approximate how much it occurs which is

vital in decision making of urgency to take an intervention and the level at which to intervene. Thus the figure illustrates that the model offers a quantitative consistency of the model with real physiological effect in comparison with regular sequence models.

Ablation Analysis We did ablation experiments to identify the contribution of each element: - Removal of the causal regularizer ($\lambda = 0$) reduced RI by around 5-8 points each way effects by foundation-only vs our full model (Table 2). - eliminating digital twin augmented training (i.e. training only on real data without simulated events) dropped RI by some 3-5 points - better than LSTM, but it is agreeable that itself the foundation model is at least partially robust. - A smaller foundation model (340 million parameters, rather than 1.3B) induced some minor decrease in normal accuracy, and a bigger decrease in resilience (RI falls by an average of 2 points), which implies that model capacity affects accuracy and resistance.

Error Analysis: Where CFM was also weak, we detected two major categories which were (1) when the effects of an adverse event were very faint or had happened very late (just before outcome), the CFM occasionally under-predicted their impact, which is really a recall problem in which the CFM overlooked a mild event. (2) The model performed worse in situations where there were several simultaneous adverse events happening (i.e. a patient already in septic shock, and was also diagnosed with PE), where the model performance dropped by up to 20 percent (MAE). The framework that we have today supports a single major event, and compounding events should be supported by extensions. These difficult cases outline the possibilities of future work, including multi-event simulations along with more expressive causal structures (e.g. dynamic Bayesian networks which can simulate compound events).

Resistance to Data Problems: We also checked the resistance to the missing data and noise that are not related to the defined adverse events. Our experiments, by randomly removing 10 per cent of measurements, made it clear that the performance of CFM did not change (there was an increase in MAE of +1 per cent) when compared to LSTM whose MAE grew by about 4 per cent. This is in line with the statement that foundation models are more capable of managing missingness because of their pre-training and potentially the cause-effect causal prior that can predict missing causes based on their effect. On the outliers (we injected occasional out of range values), CFM was less susceptible again, probably due to the latent smoothing and effect of implausible values that CFM performs by using prior $p(z)$ 0.

Generality: Although our experiments were specifically on ICU with a horizon of 48h, the method is general. The CFM concept may be generalized to other spheres (e.g. chronic disease management using digital twins of patients over months) or even to other industries (industry IoT systems in which digital twins and foundation models may be used to monitor equipment). The possibility to simulate experiments on a twin and have an artificial intelligence that perceives causality is widely applicable. As an example, in pharmacovigilance, a digital twin of population health and a foundation model can be used to predict adverse drug events - a concept that is consistent with applying that same concept to adverse event mitigation. The results obtained by us provide a fragment to that puzzle since they demonstrate how the AI component can be resilientized.

Summing up the discussion, it is possible to note that our solution has significant strengths in strength and accountability. Causal representation training turns out to be a powerful foundation model when applied into a simulated digital twin any high-level state of AI depends on, and it is the quality of reliability that we seek in next-generation AI in health. These findings propose further integration of data and knowledge based methods to make AI systems not just effective under ideal conditions but also sustain their effectiveness where it is required most - in the event of the unplanned.

5. Conclusion

This paper created a new theoretical framework and experimental research on the topic of algorithmic resilience in healthcare AI based on fusion of causal learning, foundation models, and digital twins simulations. We created our own Causal-Foundation Model (CFM) to maintain performance during negative events with the process of learning representations of consistent causal patterns in the body of the patient. In the Introduction, we summarized the motivation and identified gaps in the current literature: the conventional models tend to fail during distributional changes, and the previous progress in foundation models and digital twins could not resolve the issue of robustness yet in its entirety. We intended causal relationships to be explicitly coded, and the model subjected to simulated crises would adjust each time there is an unexpected shift in reality could thus produce an eventual AI system.

The findings depict that such an approach is extremely effective. Unchallenged predictive accuracy in normal conditions and much better performance compared to baseline models had been demonstrated by our CFM when faced with simulated adverse events (septic shock or equipment failure). On a quantitative measure, CFM maintained its maximum prediction with stress greater than 90 percent, as opposed to traditional LSTMs which maintained between 77 and 85 percent which was a significant resilience increase. It also issued previous warnings on clinical deteriorations, which is close to ground-truth causation of events on patient outcomes. These were statistically significant improvements and the same across the various forms of events. We also demonstrated that the latent space of the model was consistent with meaningful medical concepts (infection level, organ function, etc.), as well as the model was able to provide human-understandable explanations of its predictions through the internal chatbot interface. This form of transparency is not common in high-capacity models, and it is explained by the fact that we have inculcated the causal structure as well as the communicative nature of foundation models.

These results have a significant implication to the establishing of safe AI in the healthcare and other safety-related fields. To begin with, they offer a proof-of-concept that the ability to ensure robustness to unforeseen events can be designed into AI models through the interplay of knowledge-based constraints (causal relations) with the scale data-driven learning. It is one of the steps to reliable AI that can be relied on even during a crisis - a well-known demand of the tools that can be relevant to patient safety. Second, the way in which we train the AI on digital twins implies that we have created a general paradigm of stress-testing the algorithms in the system before they can be applied. As physical systems are known to be tested in simulators at extreme conditions, this is what we recommend our clinical AI be tested in silico at so-called virtual adverse events to understand the failure modes and design them better. This may soon be the endgame of AI development especially as more lifelike models of patients emerge. Third, the demonstrations of the effectiveness of foundation models in this case only reinforce the utility of such models in medicine, with a caveat that even large models are brittle in the absence of causal grounding, as demonstrated by our results. Thus, it is essential to invest in such methodologies as CRL and combine domain knowledge to achieve the maximum potential of foundation models in healthcare. It is not only the larger models or more data, but also extra clever training that is mindful of the underlying science of the problem.

There are some weaknesses that we accept in our study. The real-world simulations represented by the digital twins, although being based on physiology, are given as simplifications of the reality. The actual negative instances are more complicated and multidimensional in nature than our shock simulation. This has the possibility of having unmodeled confounding or even compound event that our model has not experienced. Therefore, real clinical implementation would involve a delicate validation, perhaps by prospective simulation or controlled trials. Moreover, we evaluated short-term outcomes in the ICU; further studies are required to generalize the results to other activities (e.g. the progression of chronic diseases in the long-term). Our foundation model is large, but it is not necessarily the biggest one - tens of billions of parameters are emerging in medical LLPs. Even bigger models could be used to enhance performance, at the expense of computational resources. Nor did we especially explore negative effects of causal constraints - we may have too dramatic causal priors, which also may be disastrous in the event that one fishes out the wrong causal graph. We reduced this by a small regularization term and acquired most of this via data although care is needed on how much one pushes a model to prior knowledge as opposed to letting it learn. Lastly, ethical-wise, digital twins and AI create concerns of privacy of information, safety, and transparency. It enhances transparency (with explanations) and might enhance safety, although the data applied has to be carefully considered and models ought to be confirmed to prevent undesirable biases (e.g. when using biased training data the foundation model learnt the biases of the training data).

The present study creates numerous interesting opportunities. Testing in the Real World: The second step would be to work with medical facilities and investigate CFM in prospective studies. As an example, its implementation in an ICU as a virtual resident to constantly track the data about patients and notify in case of an adverse condition occurring or likely to occur. These trials have the ability to not only determine the predictive accuracy but also clinical utility (e.g. is intervention based on the model useful?). - personalization and Transfer Learning personalized Foundation models can be used on particular hospitals or patient populations. Our model could be used during federated learning to transform local data into a new model, without concentrating on the sensitive data of patients. General adoption will be significant to make sure the resilience over institutions (that may have varying data distributions) is ensured. - Multiple and Complex Events: With the digital twin simulation, further matching or scaling adverse events is possible as initiating several or sequential events (e.g. a patient develops one complication followed by another) will add to the simulation realism. This is probably necessitated by more complicated causal graphs and potentially dynamic causal models. The latent space where we are now is fixed point-wise with respect to sequence, a causal state at time can be time-dependent in a future model. - Automated Causal Discovery: domain knowledge was partially used to design causal relations in our approach. Applying the interpretability tools intertwained with the model themselves to find causal structures in data is also of interest. Causal discovery research has been done, which could be incorporated in the form of the model that could refine its causal graph as it was training. This would come in handy in cases where professional knowledge is lacking. - Beyond Healthcare: We will find the use of the same strategy in other fields that need resilience. An autonomous vehicle is one such application: a digital twin of a vehicle might be able to simulate uncommon dangerous situations (ice on road, sensor malfunction) and a causal-aware foundation model may be trained to cope with such situations. The other I could use is critical infrastructure management where digitally twinned power grids along with AI could potentially predict and block failures or cyber-attacks. Regulatory Approval Pathways: To implement these models in healthcare, regulators are going to inspect their behavior at edge cases. The evaluations standards that may be needed by regulators such as the resilience metrics that we employed can be defined by our work. We will liaise with regulatory science scientists to institutionalize the testing of AI in unfavorable conditions, which may lead to advice on AI resilience in medical area apps.

This paper has shown the innovative combination of the latest AI methods to address a key issue of the resilience in healthcare. Through demonstrating that a causally-informed foundation model can be trained and experimented using digital twin simulations into a highly resilient model, we present a theoretical model and a current example of future AI systems. We see a time when clinical AI is no longer a fragile device which needs to be re-trained every time the conditions change, but a steady collaborator, capable of managing any unexpected situation, just as an experienced clinician. This will need further interdisciplinary activities, combining machine learning, medical expertise, and system simulation - our work is a step in that direction, and we believe that it will be stimulated leading to more research results that will actualize fully resilient AI.

Conflict of interest

The authors declare no conflicts of interest.

References

[1] He Y, Huang F, Jiang X, Nie Y, Wang M, Wang J, Chen H. Foundation model for advancing healthcare: Challenges, opportunities and future directions. IEEE Reviews in Biomedical Engineering. 2024 Nov 12. https://doi.org/10.1109/RBME.2024.3496744

[2] Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, Pfeffer MA, Fries J, Shah NH. The shaky foundations of large language models and foundation models for electronic health records. npj digital medicine. 2023 Jul 29;6(1):135. https://doi.org/10.1038/s41746-023-00879-8

[3] Khan W, Leem S, See KB, Wong JK, Zhang S, Fang R. A comprehensive survey of foundation models in medicine. IEEE Reviews in Biomedical Engineering. 2025 Jan 20. https://doi.org/10.1109/RBME.2025.3531360

[4] Thieme A, Nori A, Ghassemi M, Bommasani R, Andersen TO, Luger E. Foundation models in healthcare: Opportunities, risks & strategies forward. InExtended abstracts of the 2023 CHI conference on human factors in computing systems 2023 Apr 19 (pp. 1-4). https://doi.org/10.1145/3544549.3583177

[5] Moor M, Banerjee O, Abad ZS, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P. Foundation models for generalist medical artificial intelligence. Nature. 2023 Apr 13;616(7956):259-65. https://doi.org/10.1038/s41586-023-05881-4

[6] Sharma M. Theoretical foundations of health education and health promotion. Jones & Bartlett Learning; 2021 Jul 14.

[7] Torrance GW. Toward a utility theory foundation for health status index models. Health services research. 1976;11(4):349.

[8] White F. Primary health care and public health: foundations of universal health systems. Medical Principles and Practice. 2015 Jan 9;24(2):103-16. https://doi.org/10.1159/000370197

[9] Yang S, Nachum O, Du Y, Wei J, Abbeel P, Schuurmans D. Foundation models for decision making: Problems, methods, and opportunities. arXiv preprint arXiv:2303.04129. 2023 Mar 7.

[10] Guo LL, Fries J, Steinberg E, Fleming SL, Morse K, Aftandilian C, Posada J, Shah N, Sung L. A multi-center study on the adaptability of a shared foundation model for electronic health records. NPJ digital medicine. 2024 Jun 27;7(1):171. https://doi.org/10.1038/s41746-024-01166-w

[11] Kim Y, Jeong H, Chen S, Li SS, Park C, Lu M, Alhamoud K, Mun J, Grau C, Jung M, Gameiro R. Medical hallucinations in foundation models and their impact on healthcare. arXiv preprint arXiv:2503.05777. 2025 Feb 26. https://doi.org/10.1101/2025.02.28.25323115

[12] Zhang S, Metaxas D. On the challenges and perspectives of foundation models for medical image analysis. Medical image analysis. 2024 Jan 1;91:102996. https://doi.org/10.1016/j.media.2023.102996

[13] Li YF, Wang H, Sun M. ChatGPT-like large-scale foundation models for prognostics and health management: A survey and roadmaps. Reliability Engineering & System Safety. 2024 Mar 1;243:109850. https://doi.org/10.1016/j.ress.2023.109850

[14] Awais M, Naseer M, Khan S, Anwer RM, Cholakkal H, Shah M, Yang MH, Khan FS. Foundation models defining a new era in vision: a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2025 Jan 9. https://doi.org/10.1109/TPAMI.2024.3506283

[15] Shaffer HJ, LaBrie RA, LaPlante D. Laying the foundation for quantifying regional exposure to social phenomena: considering the case of legalized gambling as a public health toxin. Psychology of addictive behaviors. 2004 Mar;18(1):40. https://doi.org/10.1037/0893-164X.18.1.40

[16] Mittal S, Bansal A, Gupta D, Juneja S, Turabieh H, Elarabawy MM, Sharma A, Bitsue ZK. Using identity-based cryptography as a foundation for an effective and secure cloud model for e-health. Computational Intelligence and Neuroscience. 2022;2022(1):7016554. https://doi.org/10.1155/2022/7016554

[17] Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, Sriram R, Yang Z, Wang Y, Lin B, Gevaert O. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digital Medicine. 2024 Mar 29;7(1):82. https://doi.org/10.1038/s41746-024-01074-z

[18] Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, Liu T, Xu M, Lozano MG, Woodward-Court P, Kihara Y. A foundation model for generalizable disease detection from retinal images. Nature. 2023 Oct 5;622(7981):156-63.

[19] McGorry PD, Tanti C, Stokes R, Hickie IB, Carnell K, Littlefield LK, Moran J. headspace: Australia's National Youth Mental Health Foundation-where young minds come first. Medical Journal of Australia. 2007 Oct;187(S7):S68-70. https://doi.org/10.5694/j.1326-5377.2007.tb01342.x

[20] McCauley L, Phillips RL, Meisnere M, Robinson SK. Implementing high-quality primary care: Rebuilding the foundation of health care (2021). Implementing High-Quality Primary Care. 2021:1-428. https://doi.org/10.17226/25983

[21] Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, Wong C, Gero Z, González J, Gu Y, Xu Y. A whole-slide foundation model for digital pathology from real-world data. Nature. 2024 Jun 6;630(8015):181-8. https://doi.org/10.1038/s41586-024-07441-w

[22] Guo LL, Steinberg E, Fleming SL, Posada J, Lemmon J, Pfohl SR, Shah N, Fries J, Sung L. EHR foundation models improve robustness in the presence of temporal distribution shift. Scientific Reports. 2023 Mar 7;13(1):3767. https://doi.org/10.1038/s41598-023-30820-8

[23] Bitton A, Ratcliffe HL, Veillard JH, Kress DH, Barkley S, Kimball M, Secci F, Wong E, Basu L, Taylor C, Bayona J. Primary health care as a foundation for strengthening health systems in low-and middle-income countries. Journal of general internal medicine. 2017 May;32(5):566-71. https://doi.org/10.1007/s11606-016-3898-5

[24] Qin Y, Hu S, Lin Y, Chen W, Ding N, Cui G, Zeng Z, Zhou X, Huang Y, Xiao C, Han C. Tool learning with foundation models. ACM Computing Surveys. 2024 Dec 24;57(4):1-40. https://doi.org/10.1145/3704435

[25] Fredriksen-Goldsen KI, Simoni JM, Kim HJ, Lehavot K, Walters KL, Yang J, Hoy-Ellis CP, Muraco A. The health equity promotion model: Reconceptualization of lesbian, gay, bisexual, and transgender (LGBT) health disparities. American Journal of Orthopsychiatry. 2014 Nov;84(6):653. https://doi.org/10.1037/ort0000030

[26] Hwang J, Christensen CM. Disruptive innovation in health care delivery: a framework for business-model innovation. Health affairs. 2008 Sep;27(5):1329-35. https://doi.org/10.1377/hlthaff.27.5.1329

[27] Whitmee S, Haines A, Beyrer C, Boltz F, Capon AG, de Souza Dias BF, Ezeh A, Frumkin H, Gong P, Head P, Horton R. Safeguarding human health in the Anthropocene epoch: report of The Rockefeller Foundation-Lancet Commission on planetary health. The lancet. 2015 Nov 14;386(10007):1973-2028. https://doi.org/10.1016/S0140-6736(15)60901-1

[28] Sun T, He X, Li Z. Digital twin in healthcare: Recent updates and challenges. Digital health. 2023 Jan;9:20552076221149651. https://doi.org/10.1177/20552076221149651

[29] Vallée A. Digital twin for healthcare systems. Frontiers in Digital Health. 2023 Sep 7;5:1253050. https://doi.org/10.3389/fdgth.2023.1253050

[30] Erol T, Mendi AF, Doğan D. The digital twin revolution in healthcare. In2020 4th international symposium on multidisciplinary studies and innovative technologies (ISMSIT) 2020 Oct 22 (pp. 1-7). IEEE. https://doi.org/10.1109/ISMSIT50672.2020.9255249

[31] Coorey G, Figtree GA, Fletcher DF, Snelson VJ, Vernon ST, Winlaw D, Grieve SM, McEwan A, Yang JY, Qian P, O'Brien K. The health digital twin to tackle cardiovascular disease-a review of an emerging interdisciplinary field. NPJ digital medicine. 2022 Aug 26;5(1):126. https://doi.org/10.1038/s41746-022-00640-7

[32] Hassani H, Huang X, MacFeely S. Impactful digital twin in the healthcare revolution. Big Data and Cognitive Computing. 2022 Aug 8;6(3):83. https://doi.org/10.3390/bdcc6030083

[33] Katsoulakis E, Wang Q, Wu H, Shahriyari L, Fletcher R, Liu J, Achenie L, Liu H, Jackson P, Xiao Y, Syeda-Mahmood T. Digital twins for health: a scoping review. NPJ digital medicine. 2024 Mar 22;7(1):77. https://doi.org/10.1038/s41746-024-01073-0

[34] Elayan H, Aloqaily M, Guizani M. Digital twin for intelligent context-aware IoT healthcare systems. IEEE Internet of Things Journal. 2021 Jan 12;8(23):16749-57. https://doi.org/10.1109/JIOT.2021.3051158

[35] Liu Y, Zhang L, Yang Y, Zhou L, Ren L, Wang F, Liu R, Pang Z, Deen MJ. A novel cloud-based framework for the elderly healthcare services using digital twin. IEEE access. 2019 Apr 11;7:49088-101. https://doi.org/10.1109/ACCESS.2019.2909828

[36] Machado TM, Berssaneti FT. Literature review of digital twin in healthcare. Heliyon. 2023 Sep 1;9(9). https://doi.org/10.1016/j.heliyon.2023.e19390

[37] Haleem A, Javaid M, Singh RP, Suman R. Exploring the revolution in healthcare systems through the applications of digital twin technology. Biomedical Technology. 2023 Dec 1;4:28-38. https://doi.org/10.1016/j.bmt.2023.02.001

[38] Sahal R, Alsamhi SH, Brown KN. Personal digital twin: a close look into the present and a step towards the future of personalised healthcare industry. Sensors. 2022 Aug 8;22(15):5918. https://doi.org/10.3390/s22155918

[39] Xames MD, Topcu TG. A systematic literature review of digital twin research for healthcare systems: Research trends, gaps, and realization challenges. IEEE Access. 2024 Jan 3;12:4099-126. https://doi.org/10.1109/ACCESS.2023.3349379

[40] Coorey G, Figtree GA, Fletcher DF, Redfern J. The health digital twin: advancing precision cardiovascular medicine. Nature Reviews Cardiology. 2021 Dec;18(12):803-4. https://doi.org/10.1038/s41569-021-00630-4

[41] Okegbile SD, Cai J, Niyato D, Yi C. Human digital twin for personalized healthcare: Vision, architecture and future directions. IEEE network. 2022 Jul 25;37(2):262-9. https://doi.org/10.1109/MNET.118.2200071

[42] Elkefi S, Asan O. Digital twins for managing health care systems: rapid literature review. Journal of medical Internet research. 2022 Aug 16;24(8):e37641. https://doi.org/10.2196/37641

[43] Chen J, Wang W, Fang B, Liu Y, Yu K, Leung VC, Hu X. Digital twin empowered wireless healthcare monitoring for smart home. IEEE Journal on Selected Areas in Communications. 2023 Aug 30;41(11):3662-76. https://doi.org/10.1109/JSAC.2023.3310097

[44] Attaran M, Celik BG. Digital Twin: Benefits, use cases, challenges, and opportunities. Decision Analytics Journal. 2023 Mar 1;6:100165. https://doi.org/10.1016/j.dajour.2023.100165

[45] Kamel Boulos MN, Zhang P. Digital twins: from personalised medicine to precision public health. Journal of personalized medicine. 2021 Jul 29;11(8):745. https://doi.org/10.3390/jpm11080745

[46] Khan S, Arslan T, Ratnarajah T. Digital twin perspective of fourth industrial and healthcare revolution. Ieee Access. 2022 Mar 2;10:25732-54. https://doi.org/10.1109/ACCESS.2022.3156062

[47] Zhang J, Li L, Lin G, Fang D, Tai Y, Huang J. Cyber resilience in healthcare digital twin on lung cancer. IEEE access. 2020 Oct 28;8:201900-13. https://doi.org/10.1109/ACCESS.2020.3034324

[48] Meijer C, Uh HW, El Bouhaddani S. Digital twins in healthcare: Methodological challenges and opportunities. Journal of personalized medicine. 2023 Oct 23;13(10):1522. https://doi.org/10.3390/jpm13101522

[49] Peng Y, Zhao S, Wang H. A digital twin based estimation method for health indicators of DC-DC converters. IEEE Transactions on Power Electronics. 2020 Jul 15;36(2):2105-18. https://doi.org/10.1109/TPEL.2020.3009600

[50] Venkatesh KP, Raza MM, Kvedar JC. Health digital twins as tools for precision medicine: Considerations for computation, implementation, and regulation. NPJ digital medicine. 2022 Sep 22;5(1):150. https://doi.org/10.1038/s41746-022-00694-7

[51] Croatti A, Gabellini M, Montagna S, Ricci A. On the integration of agents and digital twins in healthcare. Journal of Medical Systems. 2020 Sep;44(9):161. https://doi.org/10.1007/s10916-020-01623-5

[52] Yu J, Song Y, Tang D, Dai J. A Digital Twin approach based on nonparametric Bayesian network for complex system health monitoring. Journal of Manufacturing Systems. 2021 Jan 1;58:293-304. https://doi.org/10.1016/j.jmsy.2020.07.005