Reliable Policy Learning: Assessing Treatment Variation in Healthcare

Unnseo (Grace) Park

CMU-CS-25-110

May 2025

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Adam Perer, Chair
Zachory Erickson

Submitted in partial fulfillment of the requirements for the Masters degree.

Keywords: Healthcare Machine Learning, Medical Decision-Making, Dataset Evaluation, Dynamics Models, Transformer Models, Simulation Environment, Visualization Tool

To my family and friends — for your endless encouragement, love, and belief in me.

Abstract

Recent advances in machine learning for personalized medicine have created a need to determine when observational healthcare data can reliably inform treatment policies. This thesis examines methods for evaluating whether treatment variation in medical datasets is sufficient for developing dependable clinical policies. Through three complementary approaches, we investigate methods to detect and measure meaningful action diversity in healthcare data:

First, we analyze the MIMIC sepsis dataset using transformer-based dynamics models. Our findings reveal that including action information provides minimal improvement in outcome predictions across the entire dataset. This suggests limited meaningful treatment diversity when analyzed in aggregate. Second, in our controlled simulation experiments with a one-dimensional GridWorld environment, we demonstrate that comparing prediction performance between models with and without action inputs effectively identifies regions where treatments meaningfully impact outcomes. Finally, we present a novel interactive visualization tool that employs t-SNE dimensionality reduction and intuitive diversity metrics to help researchers explore action diversity across patient state spaces. This tool helps identify subgroups where treatment policies can be reliably learned.

Our findings demonstrate that dynamics model comparisons can effectively identify regions where treatment policies can be reliably learned, enabling more targeted and trustworthy deployment of machine learning in healthcare. This framework provides researchers with practical tools to evaluate data sufficiency before deploying treatment recommendation systems, potentially improving both the reliability of AI assistance in clinical decision-making and, ultimately, patient outcomes.

Acknowledgments

I would like to express my deepest gratitude to my mentor, Venkat Sivaraman, whose guidance, patience, and encouragement supported me through every stage of this research. From the initial ideas to the final experiments, Venkat has been an invaluable part of this journey.

I am also incredibly grateful to my advisor, Professor Adam Perer, for his mentorship and for giving me the opportunity to explore research that truly excited me. His feedback and perspective consistently challenged me to think deeper and push further.

Thank you to Professor Zackory Erickson for serving on my committee and providing thoughtful insights into my work.

Lastly, I am thankful for the friends, colleagues and my family, whose support made this all possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Current healthcare practice relies largely on standardized clinical guidelines that may not fully account for individual patient differences and contexts. This approach provides a foundation for treatment decisions, but often falls short of optimal care for individual patients. Recent advances in machine learning and the increasing availability of electronic health records have created new opportunities for data-driven personalized medicine. This wealth of observational data has sparked numerous machine learning approaches for learning treatment policies, opening the possibility of truly personalized treatment recommendations that could significantly improve patient outcomes. However, we must be certain of the reliability of any learned treatment policies as incorrect treatment recommendations could lead to adverse outcomes or patient harm. Although we have access to more medical data than ever before, we lack robust methods to determine when this data is sufficient for learning reliable treatment policies. This gap between data availability and our ability to verify policy reliability remains a critical challenge in the advancement of personalized medicine.

1.2 Problem Statement and Goal

Learning reliable treatment policies from observational data presents several fundamental challenges. In clinical practice, clinicians understandably follow standard protocols and guidelines, leading to limited exploration of alternative treatments. This creates a key methodological challenge: we can only observe the outcomes of treatments that were actually administered, not the potential outcomes of alternative choices. Moreover, even when there are variations in treatment choices, we currently lack established methods to determine whether these variations are sufficient for learning reliable policies. This makes it difficult to identify which regions of the patient state space have enough treatment diversity to support reliable policy learning. Therefore, there is a critical need for methods that can help researchers identify when and where learned treatment policies can be trusted.

1.3 Contribution

This thesis introduces several key contributions to address these challenges. First, we propose action diversity as a metric to evaluate treatment variation in medical datasets, providing a quantitative way to assess whether we have observed enough variation in treatment choices to learn meaningful policies. Second, we develop simulation methods to identify specific subgroups within datasets where action diversity exists, allowing a more targeted analysis of where treatment policies could be reliably learned. Third, we create an interactive visualization tool that allows researchers to explore action diversity across different regions of the patient state space, helping them understand where they have enough data to learn reliable policies. Together, these contributions provide researchers with systematic ways to evaluate the suitability of their datasets for learning treatment policies and identify potential limitations in their analyses.

Chapter 2

Related Works

2.1 Related Work

2.1.1 Reinforcement Learning in Healthcare

Many studies have explored reinforcement learning (RL) to derive treatment policies from observational data, particularly for critical care scenarios such as sepsis. The AI Clinician [14] applied tabular Q-learning to learn policies from retrospective ICU data, sparking significant interest in offline RL for healthcare. Extensions have incorporated deep learning methods [17, 22], physiological constraints [16], and improved sample efficiency [12, 15]. More recently, [20] proposed factored action spaces to better model the combinatorial nature of clinical decisions.

Despite these advances, several critiques have emerged. [13] showed that even high-capacity models struggle to learn meaningful state representations tied to outcomes. [9] demonstrated how reward misspecification can lead RL agents to recommend dangerous treatments. Other work has proposed better policy evaluation methods [7] and interpretability tools [10]. However, the underlying assumption in these works is that sufficient treatment variation exists to support reliable learning.

2.1.2 Causal Inference and Identifiability

Causal inference approaches aim to estimate treatment effects from observational data while addressing confounding and selection bias. A common challenge in this domain is the overlap assumption which assumes that each patient has a nonzero probability of receiving any of the available treatments [8]. When this assumption is violated, which often occurs in clinical datasets with protocol-driven treatments, treatment effect estimates become unreliable.

Recent work has focused on identifying subgroups where treatment effects differ significantly and can be more reliably estimated. For instance, Wang and Rudin [21] and Bargagli-Stoffi et al. [1] propose interpretable rule-based models to discover patient subgroups with heterogeneous treatment effects. These models help address the problem of non-uniform treatment support by isolating regions of the data where meaningful comparisons can be made. Similarly, Bayesian

Additive Regression Trees (BART) [8] provide a flexible, nonparametric approach for estimating individualized treatment effects, while accounting for uncertainty.

Our work is complementary to these efforts: rather than estimating treatment effects directly, we focus on detecting regions of the state space where treatment variation is both present and predictive of patient outcomes. This provides a diagnostic tool for assessing when observational data is suitable for learning policies, especially in the presence of limited overlap.

2.2 Visualization and Interpretability

Interpretable models and visualizations are essential in clinical settings, where understanding why a model makes a specific recommendation is often as important as the prediction itself. Several prior works have focused on improving model transparency to build clinician trust and support decision-making.

Caruana et al. [4] developed generalized additive models with pairwise interactions (GA2Ms) to predict pneumonia risk and hospital readmission, demonstrating that intelligible models can match the accuracy of black-box models while offering direct insight into feature contributions. Che et al. [6] proposed interpretable deep models using recurrent neural networks for ICU outcome prediction, incorporating mechanisms such as attention and feature-level regularization to enhance interpretability without sacrificing performance.

More recently, Caicedo-Torres and Gutierrez [3] introduced ISeeU, a deep convolutional model for ICU mortality prediction that uses coalitional game theory to provide visual explanations for predictions. Their approach offers clinicians a clearer understanding of how the model weights various input features in making life-critical decisions.

These works illustrate the growing emphasis on interpretability and visualization in clinical machine learning. Our thesis builds on this direction by introducing a t-SNE-based interactive tool that helps researchers explore where treatment effect estimation may be trustworthy, particularly by highlighting state regions with varying levels of treatment impact.

Chapter 3

Measuring Action Diversity Through Dynamics Models

3.1 Introduction

Within the broader context of learning reliable treatment policies from observational data, sepsis represents both a compelling use case and a significant challenge. Sepsis treatment recommendation is a particularly promising area for machine learning research, given both the severity of the condition and the lack of well-established treatment guidelines. Improving treatment strategies for patients with sepsis is a challenge of considerable interest in applied machine learning (ML). Sepsis is a leading cause of death in hospitals, and there is currently little clinical consensus around best practices for treatment [5]. Several recent works have applied reinforcement learning (RL) methods in efforts to support clinician decision making in sepsis patients in the intensive care unit (ICU).

While these algorithms have shown promise when evaluated using off-policy policy evaluation (OPE) methods, off-policy evaluation presents a critical challenge in this domain. These methods attempt to assess how a new policy would perform if deployed, using only historical data collected under different policies. Although this is necessary for ethical evaluation of treatment policies without patient experimentation, OPE methods are subject to important limitations []. Beyond methodological concerns with OPE, recent analyses of specific RL models for sepsis treatment have revealed more troubling issues. Several studies have shown that these models often recommend treatments that deviate significantly from clinical practice, sometimes in ways that appear potentially dangerous [9, 19].

These findings raise a fundamental question central to this thesis: Is it possible to learn effective RL treatment policies from publicly available datasets such as MIMIC (Medical Information Mart for Intensive Care)? A key limitation may be insufficient action diversity—the variation in treatment decisions for similar patient states—within these observational datasets. If clinicians consistently follow similar treatment strategies for patients with comparable conditions, the dataset may lack the exploratory richness needed for effective policy learning. As discussed

in the broader thesis introduction, for RL to learn optimal counterfactual policies, we propose that patient trajectory datasets should exhibit *diversity in observed actions* that correlates with differences in outcomes conditioned on a particular state.

In the RL formulation shown in Fig. 3.1, we assume that for a given state s_t we can estimate not only the cumulative reward of taking observed action a_t, but also the reward for taking a different action a'_t. This would allow the offline-trained RL agent to accurately choose between a_t or a'_t despite having only observed directly the results of the former action. Since directly proving

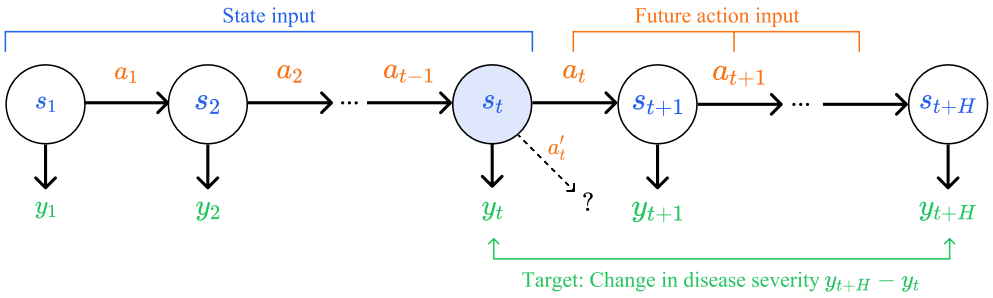

Figure 3.1: Markov decision process model for patients with sepsis in the ICU. s_t represents the patient state at time t, a_t represents a treatment action, and y_t represents a function of the state that captures the patient's disease severity. Brackets indicate how these values are used in our experiment.

the impossibility of learning effective policies is challenging, we approach this question by focusing on the relationship between patient states and treatment actions in these open datasets. Specifically, we utilize *dynamics models* built with state-of-the-art transformer architectures to investigate whether treatment actions contain additional predictive information beyond what is available in the patient state. Our reasoning follows the framework outlined in the thesis introduction: if clinician actions are diverse and have an effect on outcomes, then the action information should improve a model's ability to predict future observed disease severity.

This chapter presents a detailed exploration of action diversity in sepsis treatment datasets. By examining the predictive power of treatment actions, we provide a concrete case study demonstrating how action diversity can be measured and interpreted in a high-stakes medical domain. Our findings have important implications for the feasibility of reinforcement learning in this context and contribute directly to the thesis goal of developing methods to identify when and where learned treatment policies can be trusted.

3.2 Methodology

3.2.1 Data and Preprocessing

Patient trajectory data was extracted following [14] and [13] from MIMIC-IV [11] and the eICU Collaborative Research Database [18].[1][2] Data was aggregated at one-hour intervals, and patients with more than 14 days in the ICU were excluded. Missing data was imputed using a transformer-based autoencoder model. This resulted in a total of 2,060,446 timesteps from 33,779 patients.

We employed a transformer model to impute missing values, in contrast to prior approaches which use a combination of carrying forward the last known value and k-nearest neighbor imputation. To ensure the robustness and generalizability of the imputation model, it was trained on the available data and evaluated using artificially created missing values.

The state space for our models consisted of 60 normalized observation variables (vitals, labs, prior treatments, and fluid balances) and 35 demographic variables (age, gender, and Elixhauser comorbidities). The action space comprised log-transformed continuous-valued dosages of IV fluids and vasopressors. Three widely-used severity metrics were used as outcomes: the Sequential Organ Failure Assessment (SOFA) score, the Systemic Inflammatory Response Syndrome (SIRS) score, and Shock Index. Actions and disease severity were z-transformed for model input and output.

3.2.2 Models

Model Selection

To assess the suitability of different models for predicting patient outcomes, we conducted a comparative study using transformer-based, RNN-based, and linear-regression-based models. This experiment was conducted using a subset of the MIMIC-III [**?**] database containing patients with sepsis.

We trained and evaluated sixteen models in total: two transformer models (with 4 and 16 attention heads), one RNN model, and one linear regression model—each tested across four different embedding sizes. We used the Adam optimizer and negative log-likelihood as the loss function, with training conducted over 20 epochs and a batch size of 32. Model performance was assessed using both training and validation loss.

[1]While previous work has generally used MIMIC-III, the AI Clinician modeling procedure has been shown to yield consistent results in the two versions of MIMIC [19].

[2]Preprocessing and modeling code available at `https://github.com/cmudig/AI-Clinician-MIMICIV`.

Dynamics models

Our experiment utilized decoder-only transformer models, where each input "token" comprised embeddings of the patient's observed state, demographics, and actions.[3] The model consisted of two transformer blocks, each comprising 4 self-attention layers, each with 16 attention heads and a total dimension of 1024. The first transformer block took the state and demographic embeddings as input, while the second transformer block added embedded clinician actions. Models were trained on the future disease severity task along with three other proxy tasks: (1) predicting the current state of the patient, (2) predicting whether the current state is the last step in the patient's trajectory, and (3) predicting whether two embeddings correspond to states that are adjacent in time. The proxy tasks were included only to improve the model's convergence and generalizability, and results for these tasks are not shown.

The dynamics models were trained using a multitask learning approach, using two regression tasks and two binary classification tasks: (1) predicting the current state of the patient, (2) forecasting future disease severity (the target of interest), (3) predicting whether the current state is the last step in the patient's trajectory, and (4) predicting whether two embeddings correspond to states that are adjacent in time. Losses from the four tasks were aggregated and used to train the dynamics model until loss did not decrease for three epochs (between 10-20 epochs per model). Both regression tasks were trained using mean squared error (MSE) loss, while binary cross-entropy loss was utilized for the binary classification tasks.

Behavior cloning

While the dynamics models described above aimed to predict the difference in disease severity as a function of states and actions, we also trained behavior cloning models to predict clinician actions as a function of states. These models utilized the first transformer block from above to encode the state observations and demographics, then applied a two-layer feedforward network to simultaneously predict fluid and vasopressor dosages at one-hour intervals up to 6 hours ahead. Models were trained with MSE loss until the loss did not decrease for three consecutive epochs.

Behavior Cloning is a method based on imitation learning, which tries to predict which action a clinician would take given a patient state. It's different from the other methods in that it doesn't try to improve on the clinician action, but rather to replicate it. We can use this as a baseline for other deep learning-based methods.

3.3 Results

3.3.1 Dynamics model Selection

As shown in Figure 3.2, transformer-based models consistently demonstrated lower training and validation losses compared to RNN and linear models. Among the transformer variants, models

[3]We conducted the same experiments with linear and recurrent networks as well as XGBoost models, but found that transformers yielded the best performance.

with 16 attention heads generally outperformed those with 4 heads. For both RNN and linear models, an embedding size of 256 minimized losses. Error bars in the RNN loss curves suggest higher variance, indicating less stable performance. Based on this analysis, we selected the transformer-based dynamics model with 16 heads and an embedding size of 512 as the most appropriate model for downstream evaluation.

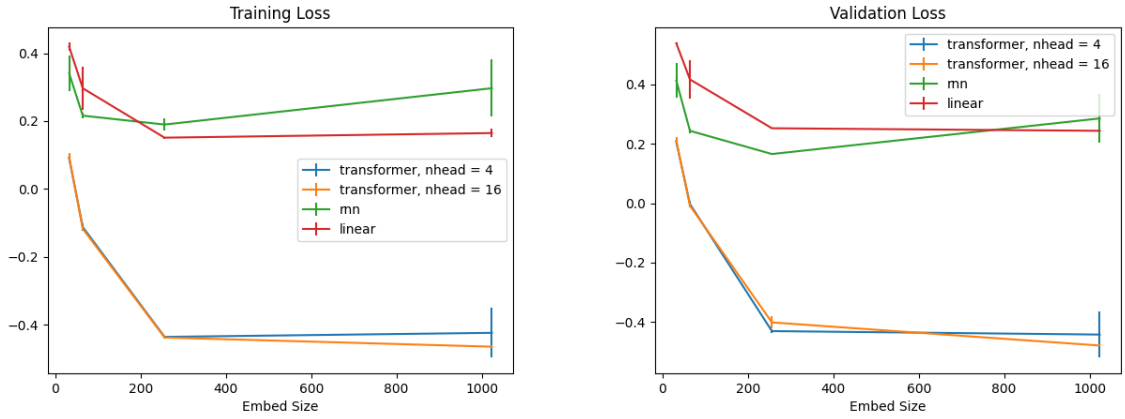

Figure 3.2: Training and validation loss for different dynamics models.

3.3.2 Influence of Action Inputs on Disease Severity Predictions

To assess the impact of incorporating clinician actions on the predictive performance of transformer-based dynamics models, we conducted an experiment comparing models trained with and without action data. Our objective was to determine whether the inclusion of treatment actions enhances the models' ability to predict future disease severity.

We trained a total of 81 dynamics models across three experimental groups to predict changes in future disease severity. The first group was trained with both patient state and future action information. The second group, with identical architectures, had all future actions set to mean values (effectively removing them from training). The last group also shared identical architectures, but was trained without the information about states.

For each configuration, we predicted disease severity changes using three metrics (SOFA, SIRS, and Shock Index) at three future time points (6, 12, and 18 hours ahead). Each model configuration was trained and evaluated across three random weight initializations. We conducted four distinct evaluations by generating predictions on different variants of the test dataset:

- **True**: Using the original treatment actions from clinical records
- **Zero**: Setting all dosage values to zero
- **Shuffled**: Using real but randomly permuted dosages
- **Mean**: Replacing all actions with mean dosage values

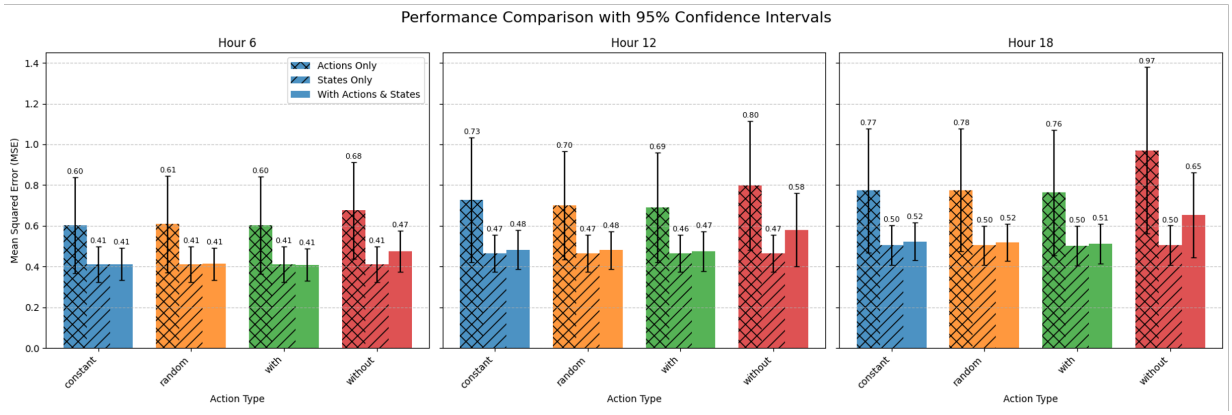

Figure 3.3: Left: RMSE (lower is better) of the predicted change in disease severity across training schemes ("Train Actions", "Train States," and "Train States + Actions") and action inputs at test time (**True**, **Zero**, **Shuffled**, and **Mean**). Error bars indicate the standard deviation across three random weight initializations. Note that all units are in z-scaled space, so an RMSE of 1 corresponds to 1 standard deviation in the severity metric. Right: example histograms comparing true and predicted changes in SOFA score at 12 hours ahead, in the **True** and **Shuffled** evaluation conditions.

Figure 3.3 presents the root mean squared error (RMSE) of these predictions in z-scaled space, along with example comparisons between model predictions and ground-truth values. Overall, RMSE remained nearly constant across training conditions and action input types, with the exception of the **Mean** condition. The **Mean** condition generally exhibited higher error and greater variance across initializations when actions were included in training, likely because consistently receiving nonzero fluids and vasopressors represents a highly unusual clinical scenario. Among the other three conditions (**True**, **Zero**, and **Shuffled**), the range of RMSEs was within 0.05 for SIRS and Shock Index, and within 0.1 for SOFA. Notably, performance in the **True** condition was highly similar whether or not actions were provided during training. This null result suggests that actions did not substantially improve model fit, consistent with our hypothesis that they are not diverse enough for policy learning. Additionally, models trained without state information showed similar trends, indicating that action information is largely redundant with patient states.

For models trained with action data, the MSEs of validation datasets with altered action information (**Zero**, **Shuffled**, **Mean**) were higher compared to the **True** validation dataset. This indicates that inaccurate or manipulated action information negatively affects predictive performance. Conversely, for models trained without action data, MSEs remained consistent across all validation datasets, confirming that these models do not utilize any information regarding clinician actions.

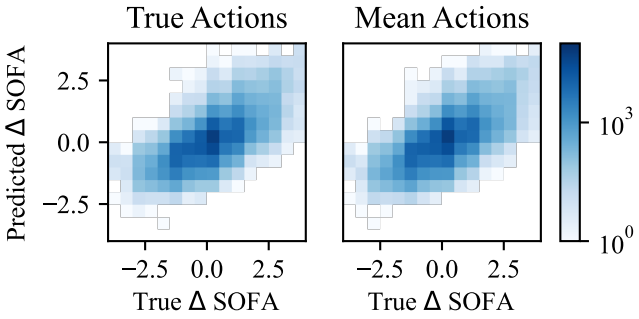

Figure 3.4: Example histograms comparing the true change in SOFA score to the predicted change at 12 hours ahead, when the model was given the **True** action compared to the **Mean** action.

3.3.3 Prediction of Future Actions with Behavior Cloning

To directly evaluate the predictability of actions from states, we trained three replicates of a behavior cloning model with different random weight initializations. If these models showed a strong fit to the data, it would suggest that actions were fully consistent and predictable across clinicians.

Fig. 3.5 shows that the average R^2 correlations between the true and predicted actions (in log-transformed and z-scaled units) were generally low, particularly after several hours. IV fluid predictions were notably less correlated with the true values than vasopressor predictions. This difference may be due to two factors: (1) vasopressors are more commonly zero than fluids, increasing the overall predictability of vasopressor use, or (2) the amount of IV fluid used is generally more clinician-dependent. The regression models also appeared to struggle with the wide range of fluid dosage values, tending to predict values within a more constrained range (Fig. 3.5, third panel).

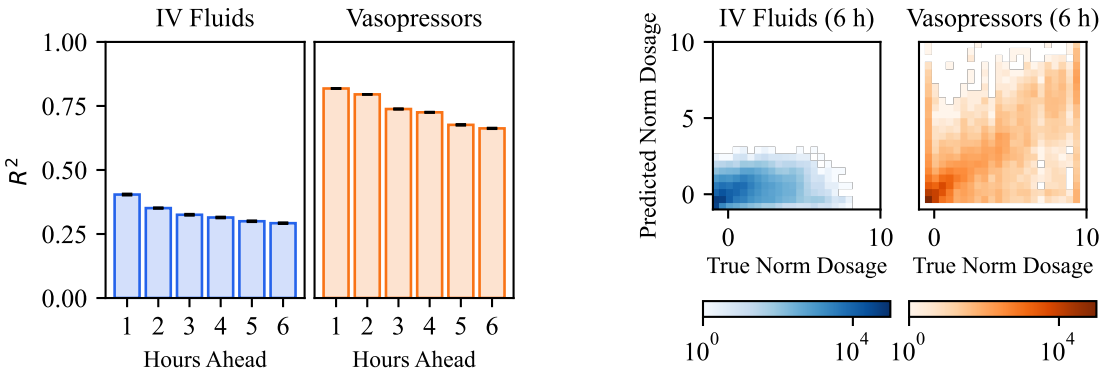

Figure 3.5: Left: correlations between true and predicted normalized actions from 1 to 6 hours ahead. Right: example histograms of correlations between true and predicted normalized actions at 6 hours.

Aside from the possible modeling issues in the IV fluid predictions, the low correlations across both treatments suggest there is in fact some diversity in clinician actions that could benefit policy learning. However, action diversity does not necessarily correspond to observable differences in outcomes, since there is likely a range of treatment dosages that correspond to similar effects for a given patient state. The results in the preceding section suggest that even when dosage differences exist, they may not yield sufficient differences in outcomes to provide a useful signal to an RL agent.

3.4 Discussion

This work explored the impact of clinician actions on the predictability of future changes in sepsis disease severity, seeking to determine whether actions have sufficient diversity to support accurate RL-based policies. Our findings revealed that action information does *not* confer substantive improvements in dynamics model fit. Transformer models could predict future disease severity almost equally well with or without true actions as input.

Taken alone, the dynamics model results in Section 3.3.2 might suggest that actions are fully predictable from patient states, and there was no need to learn from the action inputs. This observation echoes results from [2], who critique patient risk predictions as "looking over the shoulders of clinicians." However, our action prediction results (Section 3.3.3) showed fairly noisy predictions, indicating that while variation in actions exists, it is not sufficient to cause measurable differences in outcomes in our sepsis cohort. Rather, the outcome differences we observe may be more driven by unobserved patient variables or natural random variation.

The observed lack of diversity in actions within MIMIC data may stem from several inherent challenges in working with patient trajectories:

- There may only be a small number of treatment possibilities that are clinically feasible and safe, limiting the space of actions that clinicians could take
- Clinicians may follow predictable treatment patterns (such as monotonically increasing or decreasing dosages) that appear diverse yet lead to consistent outcomes
- Missing data imputation could have caused patient states and actions to appear more consistent than they truly are

These obstacles are likely to exist in any patient treatment dataset, underscoring the importance of using learning methods that are robust to missingness and a constrained action space.

Another possible explanation for our results is that our models simply didn't learn to use actions effectively, and a better model formulation might yield more pronounced differences between the "Train States" and "Train States + Actions" models. While it is impossible to determine *a priori* whether there exists a more effective way to use actions, we conjecture that if such a method exists, it would likely require more clinically-informed descriptions of actions than what has currently been explored in the literature. For instance, models could use other treatments such as antibiotics and mechanical ventilation, contextualize actions using the patient's physiological state, or limit the training data to only the most important decision points. Future work

should incorporate clinician guidance on how to meaningfully encode treatments to further test the effects of action information.

This work highlights the importance of diversity in data sources when building medical recommendation models. While it has been extremely valuable in developing and exploring ways to improve sepsis treatment, the MIMIC dataset is sourced from a single well-resourced hospital in Boston [11], where clinicians are likely to be consistent and compliant with existing practice guidelines. Human-centered ML efforts undertaken in collaboration with clinicians and medical data experts can also inspire more clinically-relevant and performant model formulations, such as focusing on the emergency department (a higher-stress environment that is less specialized towards sepsis than the ICU) or building smaller models that are relevant to specific subgroups of patients [19]. Through these research directions, applied ML efforts may be able to better utilize available observational data to improve sepsis treatment recommendation while accounting for the inherent limitations in action diversity present in clinical datasets.

Chapter 4

Analyzing Action Diversity Through Controlled Simulation

4.1 Introduction

In the previous chapter, we examined whether reinforcement learning approaches could effectively learn treatment policies for sepsis using observational data from the MIMIC dataset. Our analysis revealed a significant challenge: when evaluated across the entire dataset, treatment actions did not substantially improve the predictive performance of dynamics models for patient outcomes. This finding suggested that either clinician actions were highly predictable from patient states, or that the observable impact of diverse actions on outcomes was limited. However, this aggregate analysis may obscure important heterogeneity within the dataset. A critical question remains: Could there exist specific subgroups of patients or clinical contexts where action diversity is more pronounced and has measurable effects on outcomes?

To address this question, we require methods capable of systematically identifying and characterizing regions of action diversity. However, developing and validating such methods directly on complex clinical data presents significant challenges. The high dimensionality of patient states, the presence of unmeasured confounders, and the intricate relationships between treatments and outcomes make it difficult to isolate the effects of action diversity. Furthermore, without ground truth knowledge, it is challenging to evaluate whether our methods correctly identify regions of meaningful action diversity.

In this chapter, we take a step back and adopt a controlled simulation approach. We created a simplified GridWorld-based environment which allows explicit control over state-action relationships. This allows us to systematically study patterns of action diversity and evaluate methods for detecting these patterns.

4.2 Methodology

4.2.1 Simulation Environment Design

We implemented a one-dimensional GridWorld using the Gymnasium framework, consisting of N discrete states, where each state represents a simplified abstraction of a patient's state. States are arranged sequentially from 0 to $N-1$. The environment supports multiple actions, simulating different treatment options. Each action is assigned a specific reward value that determines the patient's next state. For example, an action with a reward of +2 would shift the patient's state two positions forward, while an action with a reward of -1 would move the patient one position backward.

We implemented two distinct state types to model different patterns of action diversity:

- **Action-diverse states:** States where patient trajectories are directly affected by action choices. In these states, the next state transition is determined by the specific reward value associated with the selected action.
- **Non-diverse states:** States where patient trajectories follow their own fixed probability pattern regardless of the action taken. In these states, the transition is determined by a state-specific reward value, and the action input is ignored.

This binary classification of states allows us to create a controlled environment with regions of clear action diversity and regions where actions have no impact on outcomes. Each state is randomly assigned one of these two types during environment initialization.

Our simulation environment defines state transition dynamics that vary depending on whether the current state is action-diverse or not. For action-diverse states, the state transition is calculated as $s_{t+1} = s_t + \text{reward}[a_t]$, where reward is a vector mapping each action to a specific reward value. These reward values are randomly initialized within the range $[-2, 2]$ and determine how far and in which direction the patient's state moves. For non-diverse states, the state transition is calculated as $s_{t+1} = s_t + \text{reward}[s_t]$. If a transition would move the patient's state outside the valid range $[0, N-1]$, the trajectory is terminated and marked with a special terminal state value. At each timestep, actions were selected with equal probability from the available action space, regardless of the state type. This approach provides a baseline scenario where all actions have equal representation in the dataset, allowing us to focus on the effects of state-dependent action efficacy rather than action selection bias.

4.2.2 Synthetic Dataset Generation

Using the defined GridWorld environment, we generated synthetic datasets to evaluate our methods for detecting action diversity. Each dataset consisted of multiple simulated trajectories, with each trajectory beginning from a randomly selected initial state. At each timestep, the current state s_t was recorded, an action a_t was selected uniformly at random, and the next state s_{t+1} was determined according to the transition dynamics. The process was repeated until the trajectory reached its maximum length or transitioned to a terminal state.

The resulting dataset captured a detailed history of each simulated patient trajectory. For every step, we recorded the trajectory ID, the timestep, the current state, the ground truth state type, the selected action, and the resulting reward. In addition to the trajectory-level data, we stored metadata describing the environment configuration, including the assignment of state types, the reward mappings, and the simulation parameters used for each dataset. This structure allowed for consistent evaluation and interpretability across different experimental runs.

4.2.3 Dynamics Model

Model Architecture

Following the same methodological approach used in Part 1 of this thesis, we trained two variants of predictive models to detect action diversity: one that incorporates action information and one that does not. The with-actions models take both state and action inputs to predict future patient states or state changes. In contrast, the without-actions models use only the state as input, with the action channel replaced by zeros. Comparing the performance of these two variants allows us to identify regions where action information contributes predictive value.

We implemented this approach using XGBoost regression models, which provide a fast and interpretable baseline for learning state transitions. Inputs were one-hot encoded, and the target variable was the change in state over a specified future interval.

To assess short- and long-term effects, both model types were trained to predict outcomes at multiple future horizons, ranging from one to four timesteps ahead. This involved shifting the target state forward and predicting the delta. This setup enabled a comprehensive evaluation of the robustness of action diversity detection across architectures and prediction horizons.

Training and Evaluation

All synthetic datasets were split into training (50%) and validation (50%) sets based on trajectory IDs. XGBoost models were trained using default hyperparameters with the 'hist' tree method. The models were trained to minimize mean squared error (MSE) between predicted and actual future states or state changes.

To evaluate model performance, we computed root mean squared error (RMSE), which we calculated for both with-action and without-action models on the training set, the full validation set, and separately on action-diverse and non-diverse subsets of the validation data. By comparing the performance metric, we can identify where action information significantly improves predictions, indicating meaningful action diversity.

4.2.4 Classifying State Diversity via Predictive Model Error Differences

To further validate whether action diversity can be detected from model behavior, we trained a classifier to predict whether a given state is action-diverse or not, based on the performance of dynamics models.

Dataset Construction for Classification

We first constructed a new dataset based on the outputs of the with-actions and without-actions XGBoost models described above. For each timestep in each trajectory, we recorded the absolute prediction errors from both models. We then computed a boolean label indicating which model performed better. These features were then joined with the ground truth `StateType` labels from the simulation to create a labeled classification dataset suitable for training and evaluation.

Classifier Training

We trained a simple XGBoost binary classification model to predict whether a given state is action-diverse or non-diverse. The input feature vector included the absolute prediction errors from both models, the difference between the two errors, a binary indicator of which model had lower error, and one-hot encoded representations of the current state and action. The classifier was trained on 70% of the data, with the remaining 30% held out for validation. Model performance was evaluated using classification accuracy and the area under the ROC curve (AUC), allowing us to assess the effectiveness of model behavior as a proxy for detecting action diversity.

4.3 Results

To evaluate whether action information improves the predictability of patient state transitions, we conducted a series of experiments using synthetic data generated by our GridWorld simulation environment. Specifically, we compared two types of dynamics models—one that included action inputs and one that did not—over multiple prediction horizons. In addition, we trained a model to classify states as action-diverse or not, based on the relative performance of the dynamics models.

4.3.1 Simulation Behavior and Dataset Properties

Our simulation environment successfully generated synthetic patient trajectories with distinct action-diverse from non-diverse states. Each environment consisted of 20 discrete states, with an equal split: 10 states were randomly designated as action-diverse and 10 as non-diverse. In action-diverse states, transitions depended on the chosen action, whereas in non-diverse states, transitions were determined solely by the current state and were independent of action. This setup provided a well-defined ground truth structure for evaluating model performance.

4.3.2 Dynamics Model Performance

XGBoost Models

We evaluated XGBoost-based dynamics models by measuring their root mean square error (RMSE) in predicting state transitions across various horizons and state types. As shown in Table 4.1, models trained with action inputs consistently outperformed their counterparts in action-diverse states, but underperformed in non-diverse states.

Table 4.1: RMSE values for XGBoost models across 1–4 future timesteps.

Horizon	**Action-Diverse**		**Non-Diverse**	
	With Action	Without Action	With Action	Without Action
1-step	1.42	2.53	3.59	3.28
2-step	2.64	3.21	3.74	3.39
3-step	3.38	3.60	3.54	3.11
4-step	3.54	3.79	3.66	3.42

In action-diverse states, incorporating the action input significantly improved model performance across all time horizons. For example, at the 1-step horizon, the RMSE dropped from 2.53 (without action) to 1.42 (with action). Although the performance gap narrowed over longer horizons, the advantage of using action remained consistent; at 4 steps ahead, the RMSE was 3.54 with action compared to 3.79 without.

In contrast, the trend reversed in non-diverse states: here, models with action input tended to perform slightly worse. For instance, at the 1-step horizon, the RMSE increased from 3.28 (without action) to 3.59 (with action). This pattern was consistent across all horizons, suggesting that including irrelevant action information introduces noise, worsening predictive accuracy in states where action has no effect.

Figure 4.1 illustrates the relationship between model predictions and true values, with each point corresponding to a particular state and timestep. We can see that the predictions are more accurate and correlated for the models trained with action inputs compared to those trained without action. The scatterplots are organized into three columns: the leftmost column shows all states in the validation set, while the right two columns separate the states into action-diverse and non-diverse categories, as determined by our classifier. For action-diverse states—where varied actions lead to varied rewards—models trained with action inputs show more accurate and tightly correlated predictions. In contrast, for non-diverse states—where actions do not significantly affect outcomes—there is less correlation between predictions and true values, and the inclusion of action can even degrade performance. The clustering patterns reveal that in diverse states, actions meaningfully inform transitions, whereas in non-diverse states, they do not.

Together, these results support our central hypothesis: action inputs improve dynamics model performance only in regions where actions influence transitions. When action is relevant, models that incorporate it produce more accurate predictions. When it is irrelevant, including it

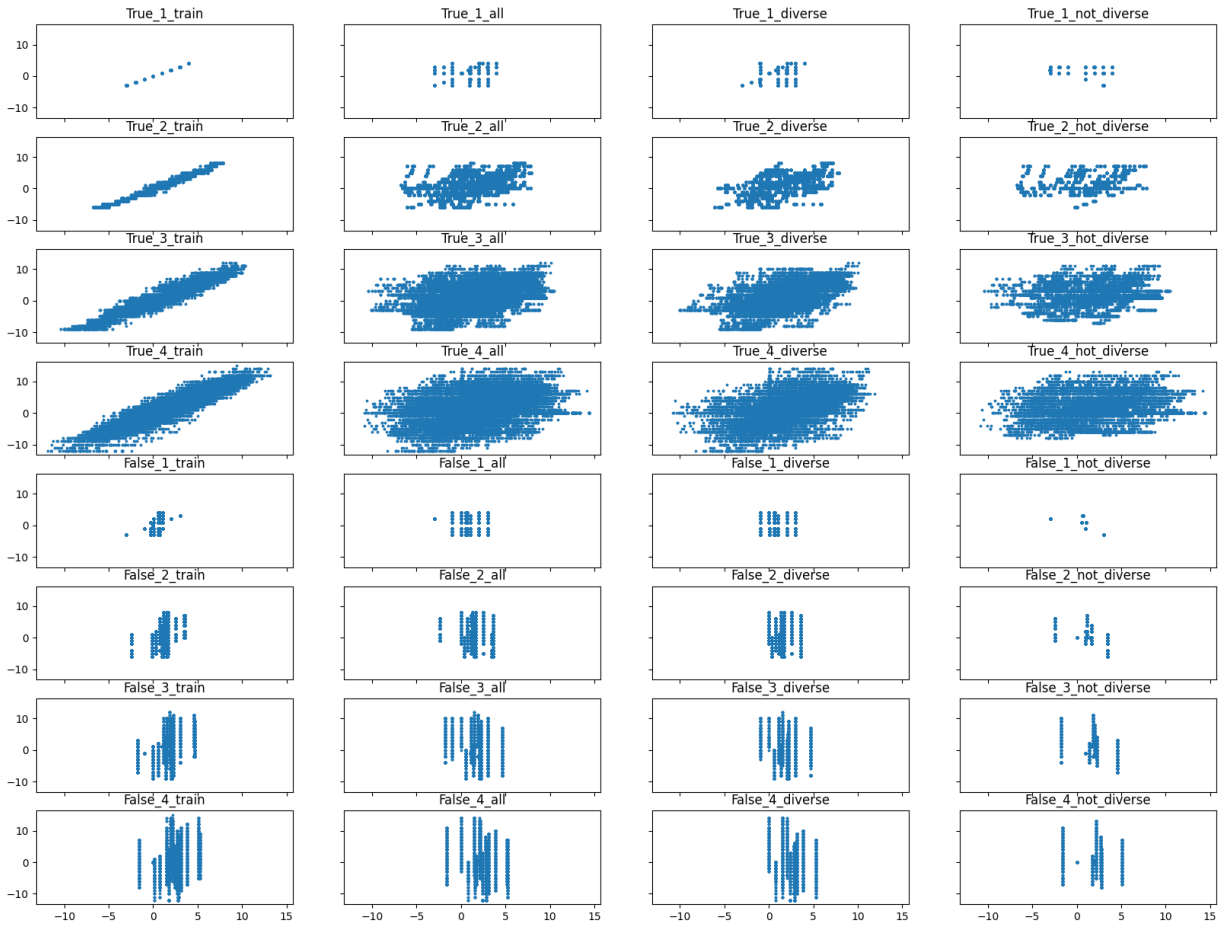

Figure 4.1: Scatterplot of model prediction errors across diverse and non-diverse states.

can harm performance. These insights emphasize the importance of selectively modeling action effects—an idea we further explore through the use of a classifier.

4.3.3 Action Diversity Classifier Performance

To test whether model performance differences can be used to identify action-diverse states directly, we trained a classifier to predict state type using features derived from dynamics model errors. For each timestep, we computed the absolute errors from both models, their difference, and an indicator of which model performed better. These were combined with one-hot encodings of the current state and action to form the feature set for an XGBoost classifier.

The classifier achieved an accuracy of 86% on a held-out validation set for 1 timestep, indicating that model behavior can be reliably used to infer underlying state properties. Table 4.2 summarizes the classification performance across different prediction horizons. These results demonstrate that action diversity can be inferred from model performance patterns. This provides a promising foundation for identifying subpopulations in real clinical data where machine

learning methods may be more effective and trustworthy.

Table 4.2: Classifier performance over different future prediction intervals.

Time (hrs)	Train Accuracy	Validation Accuracy	ROC AUC
1	0.872	0.861	0.832
2	0.803	0.791	0.769
3	0.802	0.809	0.692
4	0.796	0.819	0.686

4.4 Discussion

This section introduced a novel framework for assessing when observational medical data is suitable for learning treatment policies. Through a controlled simulation environment, we demonstrated that the difference in predictive performance between models trained with and without action input reliably signals whether a state is action-diverse.

In regions where actions affect outcomes, models incorporating action information performed better. In contrast, in states where actions had no effect on transitions, including action input provided no benefit. These patterns were consistent across model types and prediction horizons.

Building on these observations, we constructed a second-level model to classify state diversity based on dynamics model outputs. Our classifier achieved high accuracy using simple features derived from prediction error differences. This result suggests that we may be able to identify reliable subspaces for learning purely from model behavior.

While our experiments were conducted in a controlled synthetic environment, which enabled access to ground truth labels, the real-world applicability of this framework remains to be tested. Clinical datasets are noisier, less balanced, and subject to confounding factors that are not captured in our simulation. Applying this method to real data will be a key next step in validating its utility in practice. Additionally, future work could extend this framework to account for continuous state spaces.

Chapter 5

Exploring Action Diversity Through Interactive Visualization

5.1 Introduction

The previous chapters of this thesis investigated action diversity in healthcare datasets through aggregate statistics and controlled simulations. The results from Parts 1 and 2 of this thesis highlight a critical challenge in applying reinforcement learning to healthcare: the variation in treatment choices across similar patient states is often limited and unevenly distributed across the state space. While our simulation experiments demonstrated that it is possible to detect regions of meaningful action diversity using model-based signals, applying these techniques to real clinical datasets introduces additional complexity. Real-world data is high-dimensional, noisy, and lacks ground truth annotations, making it difficult to interpret where and why action diversity exists.

In this chapter, we present an interactive visualization tool designed to address this challenge. By combining dimensionality reduction techniques with intuitive visual representations of treatment patterns and model performance, this tool enables researchers to identify and explore regions where treatment diversity exists and meaningfully impacts patient outcomes. This approach complements our earlier methods by providing a more exploratory and human-centered approach to detecting action diversity.

5.2 Methodology

5.2.1 Data Preparation

The foundation of our visualization tool is the MIMIC dataset. For our analysis, we extracted sequential patient data including demographics, vital signs, laboratory measurements, medications, and SOFA scores. Each datapoint represents a specific patient state at a given timestep, identified by a unique patient ID and timestep indicator. The SOFA score, ranging from 0 to 24 with higher values indicating greater organ dysfunction, was used as our primary clinical outcome measure.

Dimensionality Reduction

A transformer-based autoencoder with 4 encoder layers, 8 attention heads, and 32-dimensional embeddings was employed to encode the high-dimensional clinical states into latent representations. These encodings were further reduced to two dimensions using t-SNE with cosine similarity as the distance metric, enabling visualization of patient states in a 2D space while preserving the complex relationships between clinical variables.

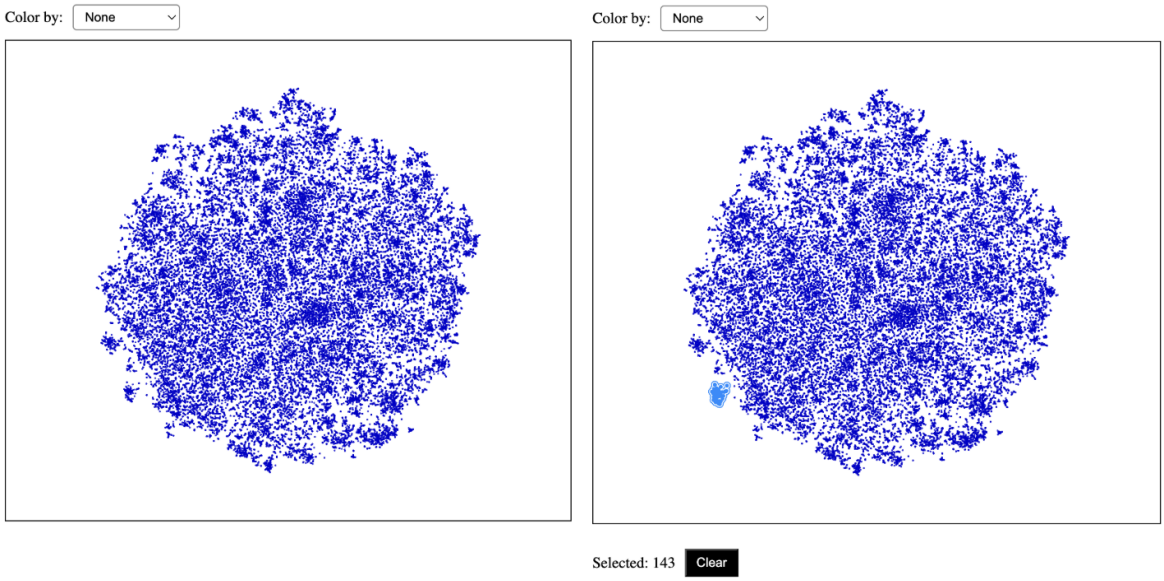

Figure 5.1: Left: t-SNE scatterplot of patient states. Right: Interactive selection of a patient subgroup.

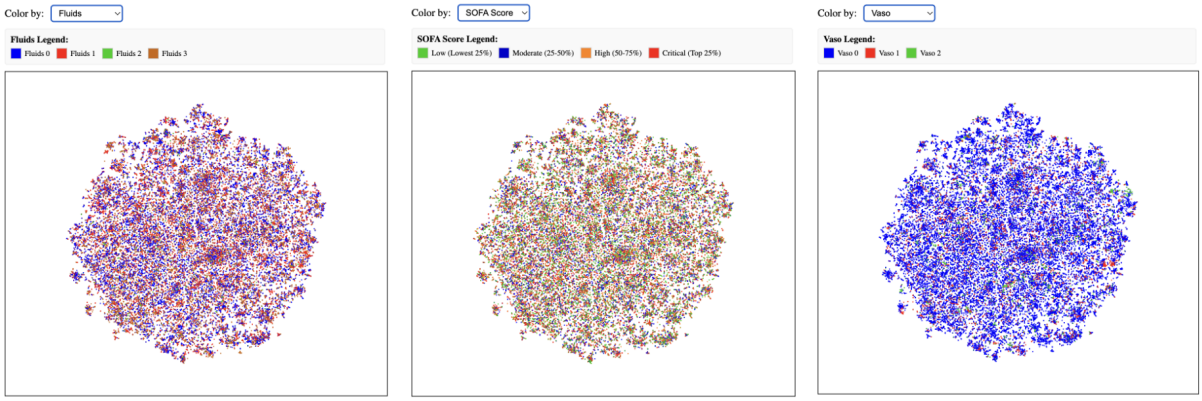

Figure 5.2: Color-coding of scatterplot by (left to right): fluid dosage, SOFA score severity, and vasopressor level.

Outcome Prediction Models

To evaluate the role of treatment actions in predictive modeling, we trained two outcome prediction models. A with-action model that predicts future SOFA scores using both the patient state and treatment action as input, and a without-action model that uses only the patient state. Each model was applied to every datapoint to generate predicted SOFA scores, which were then used to assess the impact of action inputs on model performance.

Data Integration and Storage

The processed dataset was stored in Google Cloud BigQuery to support efficient querying and scalability. Each row in the dataset includes the following fields:

- Patient ID and timestep
- 2D coordinates from t-SNE projection
- Administered treatments (fluid and vasopressor levels)
- AI-recommended treatments (from the AI Clinician)
- Predicted SOFA scores from both the with-action and without-action models
- Actual observed SOFA score

The AI Clinician recommendations were derived from a reinforcement learning model trained on the MIMIC dataset to optimize patient outcomes, as detailed in [14].

5.2.2 Visualization Components

The core of our tool is a suite of interactive visualizations that enable researchers to explore the dataset dynamically. These components update in real-time based on user interaction, allowing for intuitive exploration of patient subgroups and model behavior.

t-SNE Scatterplot

The 2D scatterplot visualizes individual patient states using the t-SNE coordinates. Each point represents a single patient-timestep. Color-coding schemes allow users to highlight different clinical attributes:

- SOFA score severity (quartile-based gradient from green to red)
- Fluid dosage level (4-level categorical scale)
- Vasopressor dosage level (3-level categorical scale)

Users can interactively select clusters via click-and-drag, which dynamically updates the treatment and model performance visualizations described below as shown in figures 5.1 and 5.2).

Treatment Heatmaps

Two side-by-side 4×3 grid heatmaps represent treatment distributions for the selected patient states:

- **Actual Treatment Distribution:** Shows the frequency of administered fluid (0–3 levels) and vasopressor (0–2 levels) combinations.
- **AI-Recommended Treatment Distribution:** Displays the treatments that the AI Clinician would have recommended for the same patient states.

Color intensity reflects normalized frequency, and the scale is shared between the two heatmaps to allow for direct visual comparison. Each cell also displays the precise count value.

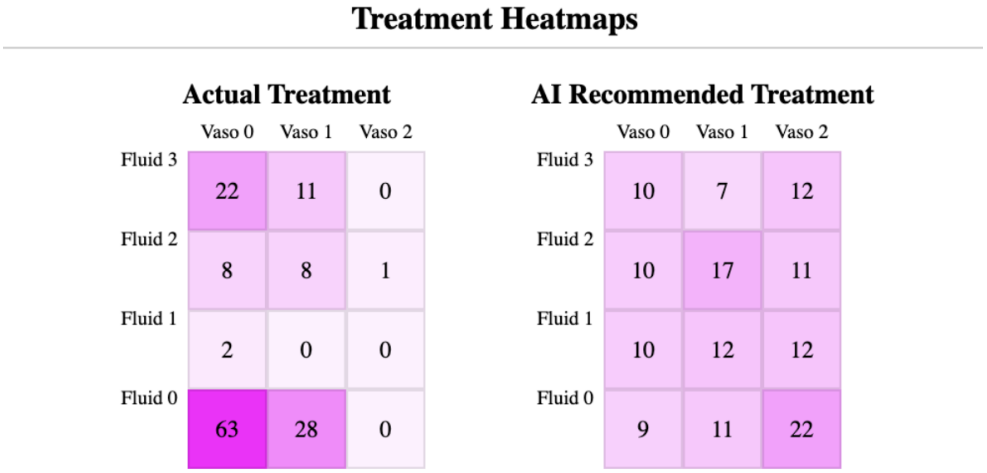

Figure 5.3: Comparison of actual vs. AI-recommended treatments across selected patient states.

Prediction Error Histograms

To compare the performance of the with-action and without-action models, we provide two histograms showing the distribution of SOFA score prediction errors:

- **With-Action Model:** Error = Actual SOFA - Predicted SOFA (with action)
- **Without-Action Model:** Error = Actual SOFA - Predicted SOFA (without action)

The x-axis represents prediction error, and the y-axis shows frequency count. Binning and axis scaling are kept consistent across the two plots to facilitate direct comparison.

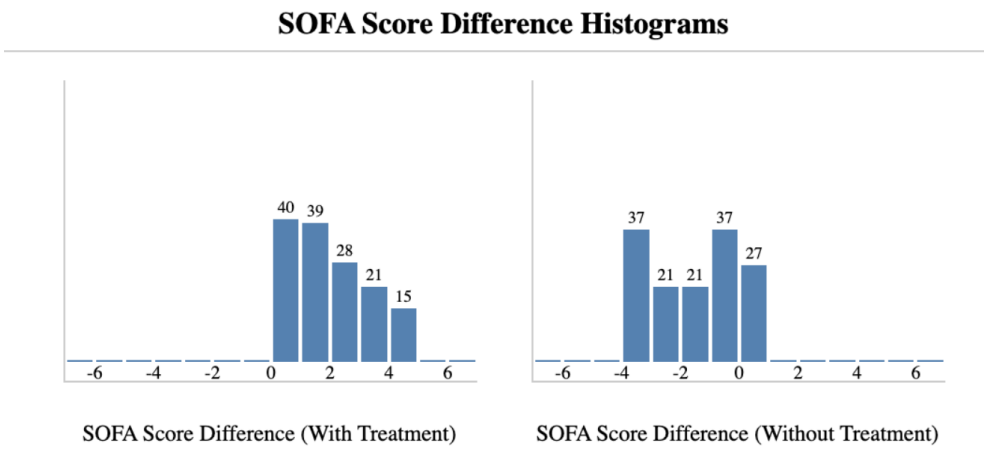

Figure 5.4: Histogram comparison of prediction errors for with-action and without-action models.

5.2.3 Implementation Details

The interactive visualization tool is implemented using a combination of web technologies:

- **Svelte** for UI reactivity and state management
- **regl** for WebGL-based high-performance rendering
- **D3.js** for data-driven SVG and DOM manipulation

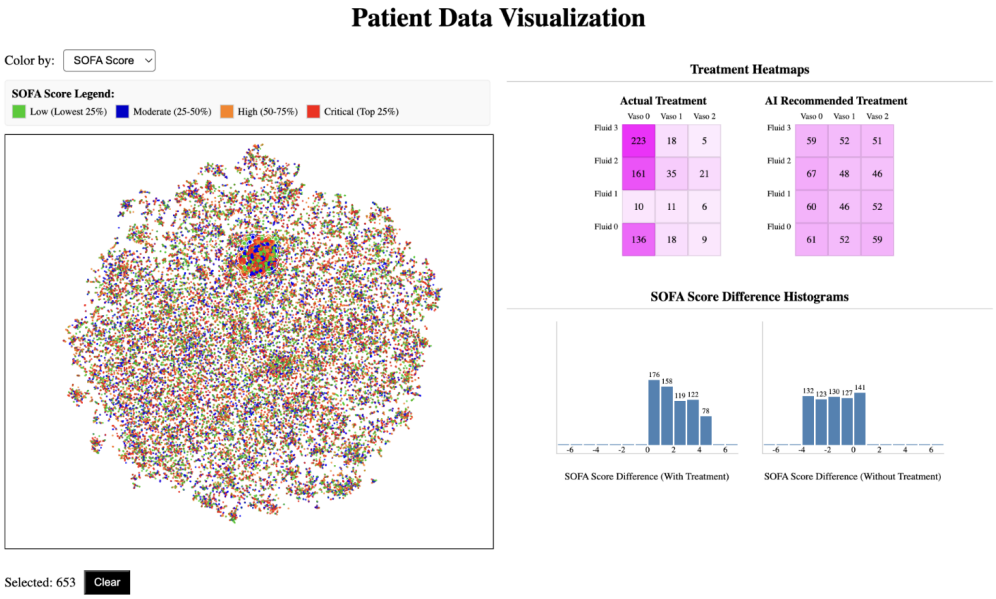

Figure 5.5: Overview of the complete visualization interface, including scatterplot, treatment heatmaps, and error histograms.

5.3 Results

We demonstrate the capabilities of our interactive visualization tool through example use cases that highlight its utility for exploring action diversity and treatment variability in clinical data. The results below illustrate how different components of the system work together to support intuitive, flexible, and interpretable analysis.

5.3.1 Use Case Demonstrations

Understanding Patient State Space

Using t-SNE projections of transformer-encoded latent states, the scatterplot component revealed coherent structure in the clinical state space. Clusters of similar patient states—particularly those with similar SOFA score ranges—emerged naturally. Color-coded overlays enabled rapid identification of high-risk regions, and interactive selection allowed focused exploration of specific subgroups. For example, users could isolate a cluster of high-severity patients (top SOFA quartile) and examine their treatment patterns and model prediction errors using the linked views.

Analyzing Treatment Patterns

The treatment heatmaps provided immediate visual feedback on discrepancies between clinician-administered treatments and AI Clinician recommendations. In selected clusters, actual treatments often skewed toward higher fluid usage, while AI suggestions leaned more conservative.

Maintaining consistent color scales across heatmaps made it easy to spot areas of agreement or divergence. This supported hypothesis generation around potential over- or under-treatment and contextualized clinician behavior.

Evaluating Model Performance

Prediction error histograms compared with-action vs. without-action models across patient subgroups. In many high-SOFA clusters, with-action models showed lower error variance—implying that treatments influenced predictive accuracy. In other areas, similar histograms indicated that actions had limited impact.

This functionality supports a deeper understanding of when treatment information is useful for outcome prediction, aligning with our broader thesis on the selective value of action data.

5.3.2 System Responsiveness and Usability

The system remained responsive with datasets exceeding 40,000 timesteps. Google Cloud BigQuery integration enabled efficient queries and dynamic updates during interaction.

Preliminary feedback from peers indicated that the visualizations made action diversity more

tangible. Users appreciated the ability to identify regions of disagreement between AI and clinician behavior, and the linked visualizations helped them interpret model performance in context.

5.4 Discussion

This work presents an interactive visualization framework designed to support exploration of action diversity in clinical datasets. By integrating a transformer-based autoencoder, dimensionality reduction via t-SNE, and interactive linked views, our tool enables researchers to identify regions of the patient state space where treatment actions significantly affect outcomes. Through intuitive interaction and real-time updates, users can isolate subgroups of interest, examine discrepancies between clinician and AI treatment strategies, and assess where treatment actions meaningfully impact predictive performance.

There are several possible directions for future improvement of the system. One opportunity is to extend the 2D scatterplot to a 3D projection using t-SNE or UMAP, which may better preserve structural relationships in the latent space and provide richer spatial context for exploration. Additional color-coding options—such as by patient ID, cluster assignment, or model performance—could offer users more flexible ways to interpret the data.

We also see potential for refining the visualization of model performance differences. For example, replacing side-by-side histograms with more integrated or compact alternatives (e.g., violin plots or density curves) may improve comparability and interpretability.

Finally, while this work focuses on tool development and internal analysis, future work could include formal user studies with clinical or machine learning researchers to assess usability, interpretability, and the tool's impact on real-world decision support or model validation workflows.

Chapter 6

Conclusion

Through three complementary approaches, we have demonstrated methods to detect and measure meaningful treatment variation in medical datasets, providing researchers with practical tools to evaluate when observational data can reliably inform treatment policies.

Our analysis of the MIMIC sepsis dataset using transformer-based dynamics models revealed limited meaningful treatment diversity when analyzed in aggregate. Including action information provided minimal improvement in outcome predictions, suggesting that the observable impact of treatment choices on outcomes may be constrained within this dataset. In our controlled simulation experiments, we successfully demonstrated that comparing prediction performance between models with and without action inputs can effectively identify regions where treatments meaningfully impact outcomes. This approach achieved high classification accuracy, providing a robust method for detecting action diversity without requiring ground truth labels. Finally, our interactive visualization tool offers researchers an intuitive way to explore action diversity across patient state spaces and identify promising subgroups for policy learning.

Together, these findings address a fundamental gap in the application of machine learning to healthcare: determining when datasets contain sufficient treatment variation to support reliable policy learning. By providing methods to identify regions of meaningful treatment variation, this work enables more targeted and trustworthy deployment of machine learning in healthcare. Researchers can focus on subpopulations where data supports reliable policy learning, potentially improving both the effectiveness of AI assistance in clinical decision-making and, ultimately, patient outcomes.

Future work can extend these methods to other clinical domains beyond sepsis, explore how to incorporate domain knowledge to enhance action diversity detection, and investigate approaches for augmenting datasets in regions with insufficient diversity. By continuing to develop robust methods for evaluating data sufficiency, we can ensure that machine learning approaches in healthcare are deployed responsibly, focusing on areas where they can provide the most reliable guidance to clinicians.

Bibliography

[1] Francesco Joseph Bargagli-Stoffi, Riccardo Cadei, Kewen Lee, and Francesca Dominici. Causal rule ensemble: Interpretable discovery and inference of heterogeneous treatment effects. *arXiv preprint arXiv:2009.09036*, 2024. URL https://arxiv.org/abs/2009.09036. 2.1.2

[2] Brett K. Beaulieu-Jones, William Yuan, Gabriel A. Brat, Andrew L. Beam, Griffin Weber, Marshall Ruffin, and Isaac S. Kohane. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *npj Digital Medicine*, 4(1): 1–6, March 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00426-3. URL https://www.nature.com/articles/s41746-021-00426-3. Publisher: Nature Publishing Group. 3.4

[3] William Caicedo-Torres and Jairo Gutierrez. Iseeu: Visually interpretable deep learning for mortality prediction inside the icu. *arXiv preprint arXiv:1901.08201*, 2019. URL https://arxiv.org/abs/1901.08201. 2.2

[4] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015. 2.2

[5] Centers for Disease Control and Prevention. What is sepsis?, 2021. 3.1

[6] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable deep models for icu outcome prediction. *AMIA Annual Symposium Proceedings*, 2016:371–380, 2016. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333206/. 2.2

[7] Omer Gottesman, Joseph Futoma, Yao Liu, Sonali Parbhoo, Leo Anthony Celi, Emma Brunskill, and Finale Doshi-Velez. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. *37th International Conference on Machine Learning, ICML 2020*, PartF16814:3616–3625, 2020. arXiv: 2002.03478. 2.1.1

[8] Jennifer Hill, Antonio Linero, and Jared Murray. Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7(1):251–278, 2020. doi: 10.1146/annurev-statistics-031219-041110. 2.1.2

[9] Russell Jeter, Christopher Josef, Supreeth Shashikumar, and Shamim Nemati. Does the "Artificial Intelligence Clinician" learn optimal treatment strategies for sepsis in intensive care? *arXiv*, November 2019. ISSN 1078-8956. doi: 10.1038/s41591-018-0213-5. arXiv:

1902.03271. 2.1.1, 3.1

[10] Christina X. Ji, Michael Oberst, Sanjat Kanjilal, and David Sontag. Trajectory Inspection: A Method for Iterative Clinician-Driven Design of Reinforcement Learning Studies. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2021(i):305–314, 2021. ISSN 1942597X. arXiv: 2010.04279. 2.1.1

[11] A Johnson, L Bulgarelli, T Pollard, S Horng, L A Celi, and R Mark. MIMIC-IV (version 1.0), 2020. 3.2.1, 3.4

[12] Song Ju, Yeo Jin Kim, Markel Sanz Ausin, Maria E. Mayorga, and Min Chi. To Reduce Healthcare Workload: Identify Critical Sepsis Progression Moments through Deep Reinforcement Learning. *Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021*, pages 1640–1646, 2021. doi: 10.1109/BigData52589.2021.9671407. Publisher: IEEE. 2.1.1

[13] Taylor W. Killian, Haoran Zhang, Jayakumar Subramanian, Mehdi Fatemi, and Marzyeh Ghassemi. An Empirical Study of Representation Learning for Reinforcement Learning in Healthcare. pages 1–22, 2020. URL http://arxiv.org/abs/2011.11235. arXiv: 2011.11235. 2.1.1, 3.2.1

[14] Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. ISSN 1546170X. doi: 10.1038/s41591-018-0213-5. URL http://dx.doi.org/10.1038/s41591-018-0213-5. arXiv: 1902.03271 Publisher: Springer US. 2.1.1, 3.2.1, 5.2.1

[15] Dayang Liang, Huiyi Deng, and Yunlong Liu. The treatment of sepsis: an episodic memory-assisted deep reinforcement learning approach. *Applied Intelligence*, 2022. ISSN 15737497. doi: 10.1007/s10489-022-04099-7. Publisher: Applied Intelligence. 2.1.1

[16] Thesath Nanayakkara, Gilles Clermont, Christopher James Langmead, and David Swigon. Unifying cardiovascular modelling with deep reinforcement learning for uncertainty aware control of sepsis treatment. *PLOS Digital Health*, 1(2):e0000012, 2022. doi: 10.1371/journal.pdig.0000012. URL http://dx.doi.org/10.1371/journal.pdig.0000012. arXiv: 2101.08477. 2.1.1

[17] Xuefeng Peng, Yi Ding, David Wihl, Omer Gottesman, Matthieu Komorowski, Li Wei H. Lehman, Andrew Ross, Aldo Faisal, and Finale Doshi-Velez. Improving Sepsis Treatment Strategies by Combining Deep and Kernel-Based Reinforcement Learning. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2018:887–896, 2018. ISSN 1942597X. arXiv: 1901.04670. 2.1.1

[18] Tom J Pollard, Alistair E W Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018. 3.2.1

[19] Venkatesh Sivaraman, Leigh A. Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. volume 1. Association for Computing Machinery, 2023.

doi: 10.1145/3544548.3581075. arXiv: 2302.00096 Publication Title: Conference on Human Factors in Computing Systems - Proceedings Issue: 1. 3.1, 1, 3.4

[20] Shengpu Tang, Maggie Makar, Michael W. Sjoding, Finale Doshi-Velez, and Jenna Wiens. Leveraging Factored Action Spaces for Efficient Offline Reinforcement Learning in Healthcare, May 2023. URL http://arxiv.org/abs/2305.01738. arXiv:2305.01738 [cs]. 2.1.1

[21] Tong Wang and Cynthia Rudin. Causal rule sets for identifying subgroups with enhanced treatment effects. *INFORMS Journal on Computing*, 34(3):1626–1643, 2022. doi: 10.1287/ijoc.2021.1143. 2.1.2

[22] Chao Yu, Guoqi Ren, and Jiming Liu. Deep inverse reinforcement learning for sepsis treatment. *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019*, pages 31–33, 2019. doi: 10.1109/ICHI.2019.8904645. Publisher: IEEE. 2.1.1