{john-mark.agosta, rhorton, Maryam.Tavakoli}@microsoft.com
http://www.springer.com/gp/computer-science/lncs

Interpreting Dynamic Causal Model Policies

John Mark Agosta[1 [0000−0001−9826−6509]], Robert Horton[1 [0000−0001−7305−354X]], and Maryam Tavakoli Hosseinabadi[1 [0000−0002−8380−5511]]

Microsoft, Redmond USA

Abstract. This project draws together work on learning dynamic causal networks from simulation data with the application of domain knowledge to improve the model. In previous work we showed how causal networks capture domain expertise and can improve simulation modelling. We demonstrate this approach in this paper as it applies to a rudimentary reinforcement learning (RL) solution. Our thesis is that a person can understand the RL solution by means of the model causal structure derived from historical data. This is work-in-progress toward the larger goal of using the combination of domain knowledge applied to causal models to develop improved dynamic treatment policies.

Keywords: Causal Modelling · Synthetic Data · Reinforcement Learning.

1 The combination of causal modelling and domain knowledge

This project draws together work on learning dynamic causal networks from simulation data [3] with the application of domain knowledge to improve the model. In our previous work, [5], we showed how causal networks capture domain expertise and can improve simulation modelling. For instance, it was shown that "confounding by indication" [9] between treatment and severity variables could be resolved by unrolling the causal model into multiple cycles, as shown in Figure 1. In this article we use this unrolled causal model to understand how human intervention may depend on interpretation of dynamic model policies. This is work-in-progress toward the larger goal of using the combination of domain knowledge applied to causal models to develop tools for improved dynamic treatment. Central to our approach is that explanation and causality are two sides of a coin [2].

To test this idea, we posit a given causal model, expressed as a network for which we'd like to discover the optimal dynamic policy. We explore several approaches that take advantage of the interplay between machine learning from data generated by the model with interpretations of the domain that can be read from the causal structure. We demonstrate this approach in this paper as it applies to a rudimentary reinforcement learning (RL) solution. **Our thesis is that a person can best understand the RL solution by means of the model causal structure derived from historical data.**

In actuality, for modelling purposes, one would only have access to data, either on-line or off-line, but not to the underlying causal model. To simulate the real world case, we use the causal model to create a data set from which we use novel tools to learn the causal structure. Because we know the true causal model we have a gold standard by which to evaluate our results.

1.1 The Domain: Simulating Clinical Episodes

We start with a model of treatment of a fictitious virus. In a hospital ICU, we posit a severe infection that requires carefully balancing the dose of a drug that is administered over the course of the infection. Either the patient survives or dies depending on the level of the drug that may vary during the episode. In this work, we explore the effectiveness, expressed as the average survival rate, when the drug dose varies over time with the patient's condition. Our intuition, to be substantiated, is that the optimal dosage should start high and decrease during the episode.

1.2 The dynamic causal network

The domain is modelled by a "Dynamic Bayes Network" [1] (DBN) show in Figure 1. Model variables are shown as nodes and dependencies by arcs to form an a-cyclic directed graph. In standard fashion, conditional probability distributions are shown by ovals, policy functions by squares, and value functions by polygons (e.g., diamonds or hexagons). Conditioning and functional arguments are shown by incident arcs; in the case of policies, these arcs indicate the observations on which the policy depends. Multi-stage models are represented by showing the two stage "unrolling" of the model. In our case the model consists of a finite, random number of stages determined by the number of stages until the final outcome happens.

The DBN is derived from the structure of a causal network learned from data whose causal structure would then be revised by expert judgment. Current network learning methods are insufficient to recover an accurate model, as is apparent by comparing Figure 1 with Figure 2. For this experiment we assume

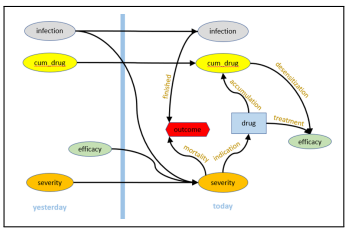


Fig. 1: Original dynamic causal model

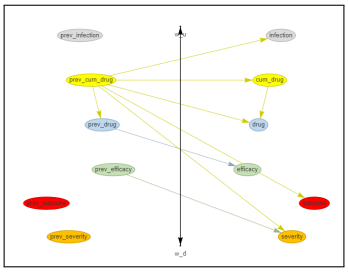


Fig. 2: Causal model learned by *RHINO*

expert judgment would lead to a revised model close to Figure 1. However, as will be seen, the apparent learned model gives insight into the discovered policies.

The modified causal network becomes one stage in the DBN to which a factored model-based RL solution can be applied. Thus by exploiting causal structure we avoid the data inefficiency that plagues RL [4].

Since we are interested in the effects of interventions, a causal understanding of the domain is necessary. We presume all dependencies shown by arcs in the true model are causal. The random variables with time transitions; *infection*, *cumulative drug*, and *severity* are understood to cause their subsequent states. *Infection* tracks the course of the infection. *Severity* is a consequence of two additional antecedents, *infection* and *efficacy*. The outcome is computed from *severity* and *infection*, which is strictly not seen as a causal relation, but just an accounting of the current reward. If the infection out-paces the disease severity, the patient survives. *Cumulative drug* is the (noisy) summation of the current dose regime, and influences *efficacy*. The arc from *severity* to *drug*—the policy node—determines observability; what state variables the policy "sees."

The structure of the unrolled DBN explains why a dynamic policy avoids confounding by indication, or more exactly confounding by severity. As in the typical confounding diagram, severity conditions both *drug* (e.g. the intervention) and the outcome, creating a "back-door"[8] to the direct effect of the intervention on outcome. Once temporal sequence is expressed in the DBN, *severity's* effect is moderated by spreading it over multiple periods. Several interventions isolate its confounding effect from the final outcome: The severity that the interventions see is not the one effecting the current outcome.[1]

1.3 Previous results: Constant dose levels

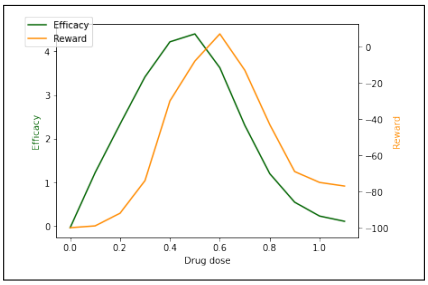
In our previous work, we showed how knowledge of the causal graph could be used to correct learned causal models, especially in extents of the domain where data was sparse. We showed this for simulations made by discrete-valued Bayes' networks, continuous valued "no-tears" style causal models, and by offline RL. These were each simulations derived by learning from simulated data (for which we knew the ground truth)—"simulations of simulations." We did this solely with constant policies; that is, policies that did not depend on time or state. By virtue of domain knowledge as expressed in the causal network, we could understand how to improve the models built from the simulation. In this work we extend this analysis to report results on the first phase of *dynamic* model investigation.

Our previous results are duplicated here in Figure 3. Survival (orange) rates increase with dose levels until efficacy (green) starts to decrease due to desensitization from the cumulative level of the drug.

1.4 Learning a dynamic causal model from data

The purpose of causal discovery is to infer the underlying causal structure from observational data. We used RHINO [3] to infer the instantaneous and one-day

[1] Demonstrating lack of confounding computationally is future work.



Testing dose levels over episodes, survival (shown as average reward per episode) reaches a maximum at a dose level of 0.7. Efficacy rises similarly with Reward, then declines, as dose accumulates over time, its efficacy declines, as the influence by *cumulative dose* on *efficacy* implies.

Fig. 3: Survival, expressed as Reward, compared to Efficacy.

lagged probabilistic structure among variables assuming a Gaussian noise component. RHINO is an end-to-end causal inference framework that learns variables' non-linear, temporal causal relationships from time-series data. It combines vector auto-regression and deep learning to model the effect of an intervention samples within a single stage, as well as causes from previous timesteps—i.e., lagged effects.

The true model network compared to the inferred network are shown in Figure 1 and Figure 2. As impressive as the ability of the learning algorithm is, the diagram illustrates that it needs to be improved by domain knowledge.

2 Policy dynamics

Short of solving for the full dynamic dosing policy we can gain substantial insight into the dosing trade-off by approximating a policy that varies linearly with time. To this end we ran a sequence of experiments with policies parameterized by constant and slope parameters, as functions of different observable variables. Since observables vary with time, and these policies vary linearly with one of the observables, they become dynamic policies. For each observable variable we searched over the range of constant and slope values to maximize average survival per episode, given the stochastic nature of the domain. The magnitude of the optimal value of the slope indicates the time-varying nature of the resulting policy. A negative sign to the slope indicates that the dosage rate declines as the observable increases. The results are shown in Figure 4.

We experimented with policy functions that depend on different observables; *day* into the infection episode, current *severity*, previous *cumulative drug*, and previous *efficacy*. **In contrast the severity-based policy stands out by its lack of any dynamic effect.** This can be explained by the structure of the causal network. The learned causal model, Figure 2, although not truthful to the true causal structure, demonstrates the primacy of the cumulative drug variable's influence on the rest—despite its inaccuracies, the learned model is an accurate representation of the variables' joint probability distribution. One can see from the model that *cumulative drug*, together with *efficacy* is sufficient to determine *severity*, so the temporal dynamics of *severity* are indirect; it only

mediates the effect of *cumulative drug*. As the network shows, *severity* is an indirect indication of the state, and inadequate for a dynamic policy.

run label	policy observable	survival rate	Δ per stage
170-14-50	day	0.963	-0.015
170-19-03	severity	0.626	0.0029
171-22-59	cum drug	0.957	**-0.156**
171-00-39	efficacy	0.968	-0.0142

To decrease *cumulative drug* we force *drug* to decrease with time. This results in an optimal declining dose policy, shown here. We would see a similar effect with *infection*. As for *efficacy*, it first increases then decreases as shown in Figure 3, so the optimal dose as function of efficacy follows a non-monotonic trend.

Fig. 4: Change in the slope of the optimal linear drug dosing policy as a function of different policy observables.

To test the information value of a *severity* based policy, we applied a standard Q-learning algorithm, $Q(s,a) \leftarrow Q(s,a) + \alpha(r + max_{a'}Q(s',a') - Q(s,a))$: See [6]. We run the Q-learner directly against the simulator that implements the true causal model, using *severity* as the state s, to generate s' as a function of (s,a). Our policy and state variables are discretized in 12 levels. As a test of calibration, we showed that with no observation, i.e. with a constant policy, the Q-learner duplicated the optimum dose shown in Figure 3. Thus when *severity* is used as the observed state, the derived policy shows no clear dependency on the observed state as opposed when cumulative drug is used. Since the learned causal model in Figure 2 infers a direct dependence of the policy on cumulative drug, *ironically this is what one would expect, despite its inaccuracy in recovering the true structure.*

Given the brittleness of RL methods, and challenges with convergence one could call this into question. Surely though this negative result is supported by the linear policy experiments, and can be explained based on the causal model structure.

3 Work in Progress: The next steps

Obviously the next step is to complete the RL solutions to determine the informativeness of the set of state variables, to validate these findings. One could argue that recent RL deep learning tools [7] would provide better accuracy, but at the cost of transparency. Our longer term goal is to integrate causal claims into RL methods. The current work in offline RL—where only historical data is available—is a promising target where causal reasoning may be applied. As this work implies, the combination of interactive (e.g. human) input with offline-RL has particular promise. [10]

Acknowledgements

None of this article was written with the aid of a generative language model.
The code for this project is in Github [2].

References

1. Dean, T., Wellman, M.: Planning and Control. Morgan Kaufmann series in representation and reasoning, M. Kaufmann Publishers (1991), https://books.google.com/books?id=cNFSAAAAMAAJ
2. Galavotti, M.C., Suppes, P., Costantini, D.: Stochastic Causality. CSLI Publications, Stanford, CA (2001)
3. Gong, W., Jennings, J., Zhang, C., Pawlowski, N.: Rhino: Deep causal temporal relationship learning with history-dependent noise. arXiv preprint arXiv:2210.14706 (2022)
4. Hester, T., Stone, P.: Learning and using models. In: Reinforcement Learning: State-of-the-Art, pp. 111–141. Springer (2012)
5. Horton, R., Hosseinabadi, M.T., Agosta, J.M.: Approaches to optimizing medical treatment policy using temporal causal model-based simulation (2022), https://openreview.net/forum?id=TptoTbkwaa
6. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. Journal of artificial intelligence research **4**, 237–285 (1996)
7. Liang, E., Liaw, R., Nishihara, R., Moritz, P., Fox, R., Gonzalez, J., Goldberg, K., Stoica, I.: Ray rllib: A composable and scalable reinforcement learning library. arXiv preprint arXiv:1712.09381 **85** (2017)
8. Pearl, J.: Causality. Cambridge university press (2009)
9. Salas, M., Hotman, A., Stricker, B.H.: Confounding by indication: an example of variation in the use of epidemiologic terminology. American journal of epidemiology **149**(11), 981–983 (1999)
10. Swazinna, P., Udluft, S., Runkler, T.: User-interactive offline reinforcement learning (2022). https://doi.org/10.48550/ARXIV.2205.10629, https://arxiv.org/abs/2205.10629

[2] https://github.com/rmhorton/bogovirus