How Consistent are Clinicians?

Evaluating the Predictability of Sepsis Disease Progression with Dynamics Models

Grace Park, Venkatesh Sivaraman, Adam Perer

Carnegie Mellon University

Human-Computer Interaction Institute

BACKGROUND

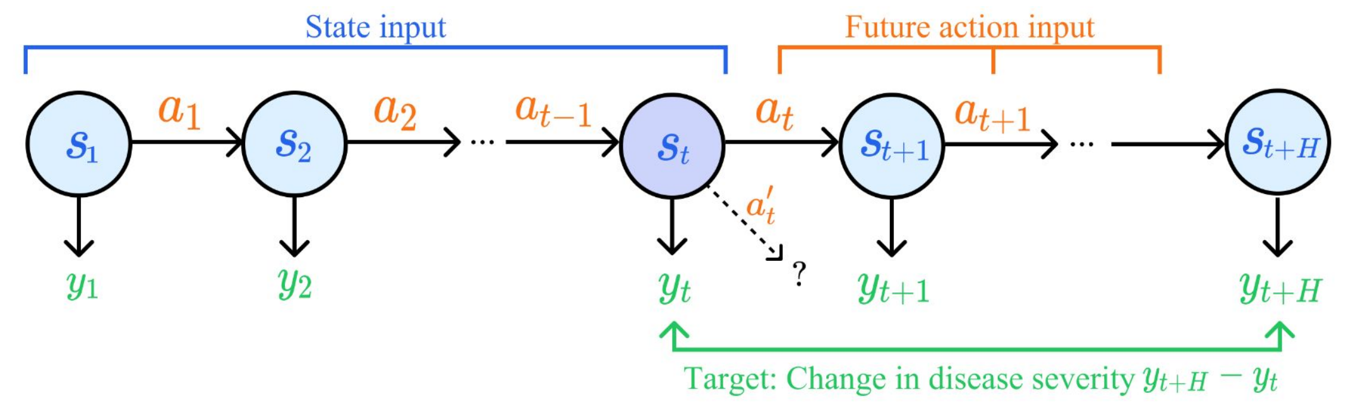
Sepsis is a leading cause of death in hospitals, and there is little clinical consensus around best practices for treatment. Several recent works have applied RL methods to support clinicians' decision-making on sepsis patients in the intensive care unit.

While these algorithms have shown promise when evaluated using off-policy policy evaluation methods, they have also been critiqued for recommending incorrect and even dangerous treatment plans.

It is currently an open question whether it is possible to derive policies from public observational datasets that truly improve current clinical practice.

OBJECTIVE

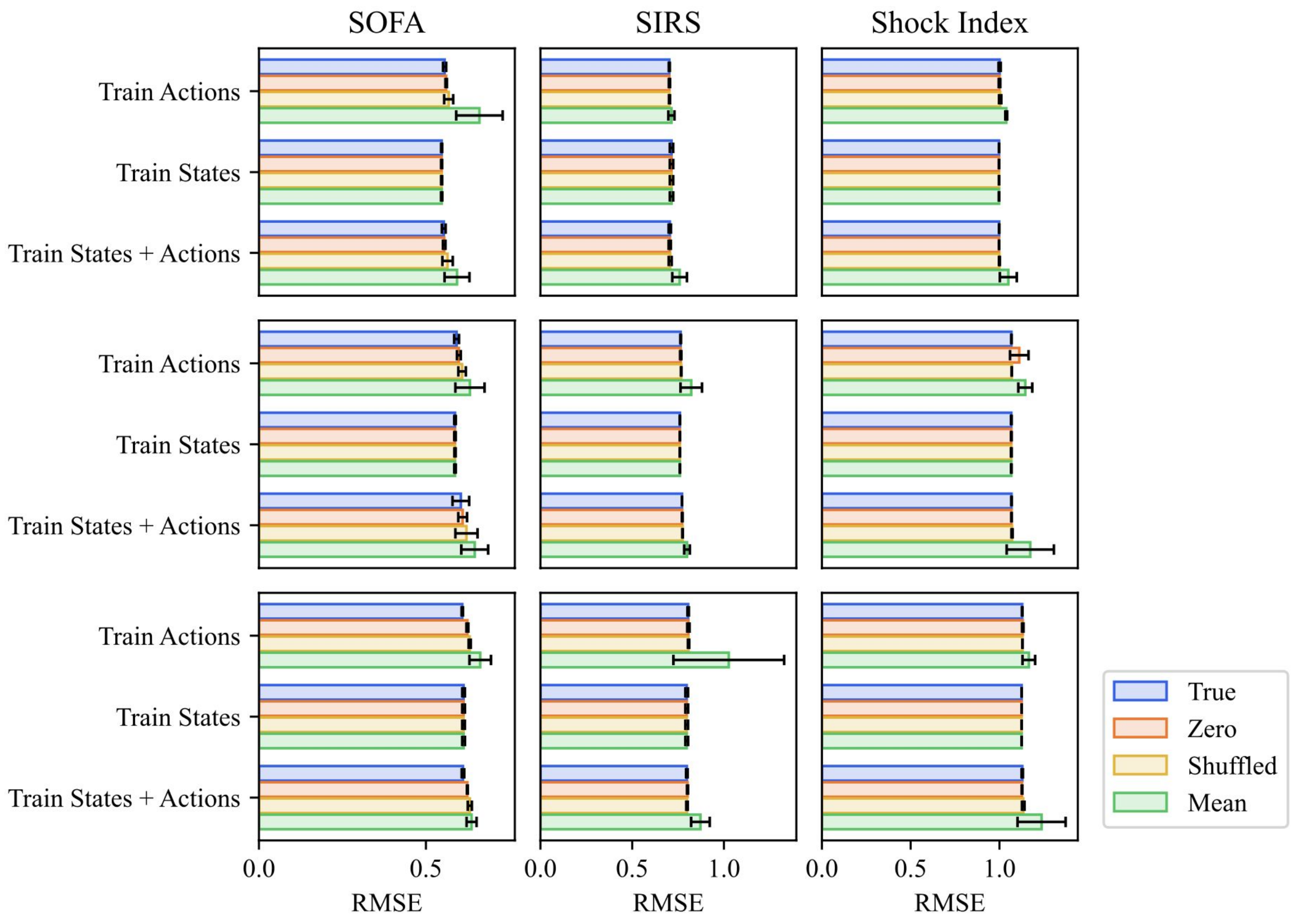
We propose that patient trajectory datasets should exhibit **diversity in observed actions** that correlates with differences in outcomes conditioned on a particular state in order to produce meaningful recommendations with adequate data support.



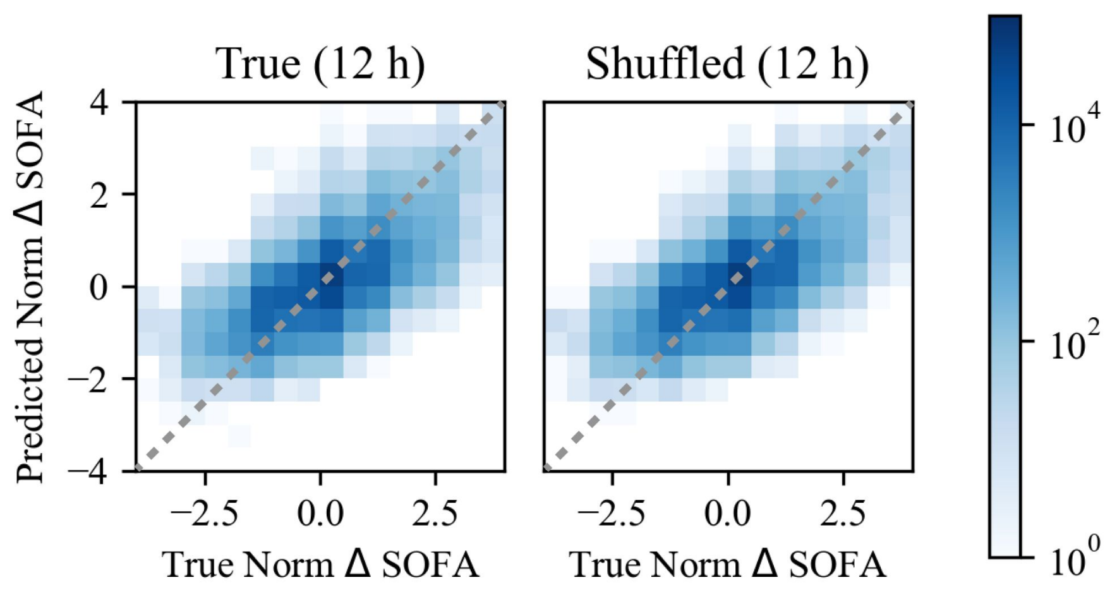
Markov decision process model for patients with sepsis in the ICU. s_t represents the patient state at time t, a_t represents a treatment action, y_t represents a function that captures the disease severity.

INFLUENCE OF ACTION INPUTS ON DISEASE SEVERITY PREDICTIONS

We trained **transformer-based dynamics model** to predict future disease severity given a patient's state and optionally the treatment actions that were taken over the subsequent hours. 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.



RMSE of the predicted change in disease severity across training schemes and action inputs at test time.

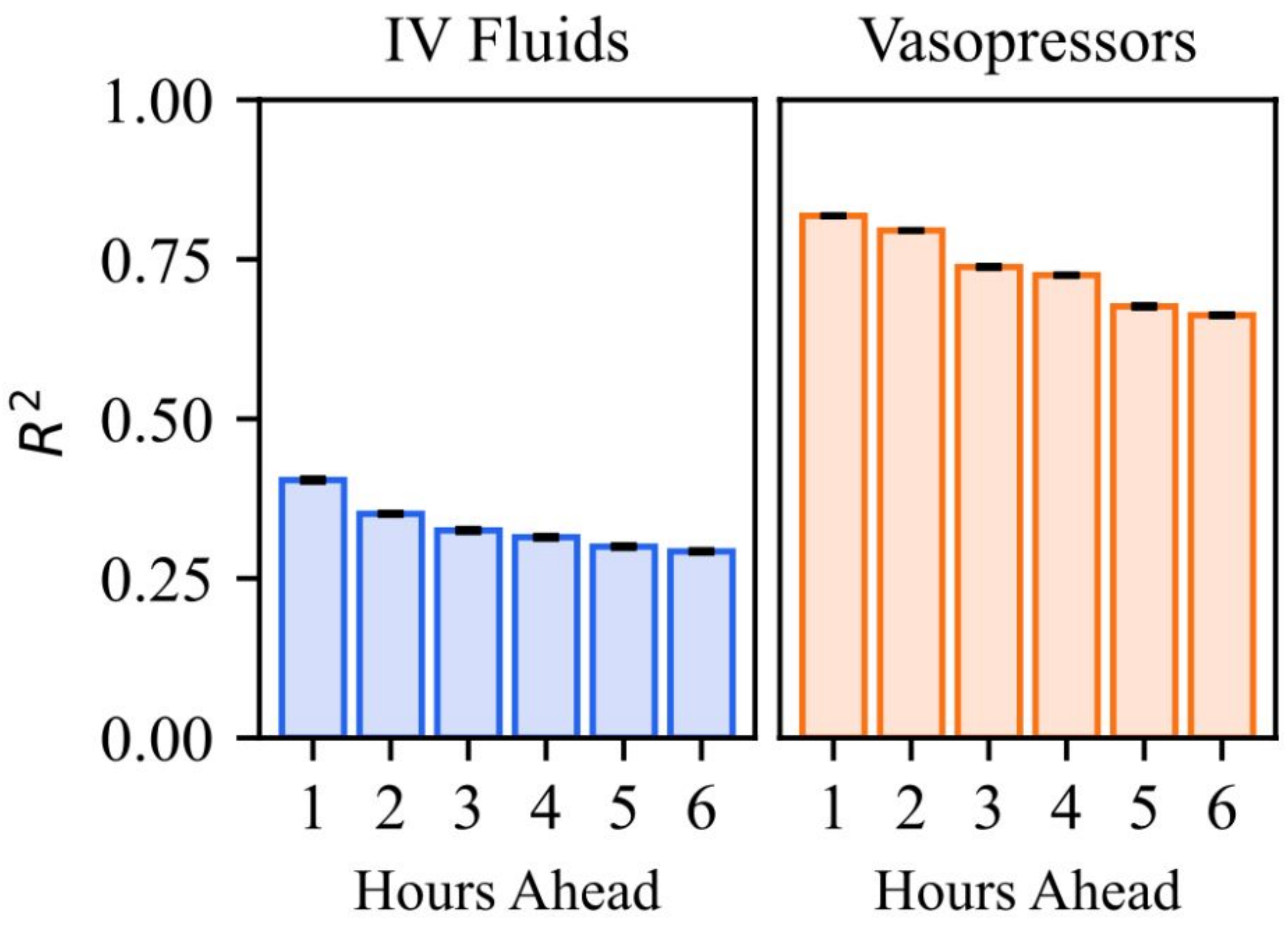


*Example histograms comparing true and predicted changes in SOFA score at 12 hours ahead, in the **True** and **Shuffled** evaluation conditions.*

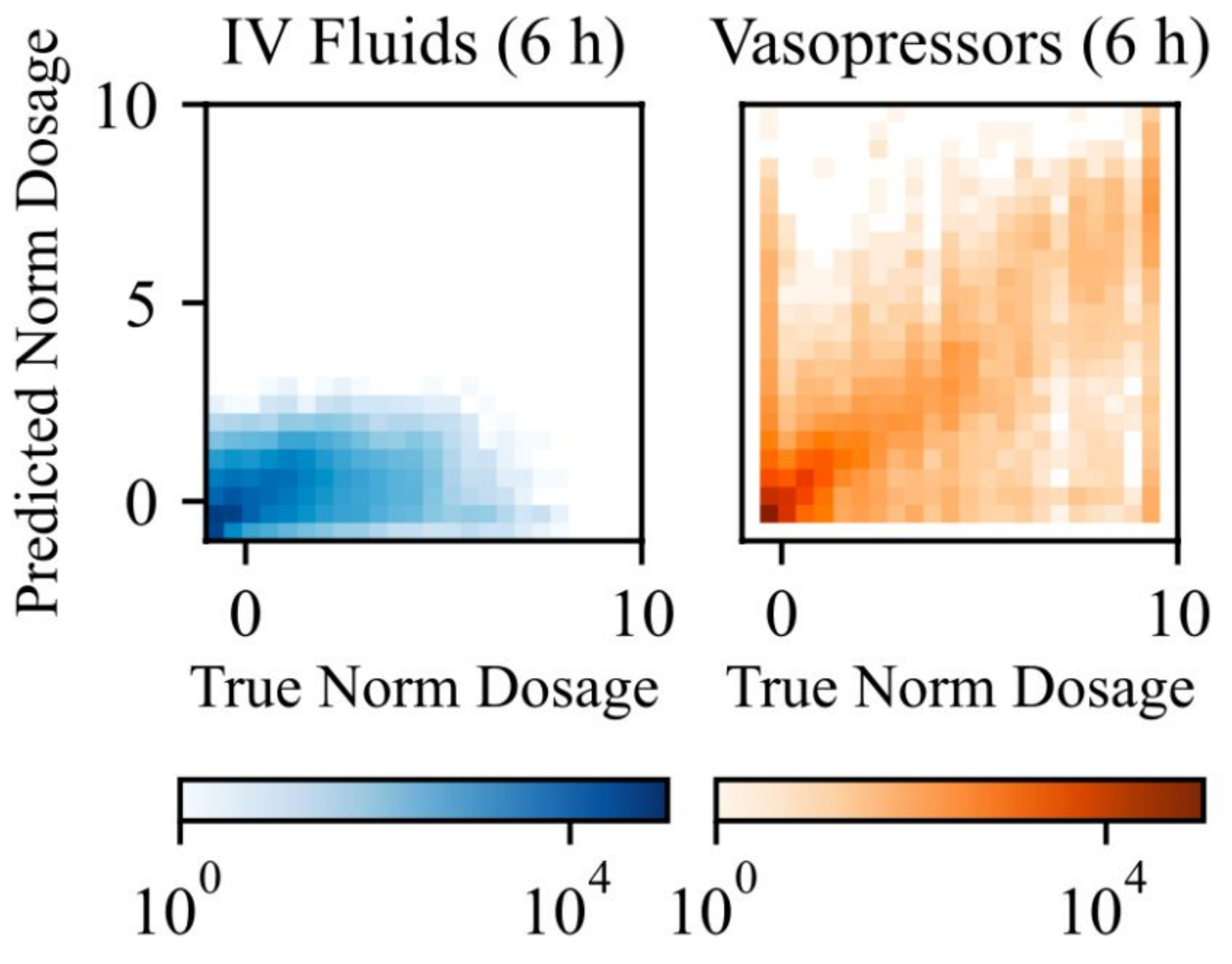
We found that **action information does not confer substantive improvements** in dynamics model fit, as our transformer models could predict future disease severity almost equally well with or without true actions as input. Taken alone, the dynamics model results might suggest that **actions are fully predictable from the states** and there was no need to learn from the action inputs.

PREDICTION OF FUTURE ACTIONS WITH BEHAVIOR CLONING

To evaluate the predictability of actions from states more directly, we trained 3 replicates of the **behavior cloning model** with different random weight initializations. If these models showed a strong fit to the data, one could infer that actions were fully consistent and predictable across clinicians.



Correlations between true and predicted normalized actions from 1 to 6 hours ahead.



Example histograms of correlations between true and predicted normalized actions at 6 hours.

Action prediction was still fairly noisy, indicating that while variation in actions exists, it is not enough 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.

DISCUSSION

Some of the observed lack of diversity in actions on MIMIC data may be due to **inherent challenges in working with patient trajectories**. For instance, there may only be a small number of treatment possibilities that are clinically feasible and safe. Clinicians may also tend to choose actions in predefined patterns, that appear diverse yet lead to consistent outcomes. Alternatively, missing data imputation could have caused patient states and actions to appear more consistent than they really 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 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.

This work highlights the importance of diversity in data sources when building medical recommendation models.