UNIVERSITY OF
CAMBRIDGE
Department of Computer
Science and Technology

Probabilistic hybrid dynamical models for temporal treatment effect estimation

Riccardo Conci

Hughes Hall

June 2024

Submitted in partial fulfillment of the requirements for the
Master of Philosophy in Advanced Computer Science

Total page count: 54

Main chapters (excluding front-matter, references and appendix): 36 pages (pp 10–45)

Declaration

I, Riccardo Conci of Hughes Hall College, being a candidate for the Master of Philosophy in Advanced Computer Science, hereby declare that this project report and the work described in it are my own work, unaided except as may be specified below, and that the project report does not contain material that has already been used to any substantial extent for a comparable purpose. In preparation of this project report I did not use text from AI-assisted platforms generating natural language answers to user queries, including but not limited to ChatGPT. I am content for my project report to be made available to the students and staff of the University.

Signed: Riccardo Conci

Date: 03/06/2024

Abstract

The bedrock of decision-making in modern medicine is high-quality evidence of treatment effect. Randomised Controlled Trials (RCTs) and their meta-analyses are the gold standard for producing this evidence. However, due to expense, limited sample sizes and high patient heterogeneity, RCTs are often not applicable, requiring evidence from observational data. Intensive Care is an exciting domain for causal machine learning (ML) methods due to the high frequency and variety of physiological data. Beyond the challenge of inferring counterfactual trajectories from confounded factual data, causal temporal ML models also need to be trustworthy to a clinical audience. The requirements include interpretability, an integration of data-driven predictions with existing knowledge, robustness in out-of-distribution settings and uncertainty quantification. In this thesis, we present a novel Hybrid Neural Stochastic Differential Equation (Hybrid SDE) model, which explicitly integrates an expert cardiovascular ordinary differential equation model with its neural stochastic counterpart. We evaluate our model against state-of-the-art neural baselines on a challenging synthetic counterfactual trajectory prediction task with visible confounders and across low-overlap settings. We show significantly better results than baselines, while also meeting the requirements stated above.

Acknowledgements

Grazie mille a Prof. Pietro Liò e Francesco Caso, who have shown me unwavering support in converting exciting ideas into actionable results. Through you, this work has managed to flourish. I am grateful to my colleagues and friends from the ACS course, who have always been willing to lend an ear and provide support. Of course, this work would not be possible without the CSD3 cluster, so shout out to Malcolm Scott and the team. Finally, thank you to my partner who has put up with many late nights.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "The healer said to me: *it worked because you thought it would.*
> I let him palm my belly and chest; sometimes he shook and
> closed his eyes. Inches above my skin, he'd sweep his hands like
> smoothing sheets I couldn't see..."
>
> — *Imagine*, by Jennifer Richter

The clinical need

Clinical decisions in modern-day evidence-based medicine are based on pathophysiological understanding, peer-reviewed research, and clinical experience [53]. These combined approaches are required to adequately model the benefits and risks of various treatments on patients. The better we understand and predict treatment effects, the better we can avoid harm and improve outcomes [30], [14].

From a causal inference perspective, double-blind randomised controlled trials (RCTs) and their meta-analyses are the gold standard to infer the average treatment effect of therapies [10]. However, RCTs have their limitations: they are expensive, often limited to patients who fit narrow inclusion criteria and cannot always be performed for ethical reasons [47]. The greatest challenge, however, is that many medications that are tested on specific patients are then used on different subgroups with minimal cross-over evidence [3]. Therefore, there is a strong need to build effective methods that can accurately estimate individualised treatment effects from observational data [16], [60], [52].

An area of medicine which has a high demand for observational causal inference and has potential for solutions is Intensive Care (ICU) [68]. Intensive care looks after the most severely unwell patients, who are typically very complex and heterogeneous. This heterogeneity is a challenge for RCTs as they require large numbers of relatively homogeneous

patients, limiting their effectiveness [24]. Therapeutic decisions in intensive care units (ICU) are, therefore, predominantly based on a first-principles understanding of pathophysiology and its application to individual patients [50], [38], [73]. To facilitate these critical decisions, patients undergo extensive invasive monitoring designed to provide real-time physiological data. Given the abundance of data and the pressing need for better predictions of personalised treatment effects, the ICU represents a key domain for the development and validation of machine learning models [64], [74], [58].

Key challenges

The central hurdle of causal inference from observational data regards confounders [49]. These are variables that impact both the treatment assignment and the outcome. People do not receive treatments at random, but rather because they have specific illnesses. Both the illness and the treatment can affect their outcome. Therefore, to infer an estimated treatment effect, the data must include enough patients who have the same baseline features that impact the outcome and receive a different treatment. This is termed overlap.

In ICU, overlap is minimal due to the greater complexity of patients and treatments. Moreover, treatments are not binary and immediate but rather continuous and dosed. Related to this, patients in ICU live in a continuous time domain, with time-dependent and time-fixed confounders and a physiology that is constantly in flux. This is inherently a much more challenging situation than the standard static setting.

A variety of Temporal Causal Inference methods have been developed for this challenge (see Section 3.1). These include both statistical methods, which require strong assumptions often lacking in clinical data, and causal machine learning methods, which harness the flexibility of neural networks to predict treatment effects.

Current limitations

The primary limitation of these methods is their failure to integrate core requirements of trustworthiness for deployment within a clinical community. These include integration of predictions with existing knowledge, interpretability, robustness to out-of-distribution settings, and uncertainty quantification. These requirements are often met by 'expert' physiological models. However, these expert models have their own limitations, such as a lack of scalability and fragility to variables that are not explicitly modelled.

Contributions

To tackle this gap in current solutions, we create a hybrid neural stochastic differential equation (SDE) (Hybrid SDE) model that attempts to fulfil both the trustworthiness requirements above while maintaining the scalability and flexibility of data-driven neural methods.

Specifically, we integrate an expert cardiovascular system of ordinary differential equations (our expert model) with a neural SDE, such that the outputs of the neural SDE act as a

learnable control that is applied during the dynamics of the expert variables. This coupling enforces the neural outputs to always interplay directly with the expert variables.

Therefore, our contributions can be summarised below:

1. We build a novel hybrid SDE decoder network that integrates expert physiological ODE models with a stochastic neural equivalent that acts as a learnable control.
2. We show that this architecture is successful in predicting counterfactual trajectories with visible confounders across low-overlap settings.
3. We show that this method of hybridisation is helpful in providing clinicians opportunities to inject prior knowledge, such as when a functional dependence of a treatment effect is known, but not the confounded functional form.
4. We evaluate whether this model is robust in out-of-distribution settings.
5. We show that the addition of the stochastic part of our neural model is key for providing uncertainty quantification of our predictions.

Chapter 2

Background

> “ Two roads diverged in a yellow wood,
> And sorry I could not travel both
> And be one traveler, long I stood
> And looked down one as far as I could
> To where it bent in the undergrowth.”
>
> — Robert Frost

2.1 Causal Inference

Causal inference focuses on the identification and estimation of causal effects from data [26]. It is a process of determining whether an association truly reflects a direct cause-effect relationship or not. Let's begin with a simple example. Let's call T_i the treatment intake for unit i.

$$T_i = \begin{cases} 1 & \text{if unit } i \text{ received the treatment} \\ 0 & \text{otherwise} \end{cases}$$

We are interested in determining the effect of this treatment T_i. Let's call Y_i the observed outcome variable for unit i. We want to know if the treatment has any effect on the outcome Y_i. However, the **fundamental challenge of causal inference** is that we can never observe the same unit with and without treatment. Instead we must infer the *potential outcome*, what would have happened to that same unit had they received a different treatment. This is also called the *counterfactual*, where instead the *factual* is the observed data.

To clarify this distinction, we add the subscript: Y_{0i} is the potential outcome for unit i without treatment and Y_{1i} is the potential outcome for *the same* unit i with treatment.

This allows us to define the **individual treatment effect** (ITE) as:

$$Y_{1i} - Y_{0i}$$

Due to the fundamental challenge of causal inference, we never have direct access to the alternate potential outcome. We need to infer this counterfactual in order to estimate the ITE. When this is too hard, however, we can revert to our baseline Average Treatment Effect:

$$ATE = E[Y_1 - Y_0]$$

where E is the expected value across multiple individuals i.

The structural causal model and confounders.

A useful way of clarifying which variables are involved when performing treatment effect estimation is to build a structural causal model (SCM), as shown in Figure 2.1 [48]. The SCM can be described by the following variables:

- $X \to Y$: the baseline change of visible variables without treatment.
- $Z \to Y$: the treatment effect.
- $X \to Z$: an observed confounder of treatment assignment.
- $U_z \to Z$: an unknown confounder of treatment assignment.
- $U_y \to Y$: an unknown variable that impacts treatment effect but not treatment assignment. This is also noise.

where **confounders** are variables that impact both treatment assignment and outcome.

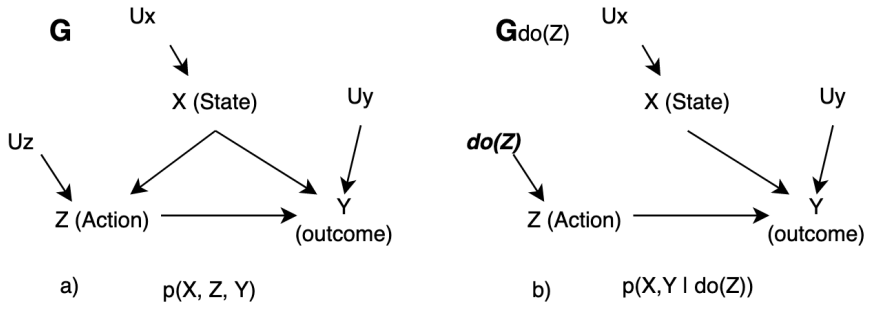


Figure 2.1: Example structural causal model

The aim of a randomised controlled trial is to break the arrow between X and Z. This can be done if Z, the treatment is assigned at random and cannot be predicted by X. This operation is also called the *do()* operation.

However, in observational data, treatments are not given at random and, therefore, are confounded. If we control for these confounders, we can still successfully estimate the treatment effect. However, this requires basic assumptions:

- Consistency: the potential outcome for a treated individual is indeed its factual outcome.
- Positivity: there is a non-zero probability for any individual to receive any treatment.
- No hidden confounders: all the confounders are present in the observed data.

The positivity assumption is often referred to as **overlap**, and can be thought of as the percentage of patients who are 'the same' who receive different treatments. 0% overlap means all the same kinds of patients receive the same treatment.

If these three assumptions are met, in a simple static causal estimation scenario, we can again attempt to estimate the average treatment effect with bias correction

$$E[Y|T=1] - E[Y|T=0] = \underbrace{E[Y_1 - Y_0|T=1]}_{\text{ATT}} + \underbrace{\{E[Y_0|T=1] - E[Y_0|T=0]\}}_{\text{BIAS}}$$

where ATT is the average treatment effect of the treated and the bias term denotes the treatment selection bias to be corrected. For the sake of this thesis, this equation is not required. Instead, we simply require awareness of the key terminology defined above.

2.2 Neural Ordinary, Controlled and Stochastic differential equation models

Neural ordinary differential equations can be written by the general equation:

$$y(0) = y_0 \quad \frac{\mathrm{d}y}{\mathrm{d}t}(t) = f_\theta(t, y(t)). \tag{2.1}$$

where $y(0)$ is the starting value of y at time =0. The function f can be learned and, in the case above, is parametrised by θ, which can represent the weights of an artificial neural network [11]. The central motivation for this is that many input-output problems can be described as a gradual dynamical transformation from a set of initial starting points to a final output.

The benefits of learning this function as an ODE are two-fold. Firstly, recall the original architecture of a residual network:

$$\mathbf{y} = F(\mathbf{x}, \{W_i\}) + \mathbf{x}$$

where $\mathbf{x}$ represents the input to the residual block, $\mathbf{y}$ represents the output, $F(\mathbf{x}, \{W_i\})$ denotes the residual mapping to be learned, with $\{W_i\}$ are the weights of the layers within the residual block. The essential nature of this residual block is the *skip connection*.

By recalling the structure of an ODE just above, we can see that by discretising an ODE via an explicit Euler method uniformly separated by times δt, we can reconstruct the residual network equation. In other words, neural ODEs are the *continuous* limit of residual networks.

The great benefits of this include the flexibility to have as input irregularly sampled data with the added prior that the $t + \delta t$ step will be a small change from the output at time t. This prior moreover fits lots of naturally occurring time-series data better than a recurrent neural network or residual network.

Neural Stochastic Differential Equations

Stochastic differential equations are widely used to model real-world time series with randomness, for example, financial markets [9], particle systems and population dynamics [2]. Integration of a random diffusion provides a natural extension to modelling dynamical systems that evolve in time with uncertainty.

SDEs have a similar drift term to ODEs, but also include a diffusion term.

$$dX_t = \mu X_t \, dt + \sigma X_t \, dW_t \tag{2.2}$$

where X_t represents the stochastic process of interest (e.g., stock price) and μ represents the drift coefficient, which influences the direction and speed of the motion without randomness. σ represents the diffusion coefficient, which determines the volatility of the process due to the random effects modelled by dW_t, a standard Brownian motion.

Being inherently random, allows these SDEs to act as generative time-series models, where we can sample trajectories. By parametrising the drift, and, potentially, diffusion term of the SDE with a neural network, we can generate a powerful model that can fit to any time-series data and then generate further samples from it [37], [67], [72].

A notable highlight to further understand neural ODEs and SDE is Patrick Kidger's PhD thesis [32], which acts as a textbook from first principles to latest research.

2.3 Hybrid models

Hybrid models aim to combine expert models with data-driven flexible machine learning (ML) models, often neural network [66]. This hybridisation can occur in multiple ways.

residual models aim to use an expert model that can issue prediction directly. This allows the ML part to fit the residual gap between the error of the experts and the true outcome

[41], [70]. Instead, *ensemble models* average the predictions of the ML and the expert models.

The expert models can also be used to directly extract latent features that are then given to the machine learning model for prediction [17].

These hybrid models are similar, but distinct from physics-inspired methods which use physical laws to guide the development of novel loss functions or architectures [65], [22]. The more famous examples of this include Hamiltonian neural networks, that aim to reflect a conservation or energy [76].

Within this panacea, our model stands as a *residual* model. As an extension in future work, it can easily integrate with a *feature* model that provides the expert model variables at the time of treatment.

2.4 Offline model-based reinforcement learning

As stated in the introduction above, a central challenge to advancing modern medicine is to successfully infer the causal treatment effects from a dataset of sequential observations, actions and outcomes. The challenge is heightened as data is confounded as treatments often cannot be randomly assigned.

Despite its seeming distance, temporal causal inference and offline reinforcement learning (RL) share exactly the same challenges [36]. Offline RL distinguishes itself from off-policy or on-policy RL in that it aims to learn the optimal policy from access to only a dataset of confounded sequences of actions. Unlike online RL, the agent does not have the possibility to take random actions to learn their effect on the environment, which is highly reminiscent of temporal causal inference for clinical settings.

Imagine the dataset D is given by $\mathcal{D} = \left\{\left(\mathbf{s}_t^i, \mathbf{a}_t^i, \mathbf{s}_{t+1}^i, r_t^i\right)\right\}$, where s is the state, a is the action and r is the reward. We can assume that this dataset is sampled from one (or multiple) policies π_β, such that $\mathbf{a} \sim \pi_\beta(\mathbf{a} \mid \mathbf{s})$. If the reward is not even present in our dataset, there is little we can do. However, if there are multiple trajectories, each imperfect that leads to the reward, then a successful Offline RL model will be able to patch these together into new and better policies.

A *distributional shift* occurs when the policies present in the dataset are still inadequate compared to the real optimal policies. The aim of offline RL algorithms is to specifically tackle this gap. Algorithmic approaches to this can be split into model-free and model-based. The most similar of which to our temporal causal inference setting being model-based.

Model-based offline RL attempts to learn a transition model explicitly from the data, and then use this to select the optimal actions. Essentially, the solved transition model will include the treatment effects of each action for each individual at any point in time,

based on the data available. When the data is not available to infer this, the model would ideally be able to quantify this epistemic uncertainty in its predictions.

However, when imperfect, the transition model can be exploited to train an agent on 'imagined' paths that are not present in the data. In other words, it is extrapolating counterfactuals. When these counterfactuals are significantly different from reality, we term this model exploitation.

A solution to this challenge is to endow the transition model with explicit uncertainties. An array of previous works has done this for online and off-policy settings [71] such as PILCO [20], MBPO [28], PETS-CEM [12]. In the offline setting, MORel [31] has shown an explicit use of uncertainty. Specifically, it trims imagined trajectories that have an uncertainty above a certain threshold, attempting, therefore, to minimise the risk of extrapolating too far.

In this context, our central aim is equivalent to building a probabilistic transition model that can help understand the treatment effects present in the confounded observational data. Specifically, in this thesis we approach this challenge by building a hybrid transition model, which incorporates expert knowledge. On the one hand, the expert model aims to go beyond the observed setting and provide greater insight into each patient by identifying important latent variables. On the other, the probabilistic SDE can help to add flexibility and learnability to this expert model.

2.5 Cardiovascular models, inverse physiology and control

A central aim in intensive care is to maintain a patient's blood pressure above a certain threshold. This threshold is typically defined as Mean Arterial Pressure (MAP) of 65 mmHg. Below this minimal pressure, the flow of oxygen around the body will not match the demand, leading to gradual organ failure and, ultimately, death. Therefore, when the blood pressure starts to decrease there are a few treatment options that attempt to target the central physiological mechanisms that drive the maintenance of this blood pressure. However, each of these options has potential upsides and risks. For example, if the MAP is low because the total intravascular volume is too low, then an infusion of intravenous isotonic saline can help to restore more normal volume and, therefore, pressure. However, the downside is that the MAP may not respond to this fluid infusion and, instead, the fluid goes to the lungs causing a worsening respiratory picture.

Inverse physiology is the challenge of identifying these latent physiological variables from limited observational findings. Clinicians use pattern recognition, guidelines and expert models to help guide which action, including investigations, is the next best to take to decrease the uncertainty around these latent values and, therefore, pick the correct

treatment.

Finally, it is worth noting that both pathology and treatment can be thought of as control variables on specific physiological variables. For example, in severe states of inflammation, a common effect is for the peripheral venous system to massively dilate, dropping the total peripheral resistance and, therefore, blood pressure. In a cardiovascular model, this change in peripheral resistance can be explicitly taken into account as driven by the 'inflammation' control. Similarly, when the correct treatment, a vasopressor agent in this case, is administered, the same latent variable will receive an opposing control driven by the treatment. This view is important to understand the motivation behind the specific model architecture that we set up in our methods.

Chapter 3

Related work

3.1 Machine learning for temporal counterfactual prediction

The methodological advancement from static to temporal causal inference was initially driven by statistical methods, including Inverse Probability Treatment Weighting (IPTW), the G-formula and Marginal Structure Models [55] [63]. These approaches extended the propensity score matching (Section 2.1) to the temporal domain. More specifically weights are assigned to each observation across time conditioned on the previous exposure history. These are then multiplied to generate a single weight for each subject, which acts to creates a pseudo-population which is used to balance the baseline covariates. However, as is the case with propensity models, when overlap is low, few subjects will dominate the weight analysis and produce an overly large variance in the predicted treatment effect estimation. Moreover, this estimation is dependent on the correct specification of the conditional probability of treatment assignment, which is complex to compute in practice.

Recurrent neural networks

In 2018, a neural extension of MSMs was built, called the Recurrent Marginal Structure Networks (RMSN) [39], which aimed to use recurrent neural networks to learn the complex dependencies on covariate history and use the latent representations as a weighted baseline for counterfactual prediction. Specifically, RMSN encompassed three modules: a propensity network, an encoder and a decoder. The propensity network was pre-trained to predict the probability of treatment at each point in time given a patient's covariates. The encoder was trained to predict one step ahead following treatment, and the decoder continued this prediction up to a certain horizon window.

This work continued onto the Time Series Deconfounder [6] and the Counterfactual Recurrent Network (CRN) [7]. These models added to the central ideas of the RMSN an adversarial training in the encoder to replace the propensity network. Specifically, the

CRN encoder takes in the patient history and outputs a final hidden representation. This is then trained to both *not* be able to predict the treatment assignment and to predict the first outcome following treatment. This dual learning encourages the learning of a latent representation given to the decoder that is unbiased and 'balanced'. Finally, the decoder rolls forward in time and predicts the outcomes given the treatment. This work was closely followed by the Disentangled CRN [5] which split the final RNN latent representation into three entities: the 'outcome factor', the 'confounding factor' and the 'treatment factor', providing a window for potential interpretability.

The main limitations across these works are the requirements of positive overlap, the lack of uncertainty quantification and the assumption of regularly sampled time series. These challenges motivated the application of a novel class of methods, neural differential equations, as a way to model counterfactual trajectories.

Neural differential equations

Neural Differential equations (introduced in Section 2.2) have been successfully used as an alternative to recurrent neural networks especially in observational datasets with irregular sampling, [57], and successfully applied to clinical datasets [18], [35].

Specifically, they have been used for treatment effect estimation, most notably with the IMODE model [25] and the CF-ODE model [19], both of which I use as benchmark models for comparison. Given their relevance, they are described in more detail in Section 3.2. The underlying assumption of both IMODE and CF-ODE is the integration of treatments and observations into a latent continuous dynamical system, that is either deterministic in IMODE or stochastic in CF-ODE.

CF-ODE was specifically built to have a low overlap setting in mind aided by a probabilistic approach. Further work, such as TE-CDE [59] returned to a similar format to the Counterfactual Recurrent Network by replacing the recurrent network architecture both in the encoder and decoder with neural controlled differential equations (CDE). CDEs can be thought of as a continuous time version of RNNs, and are therefore useful with modelling irregularly sampled data. Similar to CRN, they use adversarial training to adjust the encoder outputs and 'rebalance' them for counterfactual estimation. They do this again by adding a further training loss on the final encoder representation such that it is unable to predict the treatment assignment and instead able to predict the outcome.

A limitation of the TE-CDE model is the lack of explicit uncertainty quantification. In fact uncertainty in these scenarios can be separated into outcome (epistemic) uncertainty and model (aleatoric) uncertainty. The importance for uncertainty quantification motivated a final addition to these works, the Bayesian Neural Controlled Differential Equations (BNCDE) [27]. BNCDE takes an equivalent approach to TE-CDE, with one major difference: it explicitly models the weights of the neural CDE as a neural SDE. Specifically, the neural SDE is optimised to approximate the posterior distribution of the weights of

the neural CDE given the pre-treatment observations data. The equivalent is done for the decoder. They add a final prediction head that converts the decoder output into a mean and standard deviation outcome prediction.

BNCDE does not explicitly account for any confounders through balancing representations as is done by CRN and TE-CDE. Indeed they discuss in detail how these balancing methods do not in fact reduce the *confounding bias*, but rather the *estimation variance* [61], and may in fact introduce an infinite data bias [42]. Because of this, the authors explicitly decide not to implement the adversarial training for latent rebalancing. They reiterate that estimation bias due to confounding in time series observational datasets is still an unanswered research question.

Hybrid models

All of the methods above, however, fall short of the same challenges: interpretability and integration with existing knowledge. Hybrid models are an explicit way to integrate data-driven methods with expert knowledge (see Section 2.3). A related work that combines neural differential equations with expert models is the Latent Hybridisation Model (LHM) [51]. In this work, the authors combine an expert pharmacological ODE model with a neural ODE. Specifically, they set up an encoder-decoder architecture, with the encoder being a recurrent neural network trained to output a mean and standard deviation for the expert ODE variables at the point of treatment and trained with a variational loss. The decoder is made up of the expert system of ODEs and a **separate** neural ODE, which does not at all interact with the differential equations. Finally, a learnable output function takes in the predictions from the expert model and the dynamics of the neural ODE and combines them to match the factual output.

As we shall see in detail in Section 4.2 our work departs from the LHM in three significant ways because of a variety of challenges below:

1. In LHM, the neural ODE and the expert trajectories are both given to a learnable output function to predict the output. This has a major risk: as these learnable functions scale in parameters, the neural ODE can go from predicting the residual of the expert model to completely disregarding the expert model as it does not require it to perform adequate predictions.

2. Because of this same hybridisation, the outputs from the neural ODE fail to clarify which aspects of the expert dynamics are insufficient for predicting the output trajectory. Instead, it again acts as a black box to predict the output without actually integrating with existing knowledge.

3. LHM does not attempt an explicit uncertainty quantification in the decoder. Instead, they have a variational encoder to help identify the latent expert variables.

The HybridSDE model

Instead, in HybridSDE all these challenges are mitigated. Specifically:

1. There is no learnable output function to combine the neural SDE and the expert model. Instead, the integration occurs *within* the expert model and *during* the dynamic, with neural SDE outputs acting as a *control* to specific latent variables. Because of this, all information has to flow through the expert model and scaling the networks does not lead to ignoring the expert ODE, but rather improved control to better fit the real trajectory.

2. The exact functional form and dependency of the integration between the neural SDE output and the expert variables can be adjusted and chosen to fit prior knowledge. For example, in our experiments below, we pick the control integration as an addition operation, enforcing a level of interpretability on the dynamics.

3. Finally, the output of the SDE network has both a drift and a diffusion term. Similar to CF-ODE, the diffusion will increase in situations with low overlap, providing an uncertainty quantification to the counterfactual trajectories.

3.2 Baseline comparison models

IMODE. Observations $\mathbf{X}$ and interventions $\mathbf{A}$ occur in a sequential, irregularly sampled order. A latent neural dynamical model $\mathbf{h}$ is set up to represent the latent interaction of the underlying system with the actions taken on it. To represent this, the process $\mathbf{h}$ is updated according to the representations z_a and z_x which are learned representations of the observations and actions at each sample time. Specifically, in continuous dynamics, the functions are updated by:

$$\begin{aligned} \dot{\mathbf{h}} &= f_{\psi}^{h}\left(\mathbf{h}, \mathbf{z}_x, \mathbf{z}_a\right) \\ \dot{\mathbf{z}}_x &= f_{\theta}^{x}\left(\mathbf{z}_x\right) \qquad \text{if } t \neq t_k \\ \dot{\mathbf{z}}_a &= f_{\phi}^{a}\left(\mathbf{z}_a\right) \end{aligned}$$

where f_{ψ}^{h}, f_{θ}^{x} and f_{ϕ}^{a} are learnable neural networks. These are learned using a reconstruction loss defined by:

$$\mathcal{L} := \frac{1}{K}\sum_{k=i}^{K} \left\|\mathbf{x}_{t_k} - \hat{\mathbf{x}}\left(t_k\right)\right\|_2^2 = \frac{1}{K}\sum_{k=i}^{K} \left\|\mathbf{x}_{t_k} - \ell_{\omega}\left(\mathbf{h}\left(t_k\right)\right\|_2^2$$

where ℓ_{ω} is another learnable decoder or simply an identity function.

For the sake of the benchmark comparisons in the evaluation (Section 5), the model is adjusted such that z_a is computed as the output of a Gated Recurrent Unit (GRU) recurrent neural network [13], only one treatment is given at a specific time point, and

finally treatment is then inverted explicitly to infer the counterfactual trajectories. The trajectory is, therefore, still informed by the neural ODE **h**.

CF-ODE. Observed data **X** again modelled by a latent continuous-time process $h(t)$ whose dynamics are characterised by an ordinary differential equation. When a treatment is given, this impacts also on the latent process via the function $u_T(t)$, where T indexes the treatment assignment (binary). Finally, the outcome $Y(t)$ is a function of both $u_T(t*)$ and $h(t*)$, where $t*$ represents the time of treatment.

Importantly, De Brouwer models the process $h(t)$ via a neural Stochastic Differential Equation (see Section 2.2), motivated by the need to quantify uncertainty in the counterfactual predictions. Specifically, even though the diffusion term σ is preset, the actual noise variance increases in predictions with minimal overlap, untethering the model from requiring a positivity assumption.

Their loss function is defined as

$$\mathcal{L}(\mathcal{D}, \theta, \phi) = \mathbb{E}_{q_\theta(\mathcal{H}|\mathcal{S}_{t^*})} \left[\log p_\theta(Y \mid \mathcal{H})\right] - KL_{q_\theta(\mathcal{H}|\mathcal{S}_{t^*})} \left(q_\theta \left(\mathcal{H} \mid \mathcal{S}_{t^*}, \phi\right) \| p_0 \left(\mathcal{H} \mid \mathcal{S}_{t^*}, \phi\right)\right) \quad (3.1)$$

with

$$\begin{aligned} q_\theta \left(\mathcal{H} \mid \mathcal{S}_{t^*}\right) \sim \\ dh(t) = f_{\theta_f} \left(h(t), u_{T,\theta_u} \left(t - t^*\right)\right) dt + g_{\phi_q}(h(t)) dW_t \end{aligned} \quad (3.2)$$

and

$$p_0 \left(\mathcal{H} \mid \mathcal{S}_{t^*}\right) \sim dh(t) = f_0(h(t))dt + g_0(h(t))dW_t, \quad (3.3)$$

where S_{t*} is the final hidden representation of the GRU encoder that takes in **X** from $t = 0$ to $t = t*$ at the point of treatment. u_T the treatment function, g_ϕ the output function, and f_θ the ODE function are all multi-layer perceptrons. The prior p_0 is simply the initial parameter weights of these neural networks, making this formulation akin to a Bayesian neural network.

In our benchmarks we denote the CF-ODE model as a 'Neural SDE' to separate it from its non-stochastic version with a diffusion $\sigma = 0$ that we term the Neural ODE. This non-stochastic variant is used by the authors themselves in their own benchmarks. We ran these models directly on the author's code base [1], with the only change being the datasets that we created.

Chapter 4

Methodology

Our central motivating theme is to improve counterfactual prediction of treatments for ICU patients in low overlap temporal settings. This is required to ultimately identify the Individualised Treatment Effect (ITE) of various treatment options (see Section 2.1). The main case study for this thesis, as discussed in the introduction, is the intensive care setting. This is because it has often continuously monitored data from multiple physiological sources, patients are highly complex and heterogeneous and, finally, decisions are often made from a first-principles pathophysiological approach as guidelines evidence is not always applicable.

More concretely, as discussed in Section 2.5 a central aim in intensive care is to maintain a patient's MAP above 65 mmHg. When this starts to deteriorate, clinicians use their expert knowledge to infer latent physiological variables that offer insight into what treatments are likely most effective.

A further challenge arises if the patient does not respond to a treatment as expected by the clinician's internal model. This unexpected treatment effect can, moreover, be confounded: only certain patients who are likely to receive this treatment are likely to respond in an unexpected way. Unless the clinician can identify the reasons for this unexpected treatment effect, this group of patients will always be at risk of receiving inappropriate treatment.

A clinically useful data-driven solution to this challenge, however, has specific needs, as shown by abundant literature on clinical artificial intelligence methods [15], [58], [64], [74]. Among these are **interpretability** and **integration** with existing knowledge, **robustness** to out-of-distribution settings, and **uncertainty quantification**.

These model requirements motivate the creation of our Hybrid Neural Stochastic Differential Equation model (HybridSDE). In the following sections, we formally define this challenge stated above (Section 4.1) and the model formulation and architecture (Section 4.2). Finally, we explain the data creation process to train and evaluate our model and

baselines (Section 4.4).

4.1 Problem formulation

Setup: We consider an observational dataset $\mathcal{D} = \left\{ \left\{ \mathbf{X}^{(i)}_{t<t*}, \mathbf{A}^{(i)}_{t*}, \mathbf{Y}^{(i)}_{t\geq t*} \right\}_{t=0}^{T^i} \right\}_{i=1}^{N}$ with data from N independently sampled patients from time $t = 0$ to T. Each patient (i) is characterised by a d-dimensional time series of *observable* data before treatment assigned at $t = t*$, $\mathbf{X}_{\text{obs}} : [0, t*] \to \mathbb{R}^d$. Each patient i receives a treatment sampled from $\{A\}$. In this thesis, $\{A\}$ only includes a^* and $\neg a^*$, where $\neg a^*$ means no treatment. a^* is a continuous fixed-dose treatment. The outcome trajectory Y_i is impacted both by the baseline X_i of each individual and the specific treatment sampled from $\{A\}$.

Estimation task: Our objective is to predict a probabilistic estimate of two outcome trajectories for each individual i: $Y_i(a^*)$ and for $Y_i(\neg a^*)$. Specifically, if patient m received treatment $\neg a^*$, $Y_m(\neg a^*)$ will be called the *factual* trajectory, and $Y_m(a^*)$ is the *counterfactual* trajectory. The challenge lies in that only the factual trajectory is observed for each patient.

In a standard causal inference setting, following the Robins and Neyman potential outcomes framework [44] [56], this trajectory is identifiable following three assumptions that help us deal with confounders in observational datasets:

1. **Consistency:** For an observed treatment $A = a$, the potential outcome under this treatment trajectory is the same as the factual outcome $Y(a) = Y$.

2. **Overlap:** The probability of receiving treatment $A = a$ for any individual i is never 1 or 0.

3. **No hidden confounders:** The probability of receiving treatment depends on the observed covariates X and, conditional on X, does not further depend on any unmeasured, common causes of treatment and outcome.

Unfortunately, the overlap and no hidden confounders are often unrealistic in real-world datasets [45]. Therefore, in this thesis, we do not assume overlap but rather attempt to quantify the uncertainty in our predictions.

4.2 Model formulation

We begin by viewing the human body as a complex latent dynamical system $\mathbf{H}$ that is mostly invisible to us [62]. Our observations X are derived from this hidden process. Moreover, when a treatment occurs, it impacts this latent system. Thankfully, through centuries of physiological study, we have some expert physiological models $\mathbf{E}$ that can open small windows into this latent process. These expert models are useful, but limited

in their scope. By their very design, they ignore the other latent variables that are not explicitly modelled, making them fragile to hidden dynamics.

We want to make use of these expert models but find a way to connect them through data-driven functional approximations to the unknown process **H**.

We have a total set of working variables defined by X^d_{obs} and X^m_{csv}, where csv stands for our cardiovascular expert model, and m is the total number of the variables in the expert model. The variables in these expert models are split between those that are easily observable m^o and those that are latent m^l, which need to be inferred. In this work, we assume that a separate encoder model, not included in our HybridSDE, does the work of identifying the latent variables m^l. This means that all the variables given to the HybridSDE, which we assume are now *all* observed, have a final dimension $d + m^o + m^l$. The joined variables are now simply described as X unless further specification.

Finally, the dynamics of our hybrid model can be described as:

$$X_{t+\Delta t} = X_t + \int_t^{t+\Delta t} f_{csv}(X_t)\, dt + \int_t^{t+\Delta t} f_{hid}(X_t, \theta)\, dt. \tag{4.1}$$

where f_{csv} is known expert dynamical ODE model, and f_{hid} is instead a learnable function that models the interaction of the expert model with the broader hidden process **H**.

We model the function f_{hid} using a neural stochastic differential equation parametrised by θ. Specifically,

$$X_{t+\Delta t} = \mu(t, X_t, \theta)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}W. \tag{4.2}$$

which is composed of a learnable drift function μ and a diffusion Weiner process $\mathrm{d}W$ the size of which is defined by σ.

Implied in equation 4.2, is an integration between f_{csv} and f_{hid} that can be thought of as a control in the standard dynamical systems literature. As a concrete example, let us imagine the following dynamical system, where the control u is the output of the neural stochastic model.

$$\begin{aligned}
&\text{Initial Conditions:} && \text{Equations:}\\
&x(0) = x_0, && \frac{dx}{dt} = y,\\
&y(0) = y_0, && \frac{dy}{dt} = -k \cdot y + u,\\
&u(0) = 0, && \frac{du}{dt} = \begin{cases} 0, & \text{if } t \le t^*,\\ \mu(t, x, y, \theta) + \sigma dW_t, & \text{if } t > t^*. \end{cases}
\end{aligned} \tag{4.3}$$

As shown in equation 4.3, we can set our SDE network μ to begin once a certain time $t = t^*$ is reached. As we will discuss below, $t = t^*$ at the start of the treatment sampled from $\{A\}$.

We show a simple diagram of the hybridSDE architecture in Figure 4.1. Finally, in our HybridSDE, drift function μ is defined as a multi-layer perception with 6 layers of 400 units each and *tanh* activations between layers, totalling around 900 thousand parameters. These specific values were found by hyperparameter optimisation.

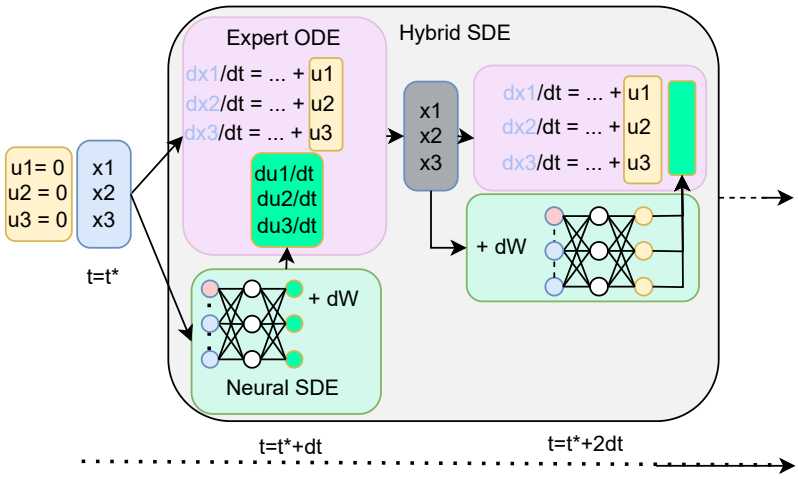


Figure 4.1: HybridSDE unrolled architecture. $x1, x2, x3$, the physiological variables at the time of treatment t^* are passed both to the expert ODE model and the neural SDE. As the dynamical system evolves over time, the network learns to apply a control on the dynamics of the physiological variables. This control can be viewed as the unknown treatment effect, or also a dynamic pathology.

4.3 HybridSDE for counterfactual estimation.

As specified in the problem formulation, we care about predicting both the factual and the counterfactual trajectories for each individual i in our dataset $\mathcal{D}$. To perform this task, we clarify the confounder assumptions that are required to perform causal inference.

- **Assumption 1.** The confounder variable is present in the observed variables X at the time of treatment, making it a visible confounder.
- **Assumption 2.** The probability of assigning a specific treatment from the set $\{A\}$, based on the observed data X, is given by $\Pr(\text{Treatment} = a \mid X_{t^*}) = \tau(X_{t^*})$, where $a \in \{A\}$ is the treatment sampled.
- **Assumption 3.** There are no other hidden confounders.

The loss function, denoted as $\mathcal{L}$, for the model is defined as the negative Gaussian log-likelihood of the predicted *factual* trajectories given the true *factual* trajectories. It is expressed as:

$$\mathcal{L}(\mathcal{D}, \theta) = -\sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \hat{y}_i(\theta))^2}{2\sigma^2}}\right)$$

where y_i represents the true factual trajectory values, $\hat{y}_i$ represents the predicted factual trajectory values, σ is the standard deviation of the prediction errors, assumed to be constant and N is the number of data points.

Software engineering: In this project, best-practice software engineering practice was maintained. Software engineering includes the design, development, testing and maintenance of software applications. These best practices include regular version control and reproducibility of results.

4.4 Data creation

We perform all of our experiments on synthetic data that we create using a cardiovascular system model inspired by Zenker et al. (2007) [75]. We do this because of three main reasons.

1. Access to ground truth counterfactual trajectories allows us to thoroughly evaluate this novel methodological advancement compared to other models. Moreover, control of the data creation allows us to simulate increasingly challenging confounding with decreasing overlap scenarios, allowing us to assess the robustness of our model across various clinical settings.

2. There is a close affinity between the data created by the model and real-world intensive care scenarios. As discussed above, we can simulate the infusion of intravenous fluids in states of low blood pressure and use our model to infer the counterfactual blood pressure trajectories.

3. Finally, the Zenker model has been used across closely related literature [40] [19], with strong benchmarks already set. Using the same synthetic dataset allows us to compare our methods to state-of-the-art solutions.

4.4.1 Cardiovascular model

We distilled the Zenker model into a core set of differential equations defined below:

$$\frac{dP_a(t)}{dt} = \frac{1}{C_a \times 100}\left(\frac{P_a(t) - P_v(t)}{R_{TPR}(S)} - SV \cdot f_{HR}(S)\right) \tag{4.4}$$

$$\frac{dP_v(t)}{dt} = \frac{1}{C_v \times 10}\left(-C_a \frac{dP_a(t)}{dt} + I_{\text{control}}(t)\right) \tag{4.5}$$

$$\frac{dS(t)}{dt} = \frac{1}{\tau_{\text{Baro}}}\left(1 - \frac{1}{1 + e^{-k_{\text{width}}(P_a(t) - P_{\text{a set}})}} - S\right) \tag{4.6}$$

$$\frac{dSV(t)}{dt} = 0.001 \times I_{\text{control}}(t) \tag{4.7}$$

where

$$R_{TPR}(S) = S(t)(R_{TPR_{\text{Max}}} - R_{TPR_{\text{Min}}}) + R_{TPR_{\text{Min}}} + R_{TPR_{\text{control}}} \tag{4.8}$$

$$f_{HR}(S) = S(t)(f_{HR_{\text{Max}}} - f_{HR_{\text{Min}}}) + f_{HR_{\text{Min}}} \tag{4.9}$$

In Table 4.1, we show the variables that are dynamically updated each time step as defined by the equations 4.4, 4.5, 4.6 and 4.7.

Symbol	**Description**	**Typical Values**
P_a	Pressure in arterial compartment	90-130 mm Hg
P_v	Pressure in venous compartment	40-90 mm Hg
S	Baroreflex sensitivity: sympathetic activation	(dimensionless)
SV	Stroke volume: volume of blood ejected during one cardiac cycle	50-100 ml

Table 4.1: Dynamic variables of the cardiovascular ODE model with typical 'healthy' values.

These dynamics are, however, ultimately set by the initial values of the operational variables, detailed in Table 4.2. These are static and depend in part on the health status of each individual. The values have been set by population studies and laboratory experiments and taken directly from the Zenker model.

Finally, we describe the dynamic control variables in Table 4.3, which are present to model the control of both pathology and treatments. For example, by setting I_{control} to a negative value, we can simulate the gradual depletion of total intravascular volume (dehydration) that is common in states of illness. By instead setting I_{control} to be positive, we can simulate the increase of total intravascular volume, for example, by the administration of intravenous fluids.

4.4.2 Simulating physiological confounders

Finally, observational clinical data is highly confounded: there exist variables that affect both treatment **effect** and treatment **assignment** (Section 2.1). The central challenge of causal inference and of our model will be to manage these confounders.

Symbol	Description	Typical Values
$f_{HR_{\max}}$, $f_{HR_{\min}}$	Heart rate	3-0.5 Hz (180-50 bpm)
$R_{TPR_{\max}}$, $R_{TPR_{\min}}$	Total systemic vascular resistance	2.134- 0.5335 mm Hg s/ml
C_a	Compliance of arterial compartment	4 ml/mm Hg
C_v	Compliance of venous compartment	111 ml/mm Hg
k_{width}	Constant determining the shape and maximal slope of the logistic baroreflex nonlinearity	0.1838 mm Hg^{-1}
$P_{\text{a set}}$	Set point of the baroreflex feedback loop	70 mm Hg
τ_{Baro}	Time constant of the baroreflex response	20 s

Table 4.2: Static variables of the ODE model with typical 'healthy' values.

Symbol	Description	Typical Values
I_{control}	Intravenous fluid infusion or depletion	0 ml/s
$R_{TPR_{\text{control}}}$	Modulation to R_{TPR} from pathology or treatment	0 mm Hg s/ml

Table 4.3: Control variables in simple cardiovascular model

Simulating confounded treatment effect: we set up a static confounder based on the initial value of the stroke volume (SV). It is important to note that this initial stroke volume does not change until the treatment is started. We uniformly sample the stroke volume to be between 83 and 90 ml. Then we perform a linear transform, as per 4.10, and then a non-linearity (4.12) to produce the $Conf_{multiplier}$, a value that is multiplied with the fluid infused $I_{control}$ at every time point to give a confounded fluid infusion rate.

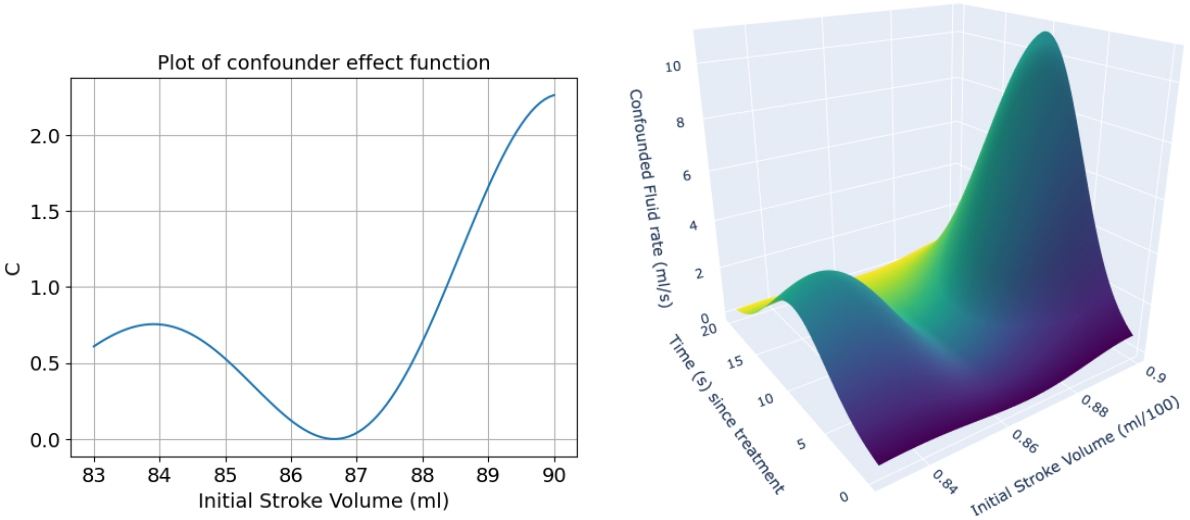
$$sv_{adj} = 0.5 + \frac{0.01 sv_{t*} - 0.19}{0.1} \tag{4.10}$$

$$C = 0.05 \left(\cos(5 * sv_{adj} - 0.2) \times (5 - sv_{adj})^2\right)^2 \tag{4.11}$$

$$I_{Conf} = I_{control} * C \tag{4.12}$$

The specific values in these equations were picked to create the confounding treatment effect function shown in Figure 4.2a. This function is a non-linear multiplier to the fluid rate, such that patients with an initial low stroke volume have a disproportionately large response to the infused fluids compared to those with a higher initial stroke volume, with certain patients having no response at all to the infused fluid. The impact of this

confounded treatment effect can be seen in Figures 4.2a and 4.2b. Figure **??** shows the non-confounded response to a standard infused fluid volume of 50ml over 20 seconds across a sample of 100 sampled patients



(a) X-axis: initial stroke volume. Y-axis: treatment effect multiplier C.

(b) Confounded fluid rate over time for each initial stroke volume.

Figure 4.2: Confounding the treatment effect based on the pre-treatment stroke volume.

Simulating confounded treatment assignment: a treatment effect multiplier is not a confounder by itself unless it also impacts treatment assignment. Without altering the treatment assignment, essentially we return to a setting of a randomised controlled trial, where the treatment effect is not known, but half of the randomly matched participants receive treatment A and the other half treatment B. Instead, by affecting the assignment, we essentially enforce that only the participants with specific characteristics (age, sex, etc.) receive treatment A and the others receive treatment B.

The positivity assumption set up in Section 4.1 only holds valid if this separation of treatment assignment is not total: at least one person with those specific characteristics receives treatment B instead of A. In observational clinical data, due to the heterogeneity of patients in complicated settings and clinical guidelines in more simple settings, overlap is typically minimal.

To flexibly simulate data with varying types of overlap, we make the treatment assignment explicitly depend on the stroke volume at the time of treatment $t*$, with the probability of treatment p is determined by a pre-specified variable γ, as per the equations below.

$$\text{norm\_sv}_{t^*} = \frac{\text{SV}_{t^*} - \text{SV}_{\text{min}}}{\text{SV}_{\text{max}} - \text{SV}_{\text{min}}} \tag{4.13}$$

$$p = 1 - \sigma(\gamma \cdot \text{norm\_sv}_{t^*}) \tag{4.14}$$

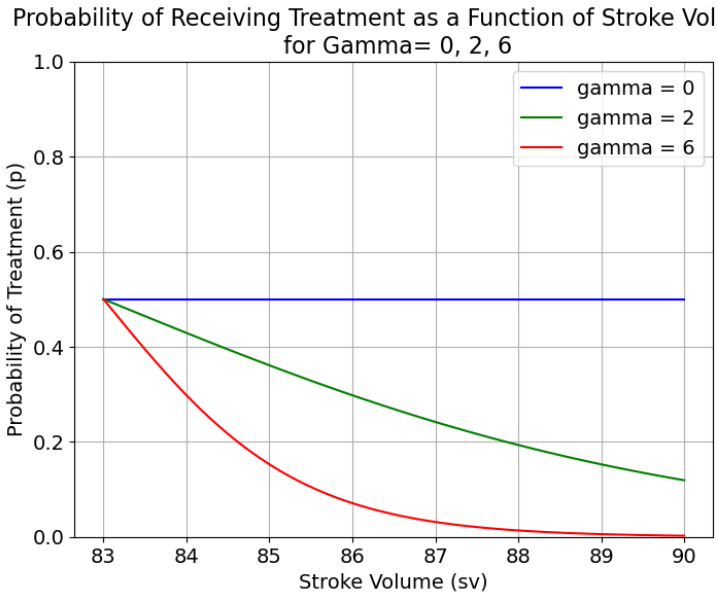


Figure 4.3: The confounded probability receiving treatment decreases with increasing gamma. The counterfactual, therefore, of receiving treatment becomes much harder to predict.

A challenging counterfactual. We set up the confounded treatment effect and treatment assignment to create a challenging setup to predict counterfactual with low overlap. In previous applications of this dataset (including in benchmark papers such as CF-ODE), the probability of treatment increased instead of decreased with increasing γ, leading to effectively all inputs receiving treatment. The counterfactual, therefore, becomes about predicting the lack of treatment. While this is especially important in real-world clinical scenarios, in this synthetic dataset, where prediction is occurring over seconds, the pre-treatment observations are relatively stable if no treatment occurs. Therefore, predicting the counterfactual of 'no treatment' becomes a much easier task.

4.4.3 Simulating out-of-distribution data

An important evaluation for models used in high-risk scenarios is their robustness to out-of-distribution data. To simulate this, we create two datasets, one to train and test in-distribution and the other to test only out-of-distribution. Specifically, we set $R_{TPR_{\text{control}}}$ to -0.5 for the training data. We can see some examples of what this data looks like in Figure 4.4, in subplot 4.4a and 4.4b, both for the non-confounded and the confounded cases. Then for the out-of-distribution testing, we create the with $R_{TPR_{\text{control}}}$ set to +0.2, as shown in subplots 4.4c and 4.4d. Throughout our experiments, we use confounded data.

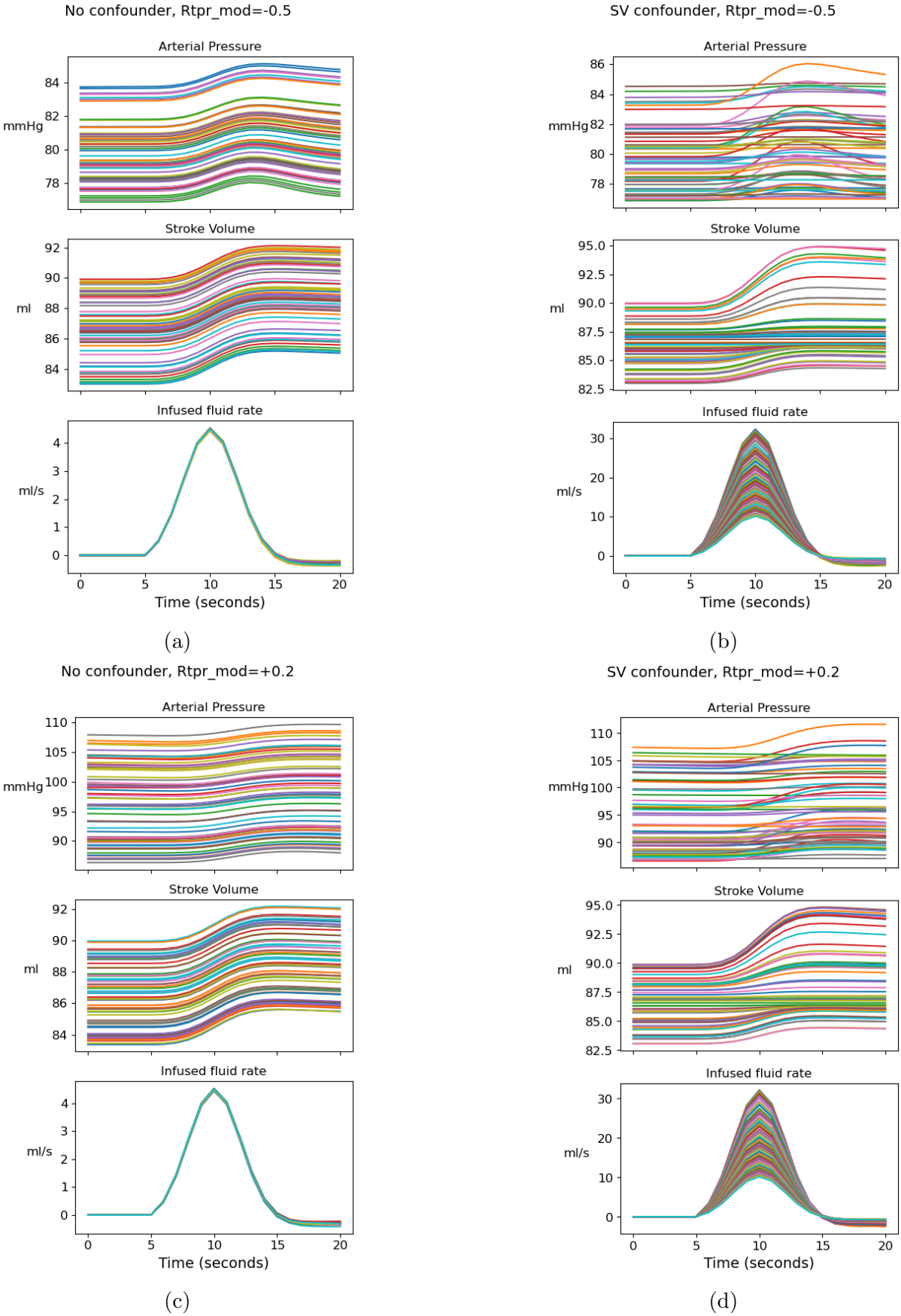


Figure 4.4: Comparison of arterial pressure, stroke volume, and fluid rate with and without confounding treatment effects. Subplot (b) shows the training data, and (d) the out-of-distribution test data.

Chapter 5

Experimental design, results and evaluation

> When I heard the learn'd astronomer,
> When the proofs, the figures, were ranged in columns before me,
> When I was shown the charts and diagrams, to add, divide, and measure them,...
>
> — Walt Whitman

There are ample situations where the treatment effect that a clinician expects based on their expert knowledge does not, in fact, match with the real outcome. In these situations, there is a gap in the physician's model, that ideally can be filled by data-driven prediction methods. Artificial neural networks offer scalability and flexibility with enough data to fill this gap and predict consistently well in within-distribution settings. However, when using observational clinical data, various data-driven risks emerge. Firstly, due to confounding effects and low overlap, overlap is potentially very low, and secondly, the cost of a poor-quality prediction is high. These risks are heightened when the data-driven predictions are not easily interpretable, and when they do not integrate with existing knowledge.

These combined data and model challenges motivate the need for a hybrid method. In this thesis, our hybrid model is composed of an 'expert' ODE physiological model that interacts during its dynamics with the output of a neural SDE model. The specifics of the interaction can be altered based on prior knowledge, and the prediction gap needs to be filled. Let us imagine various scenarios where this hybrid structure may help.

Experiment motivations

1. The clinician might know that by giving a specific treatment, only a select group of physiological variables is ultimately affected: the *functional dependence* of the treatment is known. However, the *functional form*, exactly *how* the treatment affects these variables over time, is not known. In this scenario, selectively placing

the neural network output to act as a control variable within the physiological ODE model allows the neural SDE to learn a time-varying control signal that ultimately is equivalent to the unknown treatment effect. This allows clinicians to combine their prior knowledge with the flexible and more interpretable function approximator. This scenario motivates **Experiment 1** (5.1)

2. ICU patients are often incredibly complex and varied due to the interplay of multiple physiological systems that each can go wrong and impact the others. While some pathologies are much more common than others, the combinatorial explosion of the potential pathologies leads to a long tail of rarer but real pathological scenarios. A neural network that is only trained on the commonly seen data, will likely overfit these at the expense of the rarer diseases. Instead, clinicians are able to deal with these rarer situations because of the internal causal models that provide robustness across out-of-distribution predictions. **Experiment 2** (5.2) explore whether our hybrid model can also show robustness in these settings.

Baselines and ablation studies. In the results below, we train and test 5 different models. These include:

1. **HybridSDE**: our model as described in the model architecture and in the experimental design sections.
2. **HybridODE**: an ablated version of our model, with the σ of our diffusion process set to 0.
3. **NeuralSDE**: this is the same model as the CF-ODE stated in the related works (Section 3.1). We change the name for clarity. In its decoder form, this model is equivalent to the HybridSDE but without the expert model.
4. **NeuralODE**: is the CF-ODE model but with their diffusion process set to 0.
5. **IMODE**: In the CF-ODE paper, the IMODE model was significantly better on their cardiovascular data. Therefore, it is an important benchmark to include.

All these models were run parameter matched at around 900 thousand parameters for a maximum of 250 epochs. Each model was trained and tested on three different random seeds, each of which also impacted the random splits in the data creation, therefore acting as cross-validation splits.

Evaluation metrics. For each of these models we will show two main evaluation metrics. First, the root mean squared error (RMSE) between the predicted counterfactual trajectory and the real counterfactual trajectory. Second, the Precision in Estimation of Heterogeneous Effect (PEHE) [34], which is defined by the RMSE between the predicted Individualised Treatment Effect (ITE) and the real ITE. For completeness, the ITE is computed as the difference between the factual potential outcome and the counterfactual potential outcome (see Section 2.1).

5.1 Experiment 1: Known functional dependence, unknown functional form.

Experimental setup

Confounder: visible
Input: Full physiological information at the point of treatment.
Prediction: Continuous trajectory of arterial blood pressure (Pa)
Treatment type: Continuous, binary and equal across all training and testing data
Treatment functional dependence: known
Treatment functional form: unknown
Test set: Within training distribution

In this experiment we evaluate whether the HybridSDEN can indeed learn a correct confounded treatment effect with decreasing levels of overlap. Moreover, we showcase the flexibility in our model for an expert to set a prior functional dependence between the latent variables and the treatment. For example, a clinician may know that a fluid administration will impact the venous volume and not total peripheral resistance. Instead, they do not know exactly how this treatment will impact. Specifically, we set up the following control signal within the expert ODE system:

$$\frac{dP_v(t)}{dt} = \frac{1}{C_v}\left(-C_a\frac{dP_a(t)}{dt} + \mathrm{u}_1(t)\right) \tag{5.1}$$

$$\frac{dSV(t)}{dt} = \mathrm{u}_2(t) \tag{5.2}$$

$$\frac{d\mathrm{u}_1}{dt} = \mathrm{SDEN}_{\mathrm{out}_1} \tag{5.3}$$

$$\frac{d\mathrm{u}_2}{dt} = \mathrm{SDEN}_{\mathrm{out}_2} \tag{5.4}$$

Where u_1 and u_2 are the control signals that depict the treatment and are 0 until $t*$ when treatment starts. Therefore, the stochastic network needs to learn the derivative of these control signals.

Results

Table 5.1: Evaluation in-distribution (10^{-3})

Model	Gamma	IID CF RMSE	IID PEHE
HybridSDE	0	$\mathbf{0.44 \pm 0.05}$	$\mathbf{0.84 \pm 0.01}$
HybridODE	0	0.6 ± 0.5	1.0 ± 0.7
NeuralSDE	0	22.8 ± 0.9	7.7 ± 1.7
NeuralODE	0	23.3 ± 0.4	8.8 ± 0.7
IMODE	0	3.7	5.2
HybridSDE	2	$\mathbf{0.6 \pm 0.3}$	$\mathbf{0.9 \pm 0.4}$
HybridODE	2	1.0 ± 0.2	1.3 ± 0.2
NeuralSDE	2	24.0 ± 1.0	7.4 ± 1.8
NeuralODE	2	24.0 ± 0.6	8.8 ± 0.7
IMODE	2	4.5	5.2
HybridSDE	6	$\mathbf{3.8 \pm 1.0}$	$\mathbf{3.9 \pm 1.0}$
HybridODE	6	$\mathbf{3.4 \pm 1.1}$	$\mathbf{3.6 \pm 1.2}$
NeuralSDE	6	21.0 ± 2.4	8.7 ± 0.8
NeuralODE	6	23.5 ± 0.2	8.8 ± 0.6
IMODE	6	5.8 ± 0.2	5.5 ± 0.2

Discussion

To interpret the results in Table 5.1, it is important to remember that γ defines the level of overlap. The higher γ is, the less overlap there is, with a functional form plotted in Figure 4.3. Therefore the first rows with $\gamma = 0$ are purely a *baseline* result as they assume no treatment assignment confounding. Therefore, we must both review the absolute numbers for each level of overlap, but also the trend as γ increases.

With that in mind, we first note that our HybridSDE network has a significantly lower counterfactual RMSE and PEHE as a baseline compared to other models, with the HybridODE a close second and IMODE third. As the overlap decreases, the HybridSDE still leads. However, its error significantly worsens compared to its baseline. Instead, the error in the NeuralSDE and NeuralODE has minimal change with a worsening overlap. This likely shows that it was not learning adequately even in the baseline case.

These results, therefore, show that our HybridSDE model can successfully learn the control signals u_1 and u_2 to match a confounded treatment effect even in relatively low overlap scenarios such as in $\gamma = 2$. An example of these predictions can be seen in Figure 5.2.

Finally, an important aspect of building the HybridSDE model is to find ways to integrate data-driven insights with existing knowledge. To test this, we can see whether the control signals integrated over time shown in Figure 5.1 can be reasonably interpreted and whether

they indeed match with our known confounded controls. Unfortunately, we can see that it does not truly match it: the wiggle behaviour is present and visible in multiple examples of the predicted trajectories in Figure 5.2. It is worth noting that these images, are static snapshots taken every few epochs of training, and while this pattern does repeat, there is rapid variability that occurs. Further in-depth evaluation would be required to understand the root cause of this behaviour: whether it is data, loss, or implementation.

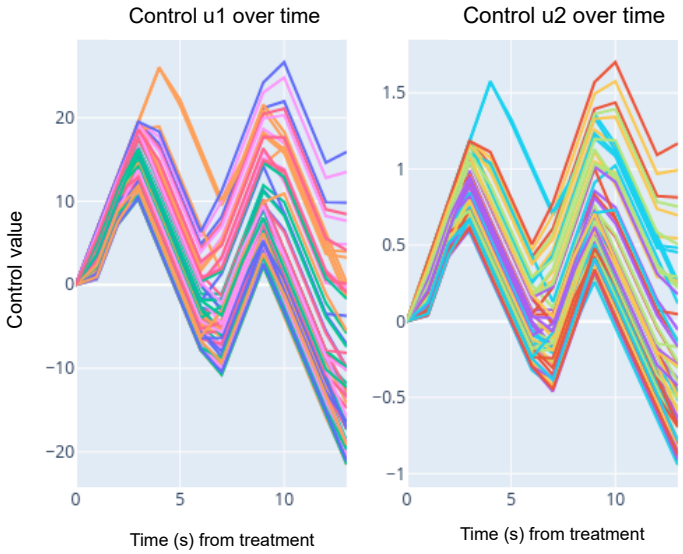


Figure 5.1: Control u_1 and u_2 learned by the SDE in $\gamma = 2$. Note that the SDE network is learning the *differential* of this control, not the control itself. Each colour depicts a different individual trajectory with the output samples averaged into one.

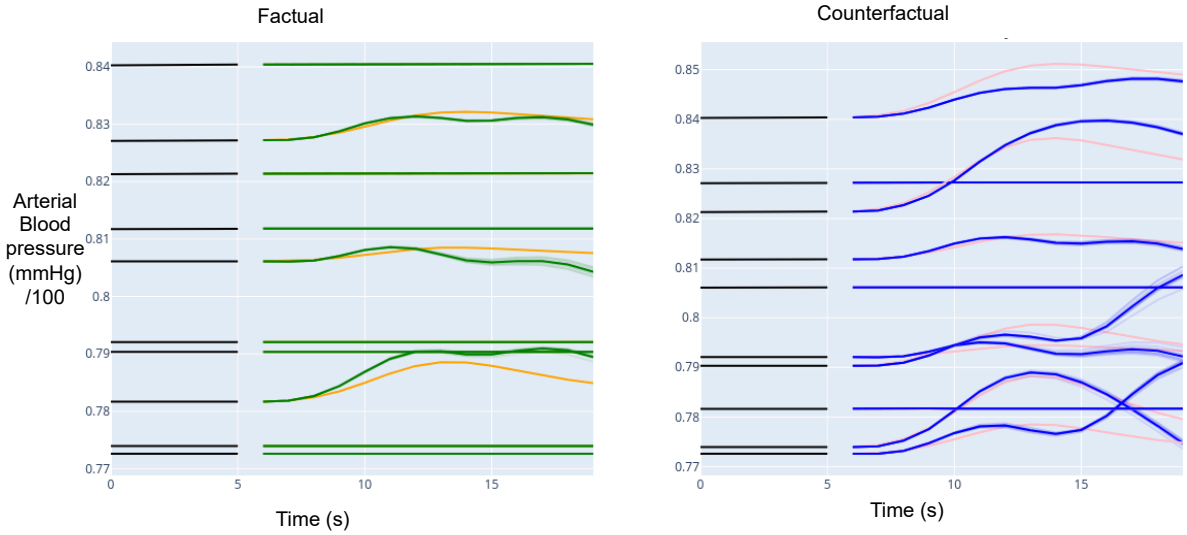


Figure 5.2: Example of predicted vs real P_a trajectories with $\gamma = 2$. In black are the factual pre-treatment trajectories. On the left plot, dark green are the *predicted factual* trajectories, which are trying to approximate the *real factual* trajectories in orange. On the left plot, the dark blue *predicted counterfactual* trajectories are trying to match the pink *real counterfactual* trajectories. Each predicted line has both a mean and 10 samples which can be seen to diverge in situations of greater predictive uncertainty, i.e. in the incorrect counterfactual predictions.

5.2 Experiment 2: Robustness in out-of-distribution settings

Experimental setup

Confounder: visible
Input: Full physiological information at the point of treatment.
Prediction: Continuous trajectory of arterial blood pressure (Pa)
Treatment type: Continuous, binary and equal across all training and testing data
Treatment functional dependence: known
Treatment functional form: unknown
Test set: Out-of-distribution

In this experiment, we maintain the same training setup as in Experiment 1, but for testing we provide the models with a previously *unseen* combination of physiological variables. Specifically, while in the training data we set $R_{TPR_{\text{control}}}$ to -0.5, in the test set we provide is as +0.2, as discussed in Section 4.4.3. This change means that these new patients are intrinsically more responsive to the fluid infusion and maintain their blood pressure higher for longer following this treatment.

Results

Table 5.2: Evaluation Out-of-distribution (10^{-3})

Model	Gamma	OOD CF RMSE	OOD PEHE
HybridSDE	0	$\mathbf{7.1 \pm 0.9}$	13 ± 0.5
HybridODE	0	7.9 ± 1.0	15.7 ± 1.7
NeuralSDE	0	18.6 ± 0.07	14.4 ± 1.3
NeuralODE	0	18.6 ± 0.04	15.3 ± 0.3
IMODE	0	162 ± 1.5	$\mathbf{11 \pm 0.4}$
HybridSDE	2	16.2 ± 5	21.6 ± 7.8
HybridODE	2	$\mathbf{10.8 \pm 3.6}$	13.3 ± 4.3
NeuralSDE	2	18.9 ± 0.12	13.9 ± 1.5
NeuralODE	2	19.0 ± 0.1	15.4 ± 0.4
IMODE	2	161 ± 1.0	$\mathbf{11.5 \pm 0.3}$
HybridSDE	6	$\mathbf{11.4 \pm 0.1}$	$\mathbf{12.8 \pm 0.3}$
HybridODE	6	12.3 ± 2.4	$\mathbf{12.7 \pm 2.6}$
NeuralSDE	6	19.0 ± 0.25	15.5 ± 0.2
NeuralODE	6	19.2 ± 0.1	15.3 ± 0.3
IMODE	6	164 ± 2	$\mathbf{12.0 \pm 0.5}$

Discussion:

The results in Table 5.2 can be interpreted in a similar way to Experiment 1, but the added comparison to Table 5.1 to see the prediction gap from within to out-of-distribution. Highlighted are the best results for each γ and column. Our hypothesis before running this experiment was that our model, which includes $R_{TPR_{\text{control}}}$ in the expert ODE system, would be less impacted by a change in its value than purely neural equivalents.

Looking at the baseline $\gamma = 0$, the HybridSDE and HybridODE have a 10-fold increase in the Counterfactual RMSE and PEHE from within-distribution to out-of-distribution. This is, however, dwarfed by the 100-fold increase present in IMODE for the Counterfactual RMSE. Interestingly, the PEHE for IMODE does increase but only 2-fold. A way to interpret this is that both the factual and the counterfactual trajectories are majorly distorted in the OOD prediction in IMODE, but relative to each other, they maintain a similar gap. Hence, the low PEHE.

The NeuralSDE and NeuralODE models improve marginally in the out-of-distribution setting. However, this improvement is from a high-loss baseline, and both in the IID and OOD settings, this level of loss suggests an inability to predict counterfactual trajectories successfully. In conclusion, we can safely say that the HybridSDE and HybridODE models, while suffering in out-of-distribution settings, are still able to predict relatively high-quality counterfactual trajectories.

5.3 Uncertainty Quantification

Finally, a central requirement when building models for high-risk environments is quantifying uncertainty. This need is a central motivator for using a neural SDE instead of a neural ODE in our model. The stochastic differential equation is split between a drift and a diffusion term, and we specifically use a neural network to learn the drift. The diffusion, instead, is not learned but fixed. While seemingly restrictive, being a stochastic *process*, it can and does increase and decrease based on the strength of the drift [8], [67]. In our case, this can be leveraged to quantify the uncertainty around the counterfactual predictions. An example of this can be seen in Figure 5.2, where in the counterfactual predictions, we can see the many samples suddenly diverging from their mean in an area where it is far from the true counterfactual trajectory.

To quantify this, we can rank each factual and counterfactual trajectory by the standard deviation of its samples and gradually remove the trajectories with the highest uncertainty. Figure 5.3 shows exactly this process with regard to the factual and counterfactual RMSE. This plot confirms that by using a stochastic model, we can decide to reject all the trajectories that reach beyond a certain uncertainty threshold or at least be aware of their decreased trustworthiness.

As a final note on the HybridSDE versus the HybridODE, the results in Table 5.1 and 5.2 show very close and similar values, with the stochastic version being slightly better than the ODE. This consistency further reinforces that the addition of a stochastic component does *not* in fact hurt performance. Instead, it maintains it while adding the added benefit of uncertainty quantification.

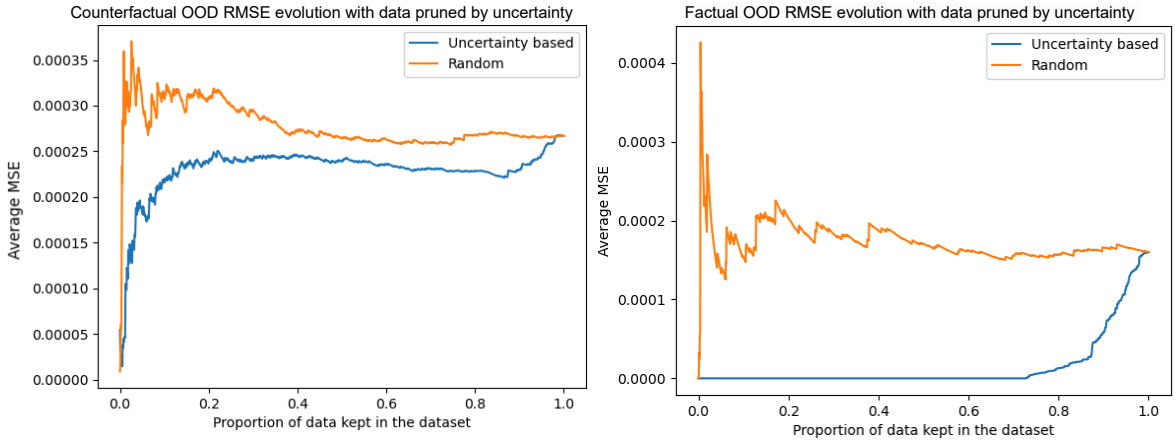


Figure 5.3: Evolution of the factual and counterfactual OOD RMSE as a function of data trimmed by the variance of the predicted samples.

Chapter 6

Discussion and conclusions

> "All models are wrong, but some save more lives."
>
> — Pietro Liò

6.1 Summary

Individualised treatment effect estimation from confounded observational temporal data is a non-trivial challenge to solve. It is also of paramount importance for the success of personalised medicine [15]. In this thesis, we were motivated by the current uncertainty over fluid administration in intensive care [4] to build a hybrid cardiovascular-neural differential equation model that can predict counterfactual blood pressure trajectories in confounded data. We argue that this hybrid model can offer the strengths of both the expert and the neural methods. However, it also encompasses the weaknesses of both methods.

6.2 Limitations and next steps

Encoder: In this work, we limit ourselves to building a hybrid decoder architecture, which we assume takes as inputs correct values for the expert ODE model. While seeming undesirable, this separation may be beneficial, as ultimately, the expert model trajectory only requires the correct starting values to always give the same dynamics, independent of previous treatments. Therefore, a dedicated expert encoder that is trained separately may be the best solution. Moreover, the variables that *do* affect treatment response can be encoded in a further neural encoder that communicates directly with the SDE network. We can show a simple configuration of this setup in Figure 6.1.

Model selection: For any expert model to be useful, it needs to be correctly selected both for the variables it does and does not model and for the time frame of its simulation. Typically, a trade-off is required between complexity and accurate representation. However, a more concerning risk of applying physiological models to the clinical outcomes

is that often the outcome of interest is months or years following an intervention. It is common for clinical trials to have all-cause mortality as one of the end-points [23], as it is important to know whether a specific therapy has an impact on mortality and morbidity benefit, rather than simply improving the specific biomarkers or symptoms of interest [21]. Converting this aim to an observational dataset with a resolution of minutes to hours is unlikely to be successful. However, potential solutions already exist in the literature. These include multi-fidelity probabilistic modelling that aims to combine a multitude of expert and data-driven simulation and surrogate models under one umbrella [54]. In this way, the model selection challenge is overcome by instead taking in multiple models and carefully engineering when each should be used and how they share information.

Unknown functional dependence: In the HybridSDE model architecture, we set the output of the neural network to learn the equivalent of control in a dynamical system of equations. If a clinician knows exactly the functional dependence of such a control, it is easy to apply the output of the network in a precise manner. However, the functional dependence of a control on the expert system may not be known. This occurs during a complex pathological process which impacts an array of physiological variables. In this case, we can adjust the network to apply multiple control outputs to every variable in the system, potentially adding a regularising effect such as an L1 norm to enforce sparsity for greater interpretability.

Realistic settings: In the current framework, the HybridSDE is applied to a specifically selected short time series, each from independently sampled synthetic patients, which all receive the same treatment. While being a useful starting point, the real aim would be to apply the model to an actual clinical time series. In this setting, each patient would receive different treatments at different times, with irregularly sampled data and with time-dependent confounding. This is the ultimate aim of these methods.

6.3 Broader perspectives and actionability in clinics

While state-of-the-art results on relatively simple datasets exist (Section 3.1), the challenge goes beyond prediction in synthetic datasets. The reality of the challenge is much greater. These are my thoughts based on personal experience as having worked as a clinician.

Multimodality for causal inference.

In order to answer any question a minimum amount of information is always required. However, this information may not always appear in the same format. Modern machine learning methods trained on modern large-scale observational clinical electronic health record (EHR) data will always be lacking, not because of the models but because of the lack of necessary information within the EHR data, driven all by confounding. For example, until the late 1990s, before the first biologics such as Infliximab and Rituximab were

approved for use in rheumatoid arthritis, severe disease was not so uncommon [43] (7.1 in the Appendix). However, as treatments and early detection have improved, that level of severity is rarely reached. In fact, its absence from the dataset is itself informative of a well-developed healthcare system. In order to know of such an absence, however, a model has to build this awareness from diverse data sources, such as medical textbooks and research papers. This is important because estimation of treatment effect requires knowledge of the 'baseline' without treatment. And the more successful clinical care becomes, the less this baseline will be directly present in electronic health records. Therefore, even for the offline task of estimating Individualised Treatment Effects from observational data, a multi-modal approach that can fill the gap that is created by highly confounded data is essential [29].

Interpretability and integration with the corpus of knowledge.

In this thesis, we attempt a specific type of multi-modality: one that combines 'expert' and neural differential equations. Despite its limitations, it is motivated by a second, broader theme: interoperability with the ongoing corpus of knowledge. Artificial Intelligence (AI) for scientific discovery [69], is about using AI methods to help with the scientific method of stating assumptions, generating hypotheses, experimenting, observing results and drawing further hypotheses. Central to this ever-turning wheel is a common set of symbols to encode this knowledge and uncertainty. Whether these be natural language, knowledge graphs or mathematics, human-AI scientists will require a shared symbolic form [33] [46].

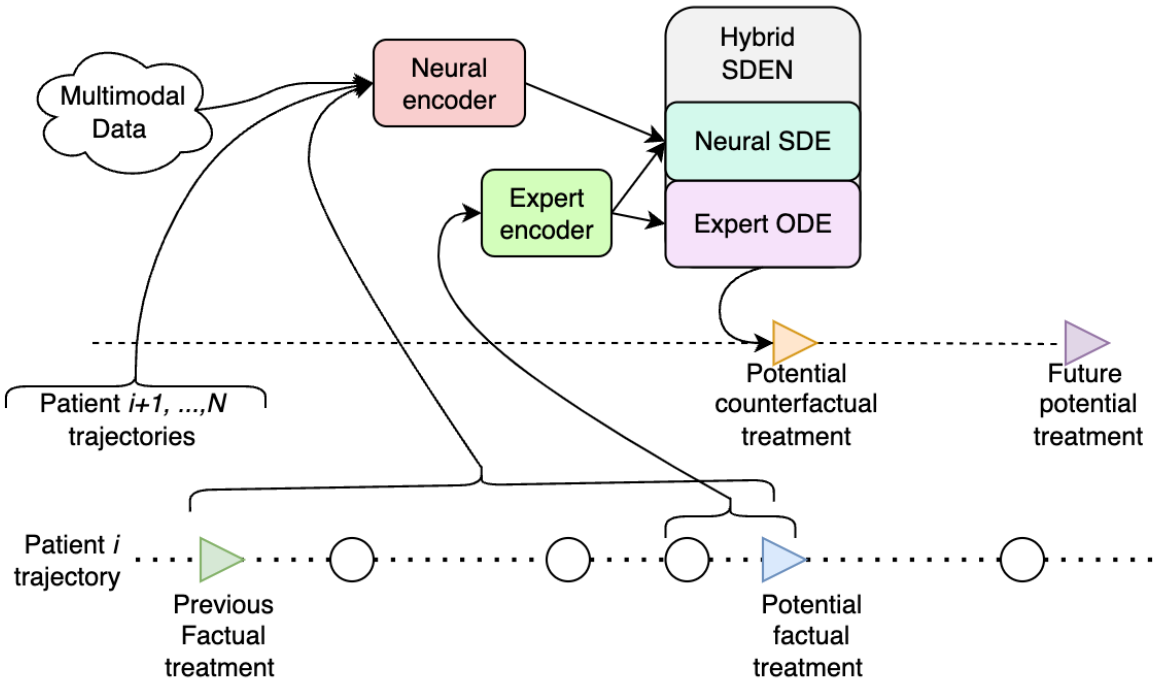


Figure 6.1: Inference using the extended Hybrid SDE model, encompassed by additional expert and neural encoders. The expert encoder provides the expert variables at the time of treatment to the ODE model and the SDE network. The neural encoder provides further inputs to the SDE network. The neural encoder can scale to data much beyond the original time series.

Bibliography

[1] cf-ode: Codebase accompanying the AISTATS 2022 paper : "predicting the impact of treatments over time with uncertainty aware neural differential equations".

[2] A famous nonlinear stochastic equation (lotka-volterra model with diffusion). *Mathematical and Computer Modelling*, 38(7-9):709–726, October 2003.

[3] Benjamin Ackerman, Ian Schmid, Kara E Rudolph, Marissa J Seamans, Ryoko Susukida, Ramin Mojtabai, and Elizabeth A Stuart. Implementing statistical methods for generalizing randomized trial findings to a target population. *Addict. Behav.*, 94:124–132, July 2019.

[4] Ashley Barlow, Brooke Barlow, Nancy Tang, Bhavik M Shah, and Amber E King. Intravenous fluid management in critically ill adults: A review. *Crit. Care Nurse*, 40(6):e17–e27, December 2020.

[5] Jeroen Berrevoets, Alicia Curth, Ioana Bica, Eoin McKinney, and Mihaela van der Schaar. Disentangled counterfactual recurrent networks for treatment effect inference over time. *arXiv [cs.LG]*, December 2021.

[6] Ioana Bica, A Alaa, and M Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. *ICML*, abs/1902.00450:884–895, February 2019.

[7] Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *Eighth International Conference on Learning Representations*, April 2020.

[8] Michelle Boué and Paul Dupuis. A variational representation for certain functionals of brownian motion. *Ann. Probab.*, 26(4):1641–1659, October 1998.

[9] Damiano Brigo and Fabio Mercurio. *Interest rate models — theory and practice*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[10] Patricia B Burns, Rod J Rohrich, and Kevin C Chung. The levels of evidence and their role in evidence-based medicine. *Plast. Reconstr. Surg.*, 128(1):305–310, July 2011.

[11] Ricky T Q Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *arXiv [cs.LG]*, June 2018.

[12] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. May 2018.

[13] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv [cs.NE]*, December 2014.

[14] Jean-Paul Collet, William McKellin, Sravan Jaggumantri, and Niranjan Kissoon. Personalized evidence of treatment effects for the practice of personalized medicine: a new model of care. *BMC Health Services Research*, 14(2):1–1, July 2014.

[15] Alicia Curth, Richard W Peck, Eoin McKinney, James Weatherall, and Mihaela van der Schaar. Using machine learning to individualize treatment effect estimation: Challenges and opportunities. *Clin. Pharmacol. Ther.*, 115(4):710–719, April 2024.

[16] Amit Dang. Real-world evidence: A primer. *Pharmaceut. Med.*, 37(1):25–36, January 2023.

[17] Arka Daw, Anuj Karpatne, William Watkins, Jordan Read, and Vipin Kumar. Physics-guided neural networks (PGNN): An application in lake temperature modeling. October 2017.

[18] Edward De Brouwer, Thijs Becker, Yves Moreau, Eva Kubala Havrdova, Maria Trojano, Sara Eichau, Serkan Ozakbas, Marco Onofrj, Pierre Grammond, Jens Kuhle, Ludwig Kappos, Patrizia Sola, Elisabetta Cartechini, Jeannette Lechner-Scott, Raed Alroughani, Oliver Gerlach, Tomas Kalincik, Franco Granella, Francois Grand'Maison, Roberto Bergamaschi, Maria José Sá, Bart Van Wijmeersch, Aysun Soysal, Jose Luis Sanchez-Menoyo, Claudio Solaro, Cavit Boz, Gerardo Iuliano, Katherine Buzzard, Eduardo Aguera-Morales, Murat Terzi, Tamara Castillo Trivio, Daniele Spitaleri, Vincent Van Pesch, Vahid Shaygannejad, Fraser Moore, Celia Oreja-Guevara, Davide Maimone, Riadh Gouider, Tunde Csepany, Cristina Ramo-Tello, and Liesbet Peeters. Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression. *Comput. Methods Programs Biomed.*, 208(106180):106180, September 2021.

[19] Edward De Brouwer, Javier Gonzalez, and Stephanie Hyland. Predicting the impact of treatments over time with uncertainty aware neural differential equations. In Gustau Camps-Valls, Francisco J R Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4705–4722. PMLR, 2022.

[20] M Deisenroth and C Rasmussen. PILCO: A model-based and data-efficient approach to policy search. *ICML*, pages 465–472, June 2011.

[21] G Filippini, F Brusaferri, W A Sibley, A Citterio, G Ciucci, R Midgard, and L Candelise. Corticosteroids or ACTH for acute exacerbations in multiple sclerosis. *Cochrane Database Syst. Rev.*, (4):CD001331, 2000.

[22] Ferdinando Fioretto, Terrence W K Mak, and Pascal Van Hentenryck. Predicting AC optimal power flows: Combining deep learning and lagrangian dual methods. *arXiv [eess.SP]*, September 2019.

[23] Jan O Friedrich, Michael O Harhay, Derek C Angus, Karen E A Burns, Deborah J Cook, Dean A Fergusson, Simon Finfer, Paul Hébert, Kathy Rowan, Gordon Rubenfeld, John C Marshall, and International Forum for Acute Care Trialists (InFACT). Mortality as a measure of treatment effect in clinical trials recruiting critically ill patients. *Crit. Care Med.*, 51(2):222–230, February 2023.

[24] Anders Granholm, Waleed Alhazzani, Lennie P G Derde, Derek C Angus, Fernando G Zampieri, Naomi E Hammond, Rob Mac Sweeney, Sheila N Myatra, Elie Azoulay, Kathryn Rowan, Paul J Young, Anders Perner, and Morten Hylander Møller. Randomised clinical trials in critical care: past, present and future. *Intensive Care Med.*, 48(2):164–178, February 2022.

[25] Daehoon Gwak, Gyuhyeon Sim, Michael Poli, Stefano Massaroli, Jaegul Choo, and Edward Choi. Neural ordinary differential equations for intervention modeling. *arXiv [cs.LG]*, October 2020.

[26] Miguel A Hernan and James M Robins. *Causal inference: What if*. CRC Press, Boca Raton, FL, April 2024.

[27] Konstantin Hess, Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bayesian neural controlled differential equations for treatment effect estimation. *arXiv [cs.LG]*, October 2023.

[28] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-Based policy optimization. June 2019.

[29] Myong Chol Jung, He Zhao, Joanna Dipnall, and Lan Du. Beyond unimodal: Generalising neural processes for multimodal uncertainty estimation. In *Thirty-seventh Conference on Neural Information Processing Systems*, November 2023.

[30] David M Kent, Ewout Steyerberg, and David van Klaveren. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ*, 363:k4245, December 2018.

[31] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL : Model-Based offline reinforcement learning. May 2020.

[32] Patrick Kidger. On neural differential equations. *arXiv [cs.LG]*, February 2022.

[33] Ross D King, Jem Rowland, Stephen G Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N Soldatova, Andrew Sparkes, Kenneth E Whelan, and Amanda Clare. The automation of science. *Science*, 324(5923):85–89, April 2009.

[34] Niki Kiriakidou and Christos Diou. An evaluation framework for comparing causal inference models. *arXiv [stat.ML]*, August 2022.

[35] Jonghyeon Lee, Edward De Brouwer, Boumediene Hamzi, and Houman Owhadi. Learning dynamical systems from data: A simple cross-validation perspective, part III: Irregularly-sampled time series. *arXiv [stat.ML]*, November 2021.

[36] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. May 2020.

[37] Xuechen Li, Ting-Kam Leonard Wong, Ricky T Q Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. *arXiv [cs.LG]*, January 2020.

[38] Geoffrey K Lighthall and Cristina Vazquez-Guillamet. Understanding decision making in critical care. *Clin. Med. Res.*, 13(3-4):156–168, December 2015.

[39] Bryan Lim. Forecasting treatment responses over time using recurrent marginal structural networks. *Neural Inf Process Syst*, pages 7494–7504, 2018.

[40] Ori Linial, Neta Ravid, Danny Eytan, and Uri Shalit. Generative ODE modeling with known unknowns. *arXiv [stat.ML]*, March 2020.

[41] Dehao Liu and Yan Wang. Multi-fidelity physics-constrained neural network and its application in materials modeling. *J. Mech. Des. N. Y.*, 141(12):1, December 2019.

[42] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. *arXiv [stat.ML]*, November 2023.

[43] Emeline Minichiello, Luca Semerano, and Marie-Christophe Boissier. Time trends in the incidence, prevalence, and severity of rheumatoid arthritis: A systematic literature review. *Joint Bone Spine*, 83(6):625–630, December 2016.

[44] Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, pages 1–51, 1923.

[45] Michael Oberst, Fredrik D Johansson, Dennis Wei, Tian Gao, Gabriel Brat, David Sontag, and Kush R Varshney. Characterization of overlap in observational studies. *arXiv [cs.LG]*, July 2019.

[46] Jeff Z Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini,

Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. Large language models and knowledge graphs: Opportunities and challenges. *arXiv [cs.AI]*, August 2023.

[47] Warren Pearce, Sujatha Raman, and Andrew Turner. Randomised trials in context: practical problems and social aspects of evidence-based medicine and policy. *Trials*, 16(1):394, September 2015.

[48] Judea Pearl. An introduction to causal inference. *Int. J. Biostat.*, 6(2):Article 7, February 2010.

[49] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal Inference in Statistics: A Primer*. Standards Information Network, Chichester, West Sussex, UK, February 2016.

[50] Michael R Pinsky, Laurent Brochard, Jordi Mancebo, and Goran Hedenstierna, editors. *Applied physiology in intensive care medicine*. Springer, Berlin, Germany, 2 edition, July 2009.

[51] Zhaozhi Qian, William R Zame, Lucas M Fleuren, Paul Elbers, and Mihaela van der Schaar. Integrating expert ODEs into neural ODEs: Pharmacology and disease progression. *arXiv [cs.LG]*, June 2021.

[52] Yumou Qiu, Jing Tao, and Xiao-Hua Zhou. Inference of heterogeneous treatment effects using observational data with high-dimensional covariates. *J. R. Stat. Soc. Series B Stat. Methodol.*, 83(5):1016–1043, November 2021.

[53] Iqbal Ratnani, Sahar Fatima, Muhammad Mohsin Abid, Zehra Surani, and Salim Surani. Evidence-based medicine: History, review, criticisms, and pitfalls. *Cureus*, 15(2):e35266, February 2023.

[54] Kislaya Ravi, Vladyslav Fediukov, Felix Dietrich, Tobias Neckel, Fabian Buse, Michael Bergmann, and Hans-Joachim Bungartz. Multi-fidelity gaussian process surrogate modeling for regression problems in physics. *arXiv [stat.ML]*, April 2024.

[55] J M Robins, M A Hernán, and B Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, September 2000.

[56] James M Robins and Miguel A Hernán. Estimation of the causal effects of time-varying exposures. 2008.

[57] Yulia Rubanova, Ricky T Q Chen, and David Duvenaud. Latent ODEs for irregularly-sampled time series. *arXiv [cs.LG]*, July 2019.

[58] Muhammad Saqib, Muhammad Iftikhar, Fnu Neha, Fnu Karishma, and Hassan Mumtaz. Artificial intelligence in critical illness and its impact on patient care: a comprehensive review. *Front. Med. (Lausanne)*, 10:1176192, April 2023.

[59] Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. *arXiv [cs.LG]*, June 2022.

[60] Jodi B Segal, Ravi Varadhan, Rolf H H Groenwold, Xiaojuan Li, Kaori Nomura, Sigal Kaplan, Shirin Ardeshirrouhanifard, James Heyward, Fredrik Nyberg, and Mehmet Burcu. Assessing heterogeneity of treatment effect in real-world data. *Ann. Intern. Med.*, 176(4):536–544, April 2023.

[61] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. *arXiv [stat.ML]*, June 2016.

[62] Arthur Sherman. Dynamical systems theory in physiology. *J. Gen. Physiol.*, 138(1):13–19, July 2011.

[63] Tomohiro Shinozaki and Etsuji Suzuki. Understanding marginal structural models for time-varying exposures: Pitfalls and tips. *J. Epidemiol.*, 30(9):377–389, September 2020.

[64] Jim M Smit, Jesse H Krijthe, Jasper van Bommel, and Causal Inference for ICU Collaborators. The future of artificial intelligence in intensive care: moving from predictive to actionable AI. *Intensive Care Med.*, 49(9):1114–1116, September 2023.

[65] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. *arXiv [cs.LG]*, February 2018.

[66] Hao Tu, Scott Moura, Yebin Wang, and Huazhen Fang. Integrating physics-based modeling with machine learning for lithium-ion batteries. *Appl. Energy*, 329(120289):120289, January 2023.

[67] Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. May 2019.

[68] Jean-Louis Vincent. We should abandon randomized controlled trials in the intensive care unit. *Crit. Care Med.*, 38(10 Suppl):S534–8, October 2010.

[69] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, August 2023.

[70] Jian-Xun Wang, Jin-Long Wu, and Heng Xiao. A physics informed machine learning approach for reconstructing reynolds stress modeling discrepancies based on DNS data. *arXiv [physics.flu-dyn]*, June 2016.

[71] Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, Shunshi Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba. Benchmarking Model-Based reinforcement learning. July 2019.

[72] Winnie Xu, Ricky T Q Chen, Xuechen Li, and David Duvenaud. Infinitely deep bayesian neural networks with stochastic differential equations. *arXiv [stat.ML]*, February 2021.

[73] Alex Yartsev. Cardiovascular system. `https://derangedphysiology.com/main/cicm-primary-exam/required-reading/cardiovascular-system`. Accessed: 2024-6-1.

[74] Joo Heung Yoon, Michael R Pinsky, and Gilles Clermont. Artificial intelligence in critical care medicine. *Crit. Care*, 26(1):75, March 2022.

[75] Sven Zenker, Jonathan Rubin, and Gilles Clermont. From inverse problems in mathematical physiology to quantitative differential diagnoses. *PLoS Comput. Biol.*, 3(11):e204, November 2007.

[76] Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ODE-net: Learning hamiltonian dynamics with control. *arXiv [cs.LG]*, September 2019.

Chapter 7

Appendix

7.1 Clinical discussion

Figure 7.1: Severe untreated Rheumatoid Arthritis