

Causal Falsification of Digital Twins

Rob Cornish*

ROB.CORNISH@NTU.EDU.SG

*College of Computing and Data Science
Nanyang Technological University
Singapore*

Muhammad Faaiz Taufiq*

MUHAMMAD.TAUFIQ@STATS.OX.AC.UK

*Department of Statistics
University of Oxford
Oxford, United Kingdom*

Arnaud Doucet

DOUCET@STATS.OX.AC.UK

*Department of Statistics
University of Oxford
Oxford, United Kingdom*

Chris Holmes

CHOLMES@STATS.OX.AC.UK

*Department of Statistics
University of Oxford
Oxford, United Kingdom*

Editor: Ilya Shpitser

Abstract

Digital twins are simulation-based models designed to predict how a real-world process will evolve in response to interventions. This modelling paradigm holds substantial promise in many applications, but rigorous procedures for assessing their accuracy are essential for safety-critical settings. We consider how to assess the accuracy of a digital twin using real-world data. We formulate this as a causal inference problem, which leads to a precise definition of what it means for a twin to be “correct”. Unfortunately, fundamental results from causal inference mean observational data cannot be used to certify a twin in this sense unless potentially tenuous assumptions are made, such as that the data are unconfounded. To avoid these assumptions, we propose instead to find situations in which the twin *is not* correct, and present a general-purpose statistical procedure for doing so. Our approach yields reliable and actionable information about the twin under only the assumption of an i.i.d. dataset of observational trajectories, and remains sound even if the data are confounded. We apply our methodology to a large-scale, real-world case study involving sepsis modelling within the Pulse Physiology Engine, which we assess using the MIMIC-III dataset of ICU patients.

Keywords: causal inference, simulation-based modelling, unmeasured confounding, hypothesis testing

*. Both authors contributed equally to this work.

1. Introduction

There is increasing interest in the use of simulation-based models for obtaining *causal* insights. Such models aim to describe what *would* occur when different actions or interventions are applied to some real-world process of interest, thereby allowing planning and decision-making to be done with a fuller understanding of the different outcomes that may result. In many applications, models of this kind are referred to as *digital twins* (Barricelli et al., 2019; Jones et al., 2020; Niederer et al., 2021). These have been considered for a wide range of use-cases including aviation (Tuegel et al., 2011), manufacturing (Lu et al., 2020), healthcare (Corral-Acero et al., 2020; Coorey et al., 2022), civil engineering (Sacks et al., 2020), and agriculture (Jans-Singh et al., 2020).

Many applications of digital twins are considered safety-critical, which means the cost of deploying an inaccurate twin to production is potentially very high. As such, methodology for assessing the performance of a twin before its deployment is essential for the safe, widespread adoption of digital twins in practice (Niederer et al., 2021). In this work, we consider the problem of assessing twin accuracy and propose a concrete, theoretically grounded, and general-purpose methodology to this end. We focus specifically on the use of statistical methods that leverage data obtained from the real-world process that the twin is designed to model. Such strategies are increasingly viable for many applications as datasets grow larger, and offer the promise of lower overheads compared with alternative strategies that rely for instance on domain expertise.

We formulate twin assessment as a problem of *causal inference* (Rubin, 1974, 2005; Pearl, 2009; Hernán and Robins, 2020). In particular, we consider a twin to be accurate if it correctly captures the behaviour of a real-world process of interest in response to certain interventions, rather than the behaviour of the process as it evolves on its own. This seems in keeping with the overall (if often implicit) design objectives that underlie many applications, including those cited above. Our approach also highlights certain pitfalls associated with conventional methods that do not account for causal factors, and that can give rise to misleading inferences about the twin as a result.

In most applications, it is desirable for an assessment procedure to be reliable and robust, and for its conclusions about the twin to be highly trustworthy. As such, our goal in this paper is to obtain a methodology that is always *sound*, even possibly at the expense of being conservative: we prefer not to draw any conclusion about the accuracy of the twin at all than to draw some conclusion that is potentially misleading. To this end, we rely on minimal assumptions about the twin and the real-world process of interest. In addition to improving robustness, this also means our resulting methodology is very general, and may be applied to a wide variety of twins across application domains.

1.1 Contribution

We begin by providing a causal model for a general-purpose twin and the data we have available for assessment. We use this to show precisely that it is not possible to use observational data to *certify* that the twin is causally accurate unless strong and often tenuous assumptions are made about the data-generating process, such as that the data are free of unmeasured confounding. To avoid these assumptions, we propose an assessment paradigm

instead based on *falsification*: we search for specific cases when the twin fails to be correct, rather than trying to quantify its overall accuracy in a more holistic sense.

To obtain a practical methodology suitable for real twins, we provide a novel set of longitudinal causal bounds that hold without additional assumptions. These bounds generalize the classical result of Manski (1990), and can be considerably more informative in comparison. We use this result as the basis for a general-purpose statistical testing procedure for falsifying a twin. Overall, our method relies on only the assumption of an i.i.d. dataset of observational trajectories: it does not require modelling the dynamics of the real-world process or any internal implementation details of the twin, and remains sound in the presence of arbitrary unmeasured confounding. We demonstrate the effectiveness of our procedure through a large-scale, real-world case study in which we use the MIMIC-III ICU dataset (Johnson et al., 2016) to assess the Pulse Physiology Engine (Bray et al., 2019), an open-source model for human physiology simulation.

1.2 Related Work

Various high-level guidelines and workflows have been proposed for the assessment of digital twins in the literature to date (Roy and Oberkamp, 2011; Grieves and Vickers, 2017; Khan et al., 2018; Corral-Acero et al., 2020; Kochunas and Huan, 2021; Niederer et al., 2021; Dahmen et al., 2022). In some cases, these guidelines have been codified as standards: for example, the ASME V&V40 Standard (ASME, 2018) provides a risk-based framework for assessing the credibility of a model from a variety of factors that include source code quality and the mathematical form of the model (Galappaththige et al., 2022). However, a significant gap still exists between these guidelines and a practical implementation that could be deployed for real twins, and the need for a rigorous lower-level framework to enable the systematic assessment of twins has been noted in this literature (Corral-Acero et al., 2020; Niederer et al., 2021; Kapteyn et al., 2021; Masison et al., 2021). We contribute towards this effort by describing a precise statistical methodology for twin assessment that can be readily implemented in practice, and which is accompanied by theoretical guarantees of robustness that hold under minimal assumptions.

In addition, a variety of concrete assessment procedures have been applied to certain specific digital twin models in the literature. For example, the Pulse Physiology Engine (Bray et al., 2019), which we consider in our empirical case study, as well as the related BioGears (McDaniel et al., 2019; McDaniel and Baird, 2019) were both assessed by comparing their outputs with ad hoc values based on available medical literature and the opinions of subject matter experts. Other twins have been assessed by comparing their outputs with real-world data through a variety of bespoke numerical schemes (Larrabide et al., 2012; Hemmler et al., 2019; Lal et al., 2021; Jans-Singh et al., 2020; Galappaththige et al., 2022). In contrast, our paper proposes a general-purpose statistical procedure for assessing twins that may be applied generically across many applications and architectures. To the best of our knowledge, we are also the first to identify the need for a causal approach to twin assessment and the pitfalls that arise when causal considerations are not accounted for.

Overall, our assessment approach follows the guidelines for empirical analysis proposed by econometrician Charles Manski (Manski, 1995), who advocates for researchers to establish a “domain of consensus” (Manski, 2009) by analysing their data initially under

minimal assumptions, and only afterwards to consider what follows as additional assumptions are introduced. To this end, Manski and others developed a literature on *partial identification* (Manski, 1989, 1990; Balke and Pearl, 1997; Manski, 1997; Manski and Pepper, 2000; Manski, 2003) to provide statistical tools with which this high-level strategy may be implemented. This literature provides a suite of tools for reasoning about models whose parameters cannot be point-identified using observational data. Similar ideas and techniques have also appeared in more recent work (Zhang and Bareinboim, 2019; Namkoong et al., 2020; Hussain et al., 2022; Zhang et al., 2022). However, to the best of our knowledge, existing results in this vast literature are largely not tailored towards assessing twin models, due in particular to their online and longitudinal structure, which complicates analysis. In particular, a direct application of Manski’s classical bounds (Manski, 1990) to our conditional longitudinal setting would not be identifiable in general. Accordingly, in Section 4 we provide a novel set of identifiable bounds on Manski’s bounds, which we use as the basis for our practical methodology in Section 5. Our bounds require minimal assumptions, and we provide various analyses showing they cannot be improved upon unless additional assumptions are added.

The partial identification literature has developed a suite of inference techniques for parameters that are not point-identified. A major line of work provides confidence sets for individual parameters in these settings (Imbens and Manski, 2004; Rosen, 2008; Romano et al., 2014), while another constructs confidence sets for the entire identification region (Chernozhukov et al., 2007; Romano and Shaikh, 2010). These methods often seek a high level of generality, which has led to various inferential machinery based on subsampling, moment selection, and related asymptotic techniques (Romano and Shaikh, 2008; Andrews and Soares, 2010) (see Molinari 2020 and Canay et al. 2023 for reviews). Our testing procedure described in Section 5 follows the same overall logic as this second line of work, but with our end goal being testing rather than obtaining confidence sets. This reduces the problem to one-sided tests of conditional expectations, for which Hoeffding’s inequality yields valid finite-sample type I error control without asymptotic approximations. A separate line of work considers specification tests for whether a partially identified region is non-empty (Bugni et al., 2015; Kitagawa, 2015; Mourifié and Wan, 2017). This is distinct from our problem, since our bounds are always guaranteed to be nonempty by construction.

2. Causal Formulation

2.1 The Real-World Process

We begin by providing a causal model of the real-world process that the twin is designed to simulate. We do so in the language of *potential outcomes* (Rubin, 1974, 2005), although we could have used the alternative framework of directed acyclic graphs and structural causal models (Pearl, 2009). We assume the real-world process operates over a fixed time horizon $T \in \{1, 2, \dots\}$. This simplifies our presentation in what follows, and it is straightforward to generalize our methodology to variable length time horizons if needed. For each $t \in \{0, \dots, T\}$, we assume the process gives rise to an *observation* at time t , which takes values in some real-valued space $\mathcal{X}_t := \mathbb{R}^{d_t}$. We also assume that the process can be influenced by some *action* taken at each time $t \in \{1, \dots, T\}$. We denote the space of actions available at time t by $\mathcal{A}_t$, which in this work we assume is always finite. For example, in a robotics

context, the observations may consist of all the readings of all the sensors of the robot, and the actions may consist of commands that can be input by an external user. In a medical context, the observations may consist of the vital signs of a patient, and the actions may consist of possible treatments or interventions. To streamline notation, we will index these spaces using vector notation, so that e.g. $\mathcal{A}_{1:t}$ denotes the cartesian product $\mathcal{A}_s \times \dots \times \mathcal{A}_t$, and $a_{s:t} \in \mathcal{A}_{s:t}$ is a choice of $a_s \in \mathcal{A}_s, \dots, a_t \in \mathcal{A}_t$. We also allow $s > t$ in this notation, so that e.g. $a_{s:t}$ denotes the empty sequence $\emptyset$, and $\mathcal{A}_{s:t}$ denotes the set $\{\emptyset\}$ containing just the empty sequence and nothing else (so we still have $a_{s:t} \in \mathcal{A}_{s:t}$).

We model the dynamics of the real-world process via the longitudinal potential outcomes framework proposed by Robins (1986), which imposes only a weak temporal structure on the underlying phenomena of interest and so may be applied across a wide range of applications in practice. In particular, for each $a_{1:T} \in \mathcal{A}_{1:T}$, we posit the existence of random variables or *potential outcomes* $X_0, X_1(a_1), \dots, X_T(a_{1:T})$, where $X_t(a_{1:t})$ takes values in $\mathcal{X}_t$. We will denote this sequence more concisely as $X_{0:T}(a_{1:T})$. Intuitively, X_0 represents data available before the first action, while $X_{1:T}(a_{1:T})$ represents the sequence of real-world outcomes that would occur if actions $a_{1:T}$ were taken successively. These quantities are therefore of fundamental interest for planning a course of actions to achieve some desired result.

As random variables, each $X_t(a_{1:t})$ may depend on additional randomness that is not explicitly modelled, and so in particular may be influenced by all the previous potential outcomes $X_{0:t-1}(a_{1:t-1})$, and possibly other random quantities. Notice however that each $X_s(a_{1:s})$ does *not* depend on the subsequent actions $a_{s+1:T}$, which amounts to saying that future actions do not affect past outcomes. This models a process whose initial state is determined by external factors, such as when a patient from some population first presents at a hospital, and where the process then evolves according to specific actions chosen from $\mathcal{A}_{1:T}$ as well as exogenous factors. It is clear that this structure applies to a wide range of phenomena occurring in practice.

2.2 The Digital Twin

We think of the twin as a computational device that, when executed, outputs a sequence of values intended to simulate a possible future trajectory of the real-world process when certain actions in $\mathcal{A}_{1:T}$ are chosen, conditional on some initial data in $\mathcal{X}_0$. We allow the twin to make use of an internal random number generator to produce outputs that vary stochastically even under fixed inputs (although our framework encompasses twins that evolve deterministically also). By executing the twin repeatedly, a user may therefore estimate the range of behaviours that the real-world process may exhibit under different action sequences, which can then inform planning and decision-making downstream.

Precisely, we model the output that the twin would produce at timestep $t \in \{1, \dots, T\}$ after receiving initialisation $x_0 \in \mathcal{X}_0$ and successive inputs $a_{1:t} \in \mathcal{A}_{1:t}$ as the quantity $h_t(x_0, a_{1:t}, U_{1:t})$, where h_t is a measurable function taking values in $\mathcal{X}_t$, and each U_s is some random variable. We will denote $\hat{X}_t(x_0, a_{1:t}) := h_t(x_0, a_{1:t}, U_{1:t})$, which we also refer to as a potential outcome. A full twin trajectory therefore consists of $\hat{X}_1(x_0, a_1), \dots, \hat{X}_T(x_0, a_{1:T})$, which we write more compactly by $\hat{X}_{1:T}(x_0, a_{1:T})$. Conceptually, $h_1, \dots, h_T$ constitute the program that executes inside the twin, and $U_{1:T}$ may be thought of as the collection of all outputs of the internal random number generator that the twin uses. We assume these

internal random numbers $U_{1:T}$ and the initial real-world observations X_0 are independent, which is mild in practice. We also assume that repeated executions of the twin give rise to i.i.d. copies of $U_{1:T}$. This means that, given fixed inputs x_0 and $a_{1:T}$, repeated executions of the twin produce i.i.d. copies of $\widehat{X}_{1:T}(x_0, a_{1:T})$. Otherwise, we make no assumptions about the precise form of either the h_t or the U_t , which allows our model to encompass a wide variety of possible twin implementations.

2.3 Correctness

Before we can consider how to assess the twin, we must first define how we want the twin ideally to behave. The following condition seems appropriate for many applications.

Definition 1 *The twin is interventionally correct if, for $\text{Law}[X_0]$ -almost all $x_0 \in \mathcal{X}_0$ and $a_{1:T} \in \mathcal{A}_{1:T}$, the distribution of $\widehat{X}_{1:T}(x_0, a_{1:T})$ is equal to the conditional distribution of $X_{1:T}(a_{1:T})$ given $X_0 = x_0$.*

Operationally, if a twin is interventionally correct, then by repeatedly executing the twin and applying Monte Carlo techniques, it is possible to approximate arbitrarily well the conditional distribution of the future of the real-world process under each possible choice of action sequence. The same can also be shown to hold when each action at each time t is chosen dynamically on the basis of previous observations in $\mathcal{X}_{0:t}$. As a result, an interventionally correct twin may be used for *planning*, or in other words may be used to select a policy for choosing actions that will yield a desirable distribution over observations at each step. We emphasise that interventional correctness does not mean the twin will accurately predict the behaviour of any *specific* trajectory of the real-world process in an almost sure sense (unless the real-world process is deterministic), but only the distribution of outcomes that will be observed over repeated independent trajectories. However, this is sufficient for many applications, and appears to be the strongest guarantee possible when dealing with real-world phenomena whose underlying behaviour is stochastic.

Definition 1 introduces some technical difficulties that arise in the general case when conditioning on events with probability zero (e.g. $\{X_0 = x_0\}$ if X_0 is continuous). In what follows, it is more convenient to consider an unconditional formulation of interventional correctness. This is supplied by the following straightforward result, which considers the behaviour of the twin when it is initialised with the (random) value of X_0 taken from the real-world process, rather than with a fixed choice of x_0 . See Section B of the Appendix for a proof.

Proposition 2 *The twin is interventionally correct if and only if, for all choices of $a_{1:T} \in \mathcal{A}_{1:T}$, the distribution of $(X_0, \widehat{X}_{1:T}(X_0, a_{1:T}))$ is equal to the distribution of $X_{0:T}(a_{1:T})$.*

2.4 Online Prediction

Our model here represents a twin at time $t = 0$ making predictions about all future timesteps $t \in \{1, \dots, T\}$ under different choices of inputs $a_{1:T}$. In practice, many twins are designed to receive new information at each timestep in an online fashion and update their predictions for subsequent timesteps accordingly (Grieves and Vickers, 2017; Niederer et al., 2021). Various notions of correctness can be devised for this online setting. We describe two

possibilities in Section C of the Appendix, and show that these notions of correctness essentially reduce to Definition 1, which motivates our focus on that notion in what follows.

3. Data-Driven Twin Assessment

There are many conceivable methods for assessing the accuracy of a twin, including static analysis of the twin’s source code and the solicitation of domain expertise, and in practice it seems most robust to use a combination of different techniques rather than relying on any single one (ASME, 2018; Niederer et al., 2021). However, in this paper, we focus on what may be called *data-driven assessment*, which we see as an important component of a larger assessment pipeline. That is, we consider the use of statistical methods that rely solely on a dataset of trajectories obtained from the real-world process and the twin. Unfortunately, as we now explain, fundamental results from the causal inference literature mean that it is not possible to obtain a data-driven assessment procedure that can *certify* that a twin is interventionally correct without further assumptions. Accordingly, we propose a strategy based on *falsifying* the twin, which we develop into a concrete statistical testing procedure in later sections.

We will assume access to a dataset of trajectories obtained by observing the interaction of some behavioural agents with the real-world process. We model each trajectory as follows. First, we represent the action chosen by the agent at time $t \in \{1, \dots, T\}$ as an $\mathcal{A}_t$ -valued random variable A_t . We then obtain a trajectory in our dataset by recording at each step the action A_t chosen and the observation $X_t(A_{1:t})$ corresponding to this choice of action. As a result, each observed trajectory has the following form:

$$X_0, A_1, X_1(A_1), \dots, A_T, X_T(A_{1:T}). \quad (1)$$

This is the standard *consistency* assumption in causal inference (Hernán and Robins, 2020), and intuitively means the potential outcome $X_t(a_{1:t})$ is observed in the data when the agent actually chose $A_{1:t} = a_{1:t}$. We model our full dataset as a set of i.i.d. copies of (1).

3.1 Certification Is Unsound in General

A natural high-level strategy for twin assessment has the following structure. First, some hypothesis $\mathcal{H}$ is chosen with the following property:

$$\text{If } \mathcal{H} \text{ is true, then the twin is interventionally correct.} \quad (2)$$

Data is then used to try to show $\mathcal{H}$ is true, perhaps up to some level of confidence. If successful, it follows by construction that the twin is interventionally correct. Assessment procedures designed to *certify* the twin in this way are appealing because they promise a strong guarantee of accuracy for certified twins. Unfortunately, the following foundational result from the causal inference literature (often referred to as the *fundamental problem of causal inference*; Holland, 1986) means that data-driven certification procedures of this kind are in general unsound, as we explain next. This is a standard and well-known result within the field of causal inference, but appears to be less widely known within the digital twin literature. For completeness, Section D of the Appendix includes a self-contained proof of this result in our notation.

Theorem 3 *If $\mathbb{P}(A_{1:T} \neq a_{1:T}) > 0$, then the distribution of $X_{0:T}(a_{1:T})$ is not uniquely identified by the distribution of the data in (1) without further assumptions.*

Since the distribution of the data encodes the information that would be contained in an infinitely large dataset of trajectories, Theorem 3 imposes a fundamental limit on what can be learned about the distribution of $X_{0:T}(a_{1:T})$ from the data we have assumed. It follows that if $\mathcal{H}$ is any hypothesis satisfying (2), then $\mathcal{H}$ cannot be determined to be true from even an infinitely large dataset. This is because, if we could do so, then we could also determine the distribution of $X_{0:T}(a_{1:T})$, since by Proposition 2 this would be equal to the distribution of $(X_0, \hat{X}_T(X_0, a_{1:T}))$. In other words, we cannot use the data alone to certify that the twin is interventionally correct.

3.2 The Assumption of No Unmeasured Confounding

Theorem 3 is true in the general case, when no additional assumptions about the data-generating process are made. One way forward is therefore to introduce assumptions under which the distribution of $X_{0:T}(a_{1:T})$ can be identified. This would mean it is possible to certify that the twin is interventionally correct, since, at least in principle, we could simply check whether this matches the distribution of $(X_0, \hat{X}_{1:T}(X_0, a_{1:T}))$ produced by the twin.

The most common such assumption in the causal inference literature is that the data are free of *unmeasured confounding*. Informally, this holds when each action A_t is chosen by the behavioural agent solely on the basis of the information available at time t that is actually recorded in the dataset, namely $X_0, A_1, X_1(A_1), \dots, A_{t-1}, X_{t-1}(A_{1:t-1})$, as well as possibly some additional randomness that is independent of the real-world process, such as the outcome of a coin toss. (This can be made precise via the *sequential randomisation assumption* introduced by Robins (1986).) Unobserved confounding is present whenever this does not hold, i.e. whenever some unmeasured factor simultaneously influences both the agent’s choice of action and the observation produced by the real-world process.

It is reasonable to assume that the data are unconfounded in certain contexts. For example, it may be possible to gather data in a way that specifically guarantees there is no confounding. Randomised controlled trials, which ensure that each A_t is chosen via a carefully designed randomisation procedure (Lavori and Dawson, 2004; Murphy, 2005), constitute a widespread example of this approach. Likewise, it is possible to show that the data are unconfounded if each $X_t(a_{1:t})$ is a deterministic function of $X_{0:t-1}(a_{1:t-1})$ and a_t , which may be reasonable to assume for example in certain low-level physics or engineering contexts. (See Section E of the Appendix for a proof.) However, for stochastic phenomena and for typical datasets, it is widely acknowledged that the assumption of no unmeasured confounding will rarely hold (Murphy, 2003; Tsiatis et al., 2019), and so assessment procedures based on this assumption may yield unreliable results in practice. Section F of the Appendix illustrates this concretely with a toy scenario.

3.3 General-Purpose Assessment via Falsification

Our goal is to obtain an assessment methodology that is general-purpose, and as such we would like to avoid introducing assumptions such as unconfoundedness that do not hold in general. To achieve this, borrowing philosophically from Popper (2005), we propose a

strategy that replaces the goal of verifying the interventional correctness of the twin with that of *falsifying* it. Specifically, we consider hypotheses $\mathcal{H}$ with the dual property to (2):

$$\text{If the twin is interventionally correct, then } \mathcal{H} \text{ is true.} \quad (3)$$

We will then try to show that each such $\mathcal{H}$ is *false*. Whenever we are successful, we will thereby have gained some knowledge about a failure mode of the twin, since by construction the twin can only be correct if $\mathcal{H}$ is true. In effect, each $\mathcal{H}$ we falsify will constitute a reason that the twin is not correct, and may suggest concrete improvements to its design, or may identify cases where its output should not be trusted.

Importantly, unlike for (2), Theorem 3 does not preclude the possibility of data-driven assessment procedures based on (3). As we show below, there do exist hypotheses $\mathcal{H}$ satisfying (3) that can in principle be determined to be false from the data alone without additional assumptions. In this sense, falsification provides a means for *sound* data-driven twin assessment, whose results can be relied upon across a wide range of circumstances. On the other hand, falsification approaches cannot provide a *complete* guarantee about the accuracy of a twin: even if we fail to falsify many $\mathcal{H}$ satisfying (3), we cannot then infer that the twin is correct. As such, in situations where (for example) it is reasonable to believe that the data are in fact unconfounded, it may be desirable to use this assumption to obtain additional information about the twin than is possible from falsification alone.

4. Longitudinal Causal Bounds

4.1 Manski's Bounds

A classical result by Manski (1990) gives a worst-case bound on the average interventional behaviour of a causal process. In our context, and using the notation established above, this result gives the following:

Theorem 4 (Manski, 1990) *Suppose $(Y(a'_{1:t}) : a'_{1:t} \in \mathcal{A}_{1:t})$ are real-valued potential outcomes defined jointly with $(X_{0:T}(a'_{1:T}) : a'_{1:T} \in \mathcal{A}_{1:T})$ and $A_{1:T}$. Suppose that for some $t \in \{1, \dots, T\}$, $a_{1:t} \in \mathcal{A}_{1:t}$, measurable $B_{0:t} \subseteq \mathcal{X}_{0:t}$, and $y_{\text{lo}}, y_{\text{up}} \in \mathbb{R}$ we have*

$$\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t}) > 0 \quad (4)$$

$$\mathbb{P}(y_{\text{lo}} \leq Y(a_{1:t}) \leq y_{\text{up}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) = 1. \quad (5)$$

It then holds that

$$\mathbb{E}[Y_{\text{lo}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}] \leq \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}] \leq \mathbb{E}[Y_{\text{up}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}], \quad (6)$$

where we denote

$$\begin{aligned} Y_{\text{lo}} &:= \mathbb{1}(A_{1:t} = a_{1:t}) Y(A_{1:t}) + \mathbb{1}(A_{1:t} \neq a_{1:t}) y_{\text{lo}} \\ Y_{\text{up}} &:= \mathbb{1}(A_{1:t} = a_{1:t}) Y(A_{1:t}) + \mathbb{1}(A_{1:t} \neq a_{1:t}) y_{\text{up}}. \end{aligned}$$

Specifically, this follows by instantiating the main result of Manski (1990) with everything conditional on the event $\{X_{0:t}(a_{1:t}) \in B_{0:t}\}$, which has nonzero probability by (4).

The potential outcome $Y(a_{1:t})$ may be thought of as some quantitative outcome of interest. For example, in a medical context, $Y(a_{1:t})$ might represent the heart rate of a patient at time t after receiving some treatments $a_{1:t}$. When defining our hypotheses below, we consider the specific form $Y(a_{1:t}) := f(X_{0:t}(a_{1:t}))$, where $f : \mathcal{X}_{0:t} \rightarrow \mathbb{R}$ is some scalar function. The bounded quantity $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}]$ in (6) then just gives the average behaviour of this outcome.

From a methodological perspective, Theorem 4 has the following well-known consequence. By taking $B_{1:t} = \mathcal{X}_{1:t}$ to be the whole space for all but the initial timestep, the bounds in Theorem 4 specialise to the following:

$$\mathbb{E}[Y_{\text{lo}} \mid X_0 \in B_0] \leq \mathbb{E}[Y(a_{1:t}) \mid X_0 \in B_0] \leq \mathbb{E}[Y_{\text{up}} \mid X_0 \in B_0]. \quad (7)$$

By Theorem 3, the middle term is in general not identified by the data since it depends on $Y(a_{1:t})$, which is not observed unless $A_{1:t} = a_{1:t}$. On the other hand, both the lower and upper bounds *are* identified, since the relevant random variables Y_{lo} , Y_{up} , and X_0 can all be expressed as functions of the observed data $X_{0:t}(A_{1:t})$ and $A_{1:t}$. In this way, Theorem 4 bounds the behaviour of a non-identifiable quantity in terms of identifiable ones. At a high level, this is achieved by replacing $Y(a_{1:t})$, whose value is only observed on the event $\{A_{1:t} = a_{1:t}\}$, with Y_{lo} and Y_{up} , which are equal to $Y(a_{1:t})$ when its value is observed (i.e. when $A_{1:t} = a_{1:t}$), and which fall back to the worst-case values of y_{lo} and y_{up} otherwise (see e.g. Imbens and Wooldridge 2009, Section 6.1; Molinari 2020, Sections 2.1–2.2). This does not require any additional causal assumptions beyond (4) and (5), and in particular remains true under arbitrary unmeasured confounding.

However, given general conditioning sets $B_{1:t} \subsetneq \mathcal{X}_{1:t}$, the bounds in Theorem 4 are usually *not* identifiable. This is because they depend on the quantity $X_{0:t}(a_{1:t})$, which is not observed in the data when $A_{1:t} \neq a_{1:t}$. In other words, Manski’s bounds do not allow for conditioning on the intermediate interventional behaviour of the real-world process: they can only be applied to bound the average interventional behaviour of the real-world process with all longitudinal information ignored except for the initial observation X_0 . For the falsification methodology we develop in Section 5, this is a significant limitation, since conditioning on intermediate observations yields a richer family of hypotheses and more granular information about the twin’s accuracy (see Section 4.3).

4.2 Identifiable Conditional Bounds

As a way forward, we provide *bounds on the bounds* (6) in terms of observed quantities, thereby permitting their use in our falsification procedure below. In particular, we have the following key result:

Lemma 5 *Consider the same setup as Theorem 4, and define*

$$N := \max\{0 \leq s \leq t \mid A_{1:s} = a_{1:s}\},$$

where this maximum is well-defined since $A_{1:0} = \emptyset = a_{1:0}$ by our convention from Section 2.1. The upper and lower bounds in (6) are then themselves bounded as follows:

$$\begin{aligned} \mathbb{E}[Y_{\text{lo}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}] &\geq \mathbb{E}[Y_{\text{lo}} \mid X_{0:N}(A_{1:N}) \in B_{0:N}] \\ \mathbb{E}[Y_{\text{up}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}] &\leq \mathbb{E}[Y_{\text{up}} \mid X_{0:N}(A_{1:N}) \in B_{0:N}]. \end{aligned} \quad (8)$$

See Appendix G for a proof. Intuitively, the random variable N encodes the largest value of $s \in \{0, \dots, t\}$ for which the prefix $a_{1:s}$ of the full action sequence $a_{1:t}$ agrees with the actions $A_{1:s}$ actually taken by the agent up to that timestep. Another way to write this is as follows:

$$N = \sum_{s=1}^t \mathbb{1}(A_{1:s} = a_{1:s}).$$

It follows that N and hence $X_{0:N}(A_{1:N})$ are always observed by construction, making the right-hand sides of (8) identifiable.

By combining Lemma 5 with (6), we immediately obtain the following identifiable counterpart of Theorem 4. This result will form the basis for the falsification procedure we develop in Section 5.

Theorem 6 *Under the same setup as Theorem 4, and with N as defined in Lemma 5, it holds that*

$$Q_{\text{lo}} \leq Q \leq Q_{\text{up}}, \quad (9)$$

where we denote

$$\begin{aligned} Q &:= \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}] \\ Q_{\text{lo}} &:= \mathbb{E}[Y_{\text{lo}} \mid X_{0:N}(A_{1:N}) \in B_{0:N}] \\ Q_{\text{up}} &:= \mathbb{E}[Y_{\text{up}} \mid X_{0:N}(A_{1:N}) \in B_{0:N}]. \end{aligned}$$

Another interpretation of this result is as follows. Analogously to Theorem 4, by applying the classical result of Manski (1990) with everything conditioned on the event $\{X_{0:N}(A_{1:N}) \in B_{0:N}\}$, we obtain directly

$$Q_{\text{lo}} \leq \mathbb{E}[Y(a_{1:t}) \mid X_{0:N}(a_{1:N}) \in B_{0:N}] \leq Q_{\text{up}}.$$

Theorem 6 shows that these bounds are also valid when the middle term here is replaced by Q . This is a more useful quantity for practical purposes than $\mathbb{E}[Y(a_{1:t}) \mid X_{0:N}(a_{1:N}) \in B_{0:N}]$, since the interaction between the random quantity N and the sequence of observations $X_{0:N}(a_{1:N})$ makes the latter harder to interpret and use in practice.

In the structural causal modelling framework (Pearl, 2009), a related result to Theorem 6 was given as Corollary 1 by Zhang and Bareinboim (2019). However, their result involves a complicated ratio of unknown quantities that makes estimation of their bounds difficult, since it is not obvious how to obtain an unbiased estimator for their ratio term. In contrast, our proposed bounds are considerably simpler, since both Q_{lo} and Q_{up} here are expressed as (conditional) expectations. This makes their unbiased estimation straightforward, which we use to obtain exact confidence intervals for both terms in Section 5.2.

4.3 Informativeness

For Theorem 6 to be useful in practice, we would like the bounds $[Q_{\text{lo}}, Q_{\text{up}}]$ to be relatively narrow compared with the worst-case bounds $[y_{\text{lo}}, y_{\text{up}}]$ that are trivially implied by (5). We can quantify the extent to which this occurs by the ratio

$$\frac{Q_{\text{up}} - Q_{\text{lo}}}{y_{\text{up}} - y_{\text{lo}}} = 1 - \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}), \quad (10)$$

where the equality here follows from the definitions of Q_{lo} and Q_{up} together with some straightforward manipulations. In other words, the (relative) tightness of our bounds is determined by the value of $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$, which is itself closely related to the classical *propensity score* in the causal literature (Rosenbaum and Rubin, 1983). Intuitively, as this probability grows, so does $\mathbb{P}(Y(a_{1:t}) = Y(A_{1:t}) \mid X_{0:N}(A_{1:N}) \in B_{0:N})$, which means the effect of unmeasured confounding on the value of Q is reduced, leading to tighter bounds.

The bounds (7) that follow directly from the original result of Manski (1990) are often regarded as quite uninformative. These correspond to the special case $B_{1:t} = \mathcal{X}_{1:t}$, and inspection of (10) shows their width is relatively large whenever $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_0 \in B_0)$ is small. This occurs in many applications, particularly for longer action sequences. On the other hand, it often seems reasonable to expect that certain longer action sequences will be fairly likely to occur when conditioned on some intermediate observations. In other words, $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$ may be large (which means $[Q_{\text{lo}}, Q_{\text{up}}]$ is narrow), even if $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_0 \in B_0)$ is small. By choosing $B_{0:t}$ carefully, we therefore have an opportunity to obtain narrower bounds than those provided by Manski’s original result. The following result provides a straightforward sufficient condition for this to hold.

Proposition 7 *Consider the same setup as Theorem 6, and suppose it also holds that $\mathbb{P}(y_{\text{lo}} \leq Y(a_{1:t}) \leq y_{\text{up}} \mid X_0 \in B_0) = 1$. Then the width of Manski’s bounds (7) exceeds that of Theorem 6, i.e.*

$$\mathbb{E}[Y_{\text{up}} \mid X_0 \in B_0] - \mathbb{E}[Y_{\text{lo}} \mid X_0 \in B_0] > Q_{\text{up}} - Q_{\text{lo}},$$

whenever we have $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}) > \mathbb{P}(A_{1:t} = a_{1:t} \mid X_0 \in B_0)$.

We note that Proposition 7 compares bounds on two different quantities, namely Q (for our bounds) and $\mathbb{E}[Y(a_{1:t}) \mid X_0 \in B_0]$ (for Manski’s bounds). However, it still appears reasonable to compare the widths of both bounds, since these two quantities share the same worst-case behaviour of $[y_{\text{lo}}, y_{\text{up}}]$ in both cases.

Beyond the appeal of tightness, the conditional nature of Theorem 6 also appears of interest simply for its own sake. In particular, Theorem 6 describes the interventional behaviour of $Y(a_{1:t})$ conditional on the behaviour of the trajectory $X_{0:t}(a_{1:t})$, thereby providing more granular information than can be obtained from Manski’s bounds (7) alone. In Section 6 below, we also demonstrate empirically that conditioning on intermediate observations leads to substantially more informative bounds in our case study.

4.4 Sharpness

An immediate question is whether our bounds in Theorem 6 are optimal. For example, an alternative approach to obtaining bounds is to consider actions taken by the agent after time t . This yields the following lower bound:

$$\sup_{a'_{t+1:T} \in \mathcal{A}_{t+1:T}} \mathbb{E}[Y(A_{1:t}) \mathbb{1}\{A_{1:T} = (a_{1:t}, a'_{t+1:T})\} + y_{\text{lo}} \mathbb{1}\{A_{1:T} \neq (a_{1:t}, a'_{t+1:T})\}] \leq \mathbb{E}[Y(a_{1:t})], \quad (11)$$

since the bound holds for any fixed $a'_{t+1:T}$ when the supremum is removed from the left-hand side. The cases of the upper bound and of nontrivial conditioning event are then also similar.

The following result shows that Theorem 6 cannot be improved without further assumptions. Intuitively speaking, the idea is that there always exists *some* family of potential outcomes that produces the same observational data as our model, but that attains the worst-case bounds Q_{lo} or Q_{up} . Therefore, we cannot rule out the possibility that the true potential outcomes achieve Q_{lo} or Q_{up} from the observational data alone. In particular this means that the alternative lower bound (11) (which only depends on the observational distribution) is in fact not tighter than Q_{lo} .¹

Proposition 8 *Under the same setup as in Theorem 6, there always exists potential outcomes $(\tilde{X}_{0:T}(a'_{1:T}), \tilde{Y}(a'_{1:t}) : a'_{1:T} \in \mathcal{A}_{1:T})$ also satisfying (4) and (5) (mutatis mutandis) with*

$$(\tilde{X}_{0:T}(A_{1:T}), \tilde{Y}(A_{1:t}), A_{1:T}) \stackrel{\text{a.s.}}{=} (X_{0:T}(A_{1:T}), Y(A_{1:t}), A_{1:T})$$

but for which $\mathbb{E}[\tilde{Y}(a_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t}) \in B_{0:t}] = Q_{\text{lo}}$. The corresponding statement is also true for Q_{up} .

4.5 Continuous Bounds Must Be Trivial

Instead of bounding Q , in some cases we may wish to consider bounding the alternative quantity $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})]$ that conditions on the *value* of $X_{0:t}(a_{1:t})$ rather than on the event $\{X_{0:t}(a_{1:t}) \in B_{0:t}\}$. To achieve this, it would be natural to generalize our assumption (5) by supposing we now have measurable functions $y_{\text{lo}}, y_{\text{up}} : \mathcal{X}_{0:t} \rightarrow \mathbb{R}$ such that

$$y_{\text{lo}}(X_{0:t}(a_{1:t})) \leq Y(a_{1:t}) \leq y_{\text{up}}(X_{0:t}(a_{1:t})) \quad \text{almost surely,} \quad (12)$$

and taking as our goal to obtain measurable functions $g_{\text{lo}}, g_{\text{up}} : \mathcal{X}_{0:t} \rightarrow \mathbb{R}$ such that

$$g_{\text{lo}}(X_{0:t}(a_{1:t})) \leq \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})] \leq g_{\text{up}}(X_{0:t}(a_{1:t})) \quad \text{almost surely.} \quad (13)$$

As we describe in Section G.4 of the Appendix, bounds of this kind can be obtained directly from Theorem 6 if $X_{0:t}(a_{1:t})$ is discrete, or by a simple modification of the proof of Theorem 6 if $X_{1:t}(a_{1:t})$ is discrete (but X_0 is possibly continuous). However, somewhat surprisingly, in general we cannot obtain nontrivial bounds of this kind without further assumptions in the continuous case. To make this precise, we will say that a given g_{lo} and g_{up} are *permissible* if (13) holds when $(X_{0:T}(a'_{1:t}), Y(a'_{1:t}), A_{1:T} : a'_{1:T} \in \mathcal{A}_{1:T})$ are replaced by any potential outcomes $(\tilde{X}_{0:T}(a'_{1:T}), \tilde{Y}(a'_{1:t}), \tilde{A}_{1:T} : a'_{1:T} \in \mathcal{A}_{1:T})$ for which

$$\text{Law}[\tilde{X}_{0:T}(\tilde{A}_{1:T}), \tilde{A}_{1:T}, \tilde{Y}(\tilde{A}_{1:t})] = \text{Law}[X_{0:T}(A_{1:T}), A_{1:T}, Y(A_{1:t})],$$

and which also satisfy (12) (mutatis mutandis). Intuitively, this means that g_{lo} and g_{up} depend only on the information we have available, i.e. the observational distribution and our assumed worst-case values. We then have the following:

1. This can also be seen directly as a consequence of the fact that $\mathbb{P}(A_{1:T} = (a_{1:t}, a'_{t+1:T})) \leq \mathbb{P}(A_{1:t} = a_{1:t})$ together with the assumption that $Y(a_{1:t}) \geq y_{\text{lo}}$ almost surely.

Theorem 9 *Suppose X_0 is almost surely constant, $\mathbb{P}(A_1 \neq a_1) > 0$, and for some $s \in \{1, \dots, t\}$ we have $\mathbb{P}(X_s(A_{1:s}) = x_s) = 0$ for all $x_s \in \mathcal{X}_s$. Then $g_{\text{lo}}, g_{\text{up}} : \mathcal{X}_{0:t} \rightarrow \mathbb{R}$ are permissible bounds only if they are trivial, i.e. almost surely we have*

$$g_{\text{lo}}(X_{0:t}(a_{1:t})) \leq y_{\text{lo}}(X_{0:t}(a_{1:t})) \quad \text{and} \quad g_{\text{up}}(X_{0:t}(a_{1:t})) \geq y_{\text{up}}(X_{0:t}(a_{1:t})).$$

Here the assumption that X_0 is constant essentially means we consider a special case of our model where there are no covariates available before the first action is taken, and serves mainly to simplify the proof. We conjecture that Theorem 9 holds more generally, provided the other assumptions are accordingly made to be conditional on X_0 also. In any case, this result shows that general-purpose bounds on $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})]$ are not forthcoming, and Theorem 6 is the best we can hope for in general. We show below that this result is nevertheless powerful enough to obtain useful information in practice.

5. Falsification Methodology

5.1 Hypotheses Derived from Causal Bounds

We now use Theorem 6 to obtain a hypothesis testing procedure that can be used to falsify the twin, and that does not rely on any further assumptions than we have already provided. To this end, we first define the hypotheses $\mathcal{H}$ satisfying (3) that we will consider. Each of these will depend on a specific choice of the following parameters:

A timestep $t \in \{1, \dots, T\}$	A sequence of actions $a_{1:t} \in \mathcal{A}_{1:t}$
A measurable function $f : \mathcal{X}_{0:t} \rightarrow \mathbb{R}$	A sequence of subsets $B_{0:t} \subseteq \mathcal{X}_{0:t}$.

In practice (and in our experiments), these parameters can be chosen using a held-out dataset via a sample splitting approach (Cox, 1975). To streamline notation, in this section, we will consider these parameters to be fixed. However, we emphasize that our construction can be instantiated for many different choices of these parameters, and indeed we will do so in our case study below. We think of f as expressing a specific outcome of interest at time t in terms of the data we have assumed. Accordingly, for each $a'_{1:t} \in \mathcal{A}_{1:t}$, we define new potential outcomes $Y(a'_{1:t}) := f(X_{0:t}(a'_{1:t}))$. For example, in a medical context, if $X_{0:t}(a'_{1:t})$ represents a full patient history at time t after treatments $a'_{1:t}$, then $Y(a'_{1:t})$ might represent the patient's heart rate after these treatments. Likewise, $B_{0:t}$ selects a subgroup of patients of interest, e.g. elderly patients whose blood pressure values were above some threshold at some timesteps before t .

For $a'_{1:t} \in \mathcal{A}_{1:t}$, we also denote by $\hat{Y}(a'_{1:t}) := f(X_0, \hat{X}_{1:t}(X_0, a'_{1:t}))$ the corresponding outcome of the twin. Supposing it holds that

$$\mathbb{P}(\hat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}) > 0, \tag{14}$$

we may then define $\hat{Q} := \mathbb{E}[\hat{Y}(a_{1:t}) \mid X_0 \in B_0, \hat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}]$, which is the analogue of Q for the twin. By Proposition 2, if the twin is interventionally correct, then $\hat{Q} = Q$. Theorem 6 therefore implies that the following hypotheses have our desired property (3):

$$\begin{aligned} \mathcal{H}_{\text{lo}}: & \text{ If (4), (5), and (14) hold, then } \hat{Q} \geq Q_{\text{lo}} \\ \mathcal{H}_{\text{up}}: & \text{ If (4), (5), and (14) hold, then } \hat{Q} \leq Q_{\text{up}}. \end{aligned}$$

Moreover, $\mathcal{H}_{\text{lo}}$ and $\mathcal{H}_{\text{up}}$ can in principle be determined to be true or false from the information we have assumed available, since Q_{lo} and Q_{up} depend only on the observational data, and $\hat{Q}$ can be estimated by generating trajectories from the twin.

When either $\mathcal{H}_{\text{lo}}$ or $\mathcal{H}_{\text{up}}$ is falsified, it immediately follows that the twin is not interventionally correct. However, even more than this, a falsification describes a concrete failure mode with various potential implications downstream, which is considerably more useful information about the twin in practice. For example, if $\mathcal{H}_{\text{lo}}$ is false (i.e. if $\hat{Q} < Q_{\text{lo}}$), it follows that, among those trajectories for which $(X_0, \hat{X}_{1:t}(X_0, a_{1:t})) \in B_{0:t}$, the mean of $\hat{Y}(a_{1:t})$ is too small. (Higher moments could also be considered by choosing f appropriately.) In light of this, a user might choose not to rely on outputs of the twin produced under these circumstances, while a developer seeking to improve the twin could focus their attention on the specific parts of its implementation that give rise to this behaviour.

5.2 Exact Testing Procedure

We now present a procedure for testing our hypotheses using a finite dataset that obtains exact control over type I error without relying on additional assumptions or asymptotic approximations. For simplicity, we specifically describe how to obtain a p-value for $\mathcal{H}_{\text{lo}}$ given a fixed choice of parameters $(t, f, a_{1:t}, B_{0:t})$. Our procedure for $\mathcal{H}_{\text{up}}$ is symmetrical, or may be regarded as a special case of testing $\mathcal{H}_{\text{lo}}$ by replacing f with $-f$. Multiple hypotheses may then be handled via standard techniques such as the method of Holm (1979) (which we use in our case study) or of Benjamini and Yekutieli (2001), both of which apply without additional assumptions.

Dataset. As above, we assume access to a finite dataset $\mathcal{D}$ of i.i.d. copies of (1). We also assume access to a dataset $\hat{\mathcal{D}}(a_{1:t})$ of i.i.d. copies of

$$X_0, \hat{X}_1(X_0, a_1), \dots, \hat{X}_t(X_0, a_{1:t}). \quad (15)$$

In practice, these copies can be obtained by initialising the twin with some value X_0 taken from $\mathcal{D}$ without replacement and supplying inputs $a_{1:t}$. If each X_0 in $\mathcal{D}$ is used to initialize the twin at most once, then the resulting trajectories in $\hat{\mathcal{D}}(a_{1:t})$ are guaranteed to be i.i.d., since we assumed in Section 2 that the potential outcomes $\hat{X}_t(x_0, a_{1:t})$ produced by the twin are independent across runs. We adopt this approach in our case study.

Antecedent conditions. Observe that $\mathcal{H}_{\text{lo}}$ is false only if (4), (5), and (14) all hold. We account for this in our testing procedure as follows. First, (5) immediately follows if

$$y_{\text{lo}} \leq f(x_{0:t}) \leq y_{\text{up}} \quad \text{for all } x_{0:t} \in B_{0:t}. \quad (16)$$

This holds automatically in certain cases, such as for binary outcomes (e.g. patient survival), or otherwise can be enforced simply by clipping the value of f to live within $[y_{\text{lo}}, y_{\text{up}}]$. We describe a practical means for choosing f in this way in Section 6.

To account for (4) and (14), we simply check whether there exists some trajectory in $\mathcal{D}$ such that $A_{1:t} = a_{1:t}$ and $X_{0:t}(A_{1:t}) \in B_{0:t}$, and some trajectory in $\hat{\mathcal{D}}(a_{1:t})$ such that $(X_0, \hat{X}_{1:t}(X_0, a_{1:t})) \in B_{0:t}$. If there are not, then we refuse to reject $\mathcal{H}_{\text{lo}}$ at any significance

level; otherwise, we proceed to test $\hat{Q} \geq Q_{\text{lo}}$ as described next. It now easily follows that there is zero probability we will reject $\mathcal{H}_{\text{lo}}$ if (4) and (14) do not in fact hold, and so our overall type I error is controlled at the desired level.

Overall testing procedure. To test $\hat{Q} \geq Q_{\text{lo}}$, we begin by constructing a one-sided lower confidence interval for Q_{lo} , and a one-sided upper confidence interval for $\hat{Q}$. In detail, for each significance level $\alpha \in (0, 1)$, we define random variables R_{lo}^α and $\hat{R}^\alpha$ in terms of $\mathcal{D}$ and $\hat{\mathcal{D}}(a_{1:t})$ such that

$$\mathbb{P}(Q_{\text{lo}} \geq R_{\text{lo}}^\alpha) \geq 1 - \frac{\alpha}{2} \quad \mathbb{P}(\hat{Q} \leq \hat{R}^\alpha) \geq 1 - \frac{\alpha}{2}. \quad (17)$$

We will also ensure that these are nested, i.e. $R_{\text{lo}}^\alpha \leq R_{\text{lo}}^{\alpha'}$ and $\hat{R}^{\alpha'} \leq \hat{R}^\alpha$ if $\alpha \leq \alpha'$. We describe two methods for obtaining R_{lo}^α and $\hat{R}^\alpha$ satisfying these conditions below.

From these confidence intervals, we obtain a test for the hypothesis $\hat{Q} \geq Q_{\text{lo}}$ that rejects when $\hat{R}^\alpha < R_{\text{lo}}^\alpha$. A straightforward argument given in Section H.1 of the Appendix shows that this controls the type I error at the desired level α . Nestedness also yields a p -value obtained as the smallest value of α for which this test rejects, i.e. $p_{\text{lo}} := \inf\{\alpha \in (0, 1) \mid \hat{R}^\alpha < R_{\text{lo}}^\alpha\}$, or 1 if the test does not reject at any level.

Obtaining confidence intervals. We now consider how to obtain confidence intervals for Q_{lo} and $\hat{Q}$ satisfying the desired conditions above. To this end, observe that both quantities are (conditional) expectations involving random variables that can be computed from $\mathcal{D}$ or $\hat{\mathcal{D}}(a_{1:t})$. This allows both to be estimated unbiasedly, which in turn can be used to derive confidence intervals via standard techniques. For example, consider the subset of trajectories in $\hat{\mathcal{D}}(a_{1:t})$ with $(X_0, \hat{X}_t(X_0, a_{1:t})) \in B_{0:t}$. For each such trajectory, we obtain a corresponding value of $\hat{Y}(a_{1:t})$ that is i.i.d. and has expectation $\hat{Q}$. Similarly, for Q_{lo} , we extract the subset of trajectories in $\mathcal{D}$ for which $X_{0:N}(A_{1:N}) \in B_{0:N}$ holds. The values of Y_{lo} obtained from each such trajectory are then i.i.d. and have expectation Q_{lo} .

At this point, our problem now reduces to that of constructing a confidence interval for the expectation of a random variable using i.i.d. copies of it. Various techniques exist for this, and we consider two possibilities in our case study. The first leverages the fact that $\hat{Y}(a_{1:t})$ and Y_{lo} are bounded in $[y_{\text{lo}}, y_{\text{up}}]$, which gives rise to R_{lo}^α and $\hat{R}^\alpha$ via an application of Hoeffding's inequality. This approach has the appealing property that (17) holds exactly, although at the expense of conservativeness. In practice, this could be mitigated by instead obtaining confidence intervals via (for example) the bootstrap (Efron, 1979), although at the expense of requiring (often mild) asymptotic assumptions. Section H.2 of the Appendix describes both methods in greater detail. Our empirical results reported in the next section all use Hoeffding's inequality, and are hence exact, but we also provide some results using bootstrapping in Section I.7 of the Appendix.

6. Case Study: Pulse Physiology Engine

We applied our assessment methodology to the Pulse Physiology Engine (Bray et al., 2019), an open-source model for human physiology simulation. Pulse simulates trajectories of various physiological metrics for patients with conditions like sepsis, COPD, and

ARDS. We describe the main steps of our experimental procedure and results below, with full details given in Section I of the Appendix. Code for replication is available at <https://github.com/faaizT/CausalTwinAssessment>.

6.1 Experimental Setup

Real-world data. We utilized the MIMIC-III dataset (Johnson et al., 2016), a comprehensive collection of longitudinal health data from critical care patients at the Beth Israel Deaconess Medical Center (2001-2012). We focused on patients meeting the sepsis-3 criteria (Singer et al., 2016), following the methodology of Komorowski et al. (2018) for selecting these. This yielded 11,677 sepsis patient trajectories. We randomly selected 5% of these (583 trajectories, denoted as $\mathcal{D}_0$) to use for choosing the parameters of our hypotheses via a sample splitting approach (Cox, 1975), with the remaining 95% (11,094 trajectories, denoted as $\mathcal{D}$) reserved for the actual testing.

We considered hourly observations of each patient over the first four hours of their ICU stay, i.e. $T = 4$. We defined the observation spaces $\mathcal{X}_{0:T}$ using a total of 17 features included in our extracted MIMIC trajectories for this time period, including static demographic quantities and patient vitals. Following Komorowski et al. (2018), the actions we considered involved the administration of intravenous fluids and vasopressors, which both play a primary role in the treatment of sepsis in clinical practice. Since these are recorded in MIMIC as continuous doses, we discretised their values via the same procedure as Komorowski et al. (2018), obtaining finite action spaces $\mathcal{A}_1 = \dots = \mathcal{A}_4$, each with 25 actions.

Hypothesis parameters. We defined a collection of hypothesis parameters $(t, f, a_{1:t}, B_{0:t})$, each of which we then used to define an $\mathcal{H}_{\text{lo}}$ and $\mathcal{H}_{\text{up}}$ to test. For this, we chose 14 different physiological quantities of interest to assess, including heart rate, skin temperature, and respiration rate (see Table 4 in the Appendix for a complete list). We used $\mathcal{D}_0$ to define a range of possible choices for each B_t via a binning procedure described in Section I.5 of the Appendix. We then selected all possible combinations of t , $a_{1:t}$, and $B_{0:t}$ that were observed for at least one patient trajectory in $\mathcal{D}_0$. We took y_{lo} and y_{up} to be the .2 and .8 quantiles of the same physiological quantity as was recorded in $\mathcal{D}_0$, and defined f as the function that extracts this quantity from $\mathcal{X}_t$ and clips its value between y_{lo} and y_{up} , so that (16) holds. We also investigated the sensitivity of our procedure to the choice of y_{lo} and y_{up} and found it to be relatively stable: see Section I.9 of the Appendix. We obtained 721 unique parameter choices, each of which produced two hypotheses $\mathcal{H}_{\text{lo}}$ and $\mathcal{H}_{\text{up}}$, leading to 1,442 hypotheses in total.

Twin trajectories. We generated data from Pulse to test the chosen hypotheses. For each $a_{1:t}$ occurring in any of our hypotheses, we obtained the dataset $\widehat{\mathcal{D}}(a_{1:t})$ as described in Section 5.2. Specifically, we sampled X_0 without replacement from $\mathcal{D}$, and used this to initialize a twin trajectory. Then, at each hour $t' \in \{1, \dots, 4\}$ in the simulation, we administered a dose of intravenous fluids and vasopressors corresponding to $a_{t'}$ and recorded the resulting patient features generated by Pulse. We describe this procedure in full in Section I.6 in the Appendix. This produced a total of 26,115 simulated trajectories.

Physiological quantity	Ours		Manski	
	Rejs.	Hyps.	Rejs.	Hyps.
Chloride Blood Concentration (Chloride)	24	94	1	46
Sodium Blood Concentration (Sodium)	21	94	9	46
Potassium Blood Concentration (Potassium)	13	94	0	46
Skin Temperature (Temp)	10	86	9	46
Calcium Blood Concentration (Calcium)	5	88	0	46
Glucose Blood Concentration (Glucose)	5	96	1	46
Arterial CO ₂ Pressure (paCO ₂)	3	70	0	46
Bicarbonate Blood Concentration (HCO ₃)	2	90	1	46
Systolic Arterial Pressure (SysBP)	2	154	0	46

Table 1: Total hypotheses (Hyps.) and rejections (Rejs.) per physiological quantity using our causal bounds, as well as those of Manski (1990)

6.2 Hypothesis Rejections

We tested the hypotheses just described using our methodology from Section 5.2. Here we report the results when using Hoeffding’s inequality to obtain confidence intervals for Q_{lo} , Q_{up} , and $\hat{Q}$. We also tried confidence intervals obtained via bootstrapping, and obtained similar if less conservative results (see Section I of the Appendix). We used the Holm-Bonferroni method to adjust for multiple tests, with family-wise error rate of 0.05.

We obtained rejections for hypotheses corresponding to 10 different physiological quantities shown in Table 1. (Table 4 in the Appendix shows all hypotheses we tested, including those not rejected.) We may therefore infer that, at a high level, Pulse does not simulate these quantities accurately for the population of sepsis patients we consider. This appears of interest in a variety of downstream settings: for example, a developer could use this information when considering how to improve the accuracy of Pulse, while a practitioner using Pulse may wish to rely less on these outputs as a result.

To assess the relative performance of our bounds from Theorem 6 compared with the unconditional bounds of Manski (1990), we also reran this analysis with each t , f , and $a_{1:t}$ chosen as before, but with each $B_{0:t}$ now set to the whole space $\mathcal{X}_{0:t}$. This in turn led to fewer hypotheses, namely 690 in total, which were evenly divided between hypotheses of the form $\mathbb{E}[\hat{Y}(a_{1:t})] \geq \mathbb{E}[Y_{\text{lo}}]$ and those of the form $\mathbb{E}[\hat{Y}(a_{1:t})] \leq \mathbb{E}[Y_{\text{up}}]$. We kept all other aspects of our procedure the same as just described, including our method for obtaining confidence intervals and controlling for multiple testing. The number of rejections that we obtained in this case is also shown in Table 1. As anticipated by our discussion in Section 4.3, the conservativeness of Manski’s bounds led to considerably fewer rejections than our more general result given in Theorem 6, even when considered as a proportion of the total hypotheses we tested.

6.3 p-Value Plots

To obtain more granular information about the failure modes of the twin just identified, we examined the p -values obtained for each hypothesis $\mathcal{H}_{\text{lo}}$ and $\mathcal{H}_{\text{up}}$ tested using our causal bounds, which we denote here by p_{lo} and p_{up} . Figure 1 shows the distributions of $-\log_{10} p_{\text{lo}}$ and $-\log_{10} p_{\text{up}}$ that we obtained for all physiological quantities for which some hypothesis

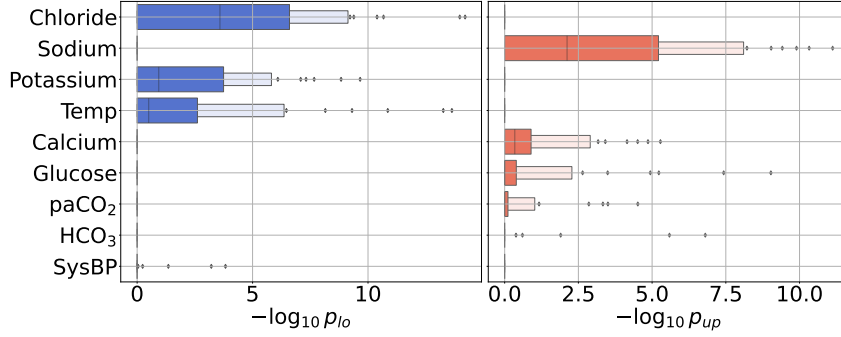


Figure 1: Distributions of $-\log_{10} p_{lo}$ and $-\log_{10} p_{up}$ across hypotheses, grouped by physiological quantity. Higher values indicate greater evidence in favour of rejection.

was rejected. (The remaining p -values are shown in Figure 6 in the Appendix.) Notably, in each row, one distribution is always tightly concentrated at $-\log_{10} p = 0$ (i.e. $p = 1$). This means that, for all physiological outcomes of interest, there was either very little evidence in favour of rejecting any $\mathcal{H}_{lo}$, or very little in favour of rejecting any $\mathcal{H}_{up}$. In other words, across configurations of $(t, f, a_{1:t}, B_{0:t})$ that were rejected, the twin consistently either underestimated or overestimated each quantity on average. For example, Pulse consistently underestimated chloride blood concentration and skin temperature, while it consistently overestimated sodium and glucose blood concentration levels. Like Table 1, this information appears of interest and actionable in a variety of downstream tasks.

6.4 Pitfalls of Naive Assessment

A naive approach to twin assessment involves simply comparing the output of the twin with the observational data directly, without accounting for causal considerations. We now show that, unlike our methodology, the results produced in this way can be potentially misleading. In Figure 2, for two different choices of $(a_{1:4}, B_{1:4})$, we plot estimates of $\hat{Q}_t$ and Q_t^{obs} for $t \in \{1, \dots, 4\}$, where

$$\begin{aligned}\hat{Q}_t &:= \mathbb{E}[\hat{Y}(a_{1:t}) \mid X_0 \in B_0, \hat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}] \\ Q_t^{obs} &:= \mathbb{E}[Y(A_{1:t}) \mid X_{0:t}(A_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t}].\end{aligned}$$

Here $\hat{Q}_t$ is just $\hat{Q}$ as defined above with its dependence on t made explicit. Each plot also shows one-sided 95% confidence intervals on Q_{lo} and Q_{up} at each $t \in \{1, \dots, 4\}$ obtained from Hoeffding’s inequality. Directly comparing the estimates of $\hat{Q}_t$ and Q_t^{obs} would suggest that the twin is comparatively more accurate for the right-hand plot, as these estimates are closer to one another in that case. However, the output of the twin in the right-hand plot is falsified at $t = 1$, as can be seen from the fact that confidence interval for $\hat{Q}_1$ lies entirely above the one-sided confidence interval for Q_{up} at that timestep. On the other hand, the output of the twin in the left-hand plot is not falsified at any of the timesteps shown, so that the twin may in fact be accurate for these $(a_{1:4}, B_{1:4})$, contrary to what a naive assessment strategy would suggest. Our methodology provides a principled means for twin assessment that avoids drawing potentially misleading inferences like this.

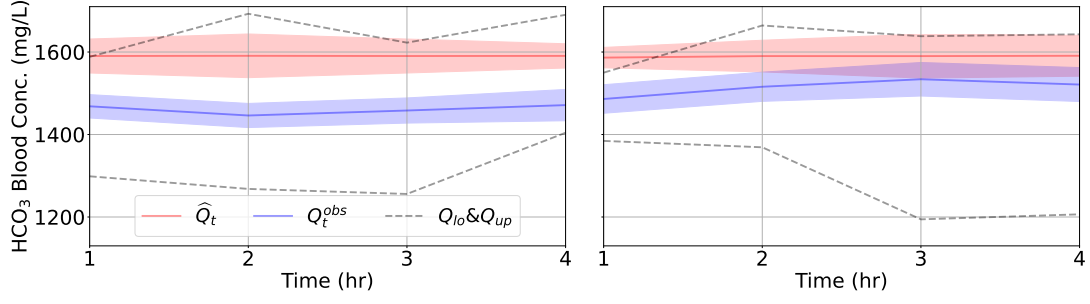


Figure 2: Estimates and 95% confidence intervals for $\hat{Q}_t$ and Q_t^{obs} at each $1 \leq t \leq 4$ for two choices of $(B_{0:4}, a_{1:4})$, where $\hat{Y}(a_{1:t})$ and $Y(a_{1:t})$ correspond to HCO₃ concentration. The dashed lines indicate lower and upper 95% confidence intervals for Q_{lo}, Q_{up} respectively.

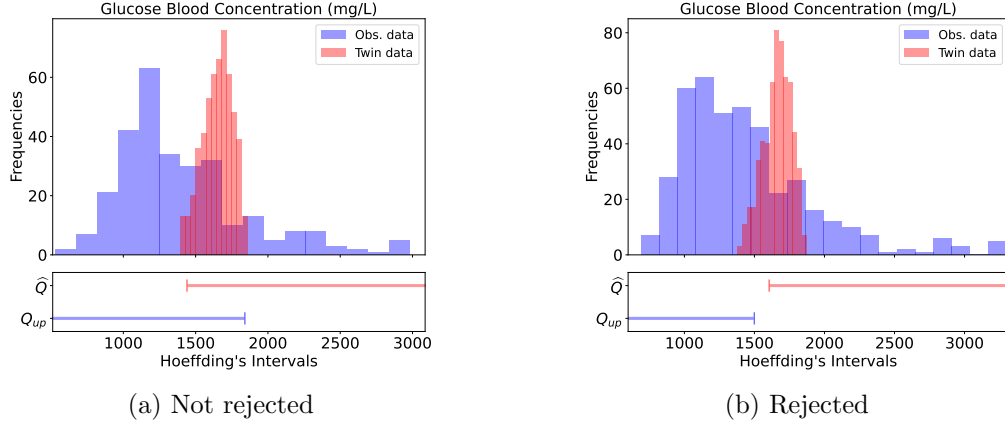


Figure 3: Raw glucose values from the observational data and twin for two choices of $(B_{0:t}, a_{1:t})$, with confidence intervals for $\hat{Q}$ and Q_{up} shown below. The horizontal axes are truncated to the .025 and .975 quantiles of the observational data for clarity. Untruncated plots are shown in Figure 7 of the Appendix.

A similar phenomenon appears in Figure 3, which for two choices of $B_{0:t}$ and $a_{1:t}$ shows histograms of raw glucose values obtained from the observational data conditional on $A_{1:t} = a_{1:t}$ and $X_{0:t}(A_{1:t}) \in B_{0:t}$, and from the twin conditional on $\hat{X}_{0:t}(a_{1:t}) \in B_{0:t}$. Below each histogram we also show 95% confidence intervals for Q_{up} and $\hat{Q}$ obtained from Hoeffding's inequality. While Figures 3a and 3b appear visually very similar, the inferences produced by our testing procedure are different: the hypothesis corresponding to the right-hand plot is rejected, since there is no overlap between the confidence intervals underneath, while the hypothesis corresponding to the left-hand plot is not. This was not an isolated case and several other examples of this phenomenon are shown in Figure 7 in the Appendix. This demonstrates that the inferences obtained from our procedure do not depend only on the distribution of observed outcomes (which is essentially the same for both cases). Instead, as discussed in Section 4.2, these also account for the worst-case effects of unmeasured confounding that may exist in the observational data.

7. Discussion

We have advocated for a causal approach to digital twin assessment, and have presented a statistical procedure for doing so that obtains rigorous theoretical guarantees under minimal assumptions. We now highlight the key limitations of our approach. Importantly, our methodology implicitly assumes that there is no distribution shift between testing and deployment time. If the conditional distribution of $X_{1:T}(a_{1:T})$ given X_0 changes at deployment time, then so too does the set of twins that are interventionally correct, and if this change is significant enough, our assessment procedure may yield misleading results. Distribution shift in this sense is a separate issue to unobserved confounding, and arises in a wide variety of statistical problems beyond ours.

Additionally, the procedure we used in our case study to choose the hypothesis parameters $B_{0:t}$ (described in Section I.5 of the Appendix) was ad hoc. For scalability, it would likely be necessary to obtain $B_{0:t}$ via a more automated procedure. It may also be desirable to choose $B_{0:t}$ dynamically in light of previous hypotheses tested, zooming in to regions containing possible failure modes to obtain increasingly granular information about the twin. We see opportunities here for using machine learning techniques, but leave this to future work.

Various other extensions and improvements also appear possible. For example, one can leverage ideas from the literature on partial identification (Manski, 2003) to obtain greater statistical efficiency, for example by building on the line of work initiated by Imbens and Manski (2004) for obtaining more informative confidence intervals. Beyond this, it may sometimes be useful to consider additional assumptions that lead to less conservative assessment results. For example, various methods for *sensitivity analysis* have been proposed that model the *degree* to which the actions of the behavioural agent are confounded (Rosenbaum, 2002; Tan, 2006; Yadlowsky et al., 2022). This can yield tighter bounds on Q than Theorem 6, albeit at the expense of less robustness if these assumptions are violated.

Acknowledgments

The authors are highly appreciative of troubleshooting and development assistance provided by the Pulse team. RC was supported by the National Research Foundation, Singapore, under a National Research Foundation Fellowship in Artificial Intelligence (Award No.: NRFFIAI1-2024-0014). In addition, RC, AD, and CH were supported by the Engineering and Physical Sciences Research Council (EPSRC) through the Bayes4Health programme [Grant number EP/R018561/1]. MFT was funded by Google DeepMind. The authors declare there are no competing interests.

Appendix A. Notation

$z_{t:t'}$	The sequence of elements $(z_t, \dots, z_{t'})$ (or the empty sequence when $t > t'$)
$\mathcal{Z}_{t:t'}$ (where each $\mathcal{Z}_i$ is a set)	The cartesian product $\mathcal{Z}_t \times \dots \times \mathcal{Z}_{t'}$ (or the set containing just the empty sequence when $t > t'$)
$Z_{t:t'}(a_{1:t'})$	The sequence of potential outcomes $Z_t(a_{1:t}), \dots, Z_{t'}(a_{1:t'})$ (or the empty sequence when $t > t'$)
$\text{Law}[Z]$	The distribution of the random variable Z
$\text{Law}[Z \mid M]$	The conditional distribution of Z given M , where M is either an event or a random variable
$Z \stackrel{\text{a.s.}}{=} Z'$	The random variables Z and Z' are almost surely equal, i.e. $\mathbb{P}(Z = Z') = 1$
$Z \perp\!\!\!\perp Z'$	The random variables Z and Z' are independent
$Z \perp\!\!\!\perp Z' \mid Z''$	The random variables Z and Z' are conditionally independent given the random variable Z''
$\mathbb{1}(E)$	Indicator function of some event E

Appendix B. Proof of Proposition 2 (Unconditional Form of Interventional Correctness)

Proof Fix any choice of $a_{1:T} \in \mathcal{A}_{1:T}$. By disintegrating $\text{Law}[X_0, \widehat{X}_{1:T}(X_0, a_{1:T})]$ and $\text{Law}[X_{0:T}(a_{1:T})]$ along their common $\mathcal{X}_0$ -marginal (which is namely $\text{Law}[X_0]$), it holds that

$$\text{Law}[X_0, \widehat{X}_{1:T}(X_0, a_{1:T})] = \text{Law}[X_{0:T}(a_{1:T})] \quad (18)$$

if and only if

$$\text{Law}[\widehat{X}_{1:T}(X_0, a_{1:T}) \mid X_0 = x_0] = \text{Law}[X_{1:T}(a_{1:T}) \mid X_0 = x_0] \quad (19)$$

for $\text{Law}[X_0]$ -almost all $x_0 \in \mathcal{X}_0$. But now, our definition of $\widehat{X}_{1:T}(x_0, a_{1:T})$ in terms of h_t and $U_{1:t}$ means we can write $\widehat{X}_{1:T}(X_0, a_{1:T}) = \mathbf{h}(X_0, a_{1:T}, U_{1:T})$, where

$$\mathbf{h}(x_0, a_{1:T}, u_{1:T}) := (h_1(x_0, a_1, u_1), \dots, h_T(x_0, a_{1:T}, u_{1:T})).$$

For all $x_0 \in \mathcal{X}_0$ and measurable $B_{1:T} \subseteq \mathcal{X}_{1:T}$, we then have

$$\begin{aligned} \text{Law}[\widehat{X}_{1:T}(x_0, a_{1:T})](B_{1:T}) &= \mathbb{E}[\mathbb{1}(\mathbf{h}(x_0, a_{1:T}, U_{1:T}) \in B_{1:T})] \\ &= \int \mathbb{1}(\mathbf{h}(x_0, a_{1:T}, u_{1:T}) \in B_{1:T}) \text{Law}[U_{1:T}](du_{1:T}). \end{aligned}$$

It is standard to show that the right-hand side is a Markov kernel in x_0 and $B_{1:T}$. Moreover, for any measurable $B_0 \subseteq \mathcal{X}_0$, we have

$$\begin{aligned}
 & \int_{B_0} \text{Law}[\widehat{X}_{1:T}(x_0, a_{1:T})](B_{1:T}) \text{Law}[X_0](dx_0) \\
 &= \int_{B_0} \left[\int \mathbb{1}(\mathbf{h}(x_0, a_{1:T}, u_{1:T}) \in B_{1:T}) \text{Law}[U_{1:T}](du_{1:T}) \right] \text{Law}[X_0](dx_0) \\
 &= \int \mathbb{1}(x_0 \in B_0, \mathbf{h}(x_0, a_{1:T}, u_{1:T}) \in B_{1:T}) \text{Law}[X_0, U_{1:T}](dx_0, du_{1:T}) \\
 &= \text{Law}[X_0, \widehat{X}_{1:T}(X_0, a_{1:T})](B_{0:T}),
 \end{aligned}$$

where the second step follows because $X_0 \perp\!\!\!\perp U_{1:T}$. It therefore follows that $(x_0, B_{1:T}) \mapsto \text{Law}[\widehat{X}_{1:T}(x_0, a_{1:T})](B_{1:T})$ is a regular conditional distribution of $\widehat{X}_{1:T}(X_0, a_{1:T})$ given X_0 , i.e.

$$\text{Law}[\widehat{X}_{1:T}(x_0, a_{1:T})] = \text{Law}[\widehat{X}_{1:T}(X_0, a_{1:T}) \mid X_0 = x_0] \quad \text{for Law}[X_0]\text{-almost all } x_0 \in \mathcal{X}_0.$$

Substituting this into (19), we see that (18) holds if and only if

$$\text{Law}[\widehat{X}_{1:T}(x_0, a_{1:T})] = \text{Law}[X_{1:T}(a_{1:T}) \mid X_0 = x_0]$$

for Law $[X_0]$ -almost all $x_0 \in \mathcal{X}_0$. The result now follows since $a_{1:T}$ was arbitrary. $\blacksquare$

Appendix C. Online Prediction

C.1 Correctness in the Online Setting

A distinguishing feature of many digital twins is their ability to integrate real-time information obtained from sensors in their environment (Barricelli et al., 2019). It is therefore relevant to consider a setting in which a twin is used repeatedly to make a sequence of predictions over time, each time taking all previous information into account. One way to formalize this is to instantiate our model for the twin at each timestep. For example, we could represent the predictions made by the twin at $t = 0$ after observing initial covariates x_0 as potential outcomes $(\widehat{X}_{1:T}^1(x_0, a_{1:T}) : a_{1:T} \in \mathcal{A}_{1:T})$, similar to what we did in the main text. We could then represent the predictions made by the twin after some action a_1 is taken and an additional observation x_1 is made via potential outcomes $(\widehat{X}_{2:T}^2(x_{0:1}, a_{1:T}) : a_{2:T} \in \mathcal{A}_{2:T})$. More generally, for $t \in \{1, \dots, T\}$, we could introduce potential outcomes $(\widehat{X}_{t:T}^t(x_{0:t-1}, a_{1:T}) : a_{t:T} \in \mathcal{A}_{t:T})$ to represent the predictions that the twin would make at time t after the observations $x_{0:t-1}$ are made and the actions $a_{1:t-1}$ are taken.

This extended model requires a new definition of correctness than our Definition 1 from the main text. A natural approach is to say that the twin is correct in this new setting if

$$\text{Law}[\widehat{X}_{t:T}^t(x_{0:t-1}, a_{1:T})] = \text{Law}[X_{t:T}(a_{1:T}) \mid X_{0:t-1}(a_{1:t-1}) = x_{0:t-1}] \quad (20)$$

for all $t \in \{1, \dots, T\}$, $a_{1:T} \in \mathcal{A}_{1:T}$, and Law $[X_{0:t-1}(a_{1:t-1})]$ -almost all $x_{0:t-1} \in \mathcal{X}_{0:t-1}$. A twin with this property would at each step be able to accurately simulate the future in light

of previous information, use this to choose a next action to take, observe the result of doing so, and then repeat. It is possible to show that (20) holds if and only if we have

$$\begin{aligned} \text{Law}[\hat{X}_{1:T}^1(x_0, a_{1:T})] &= \text{Law}[X_{1:T}(a_{1:T}) \mid X_0 = x_0] \\ \text{Law}[\hat{X}_{t:T}^t(x_{0:t-1}, a_{1:T})] &= \text{Law}[\hat{X}_{t:T}^1(x_0, a_{1:T}) \mid \hat{X}_{1:t-1}^1(x_0, a_{1:t-1}) = x_{1:t-1}] \end{aligned}$$

for all $t \in \{1, \dots, T\}$, $a_{1:T} \in \mathcal{A}_{1:T}$, $\text{Law}[X_0]$ -almost all $x_0 \in \mathcal{X}_0$, and $\text{Law}[\hat{X}_{1:t-1}^1(x_0, a_{1:t-1})]$ -almost all $x_{1:t-1} \in \mathcal{X}_{1:t-1}$. The first condition here says that $\hat{X}_{1:T}^1(x_0, a_{1:T})$ must be interventionally correct in the sense of Definition 1 from the main text. The second condition says that the predictions made by the twin across different timesteps must be internally consistent with each other insofar as their conditional distributions must align. This holds automatically in many circumstances, such as if the predictions of the twin are obtained from a Bayesian model (for example), and otherwise could be checked numerically given the ability to run simulations from the twin, without the need to obtain data or refer to the real-world process in any way. As such, the problem of assessing the correctness of the twin in this new sense primarily reduces to the problem of assessing the correctness of $\hat{X}_{1:T}^1(x_0, a_{1:T})$ in the sense of Definition 1 in the main text, which motivates our focus on that condition.

C.2 Alternative Notions of Online Correctness

An important and interesting subtlety arises in this context that is worth noting. In general it does not follow that a twin correct in the sense of (20) satisfies

$$\text{Law}[\hat{X}_{t:T}^t(x_{0:t-1}, a_{1:T})] = \text{Law}[X_{t:T}(a_{1:T}) \mid X_{0:t-1}(a_{1:t-1}) = x_{0:t-1}, A_{1:t-1} = a_{1:t-1}] \quad (21)$$

for all $a_{1:T} \in \mathcal{A}_{1:T}$, and $\text{Law}[X_{0:t-1}(a_{1:t-1}) \mid A_{1:t-1} = a_{1:t-1}]$ -almost all $x_{0:t-1} \in \mathcal{X}_{0:t-1}$, since in general it does not hold that

$$\begin{aligned} \text{Law}[X_{t:T}(a_{1:T}) \mid X_{0:t-1}(a_{1:t-1}) = x_{0:t-1}] \\ = \text{Law}[X_{t:T}(a_{1:T}) \mid X_{0:t-1}(a_{1:t-1}) = x_{0:t-1}, A_{1:t-1} = a_{1:t-1}]. \end{aligned}$$

for all $a_{1:T} \in \mathcal{A}_{1:T}$ and $\text{Law}[X_{0:t-1}(a_{1:t-1}) \mid A_{1:t-1} = a_{1:t-1}]$ -almost all $x_{0:t-1} \in \mathcal{X}_{0:t-1}$ unless the actions $A_{1:t-1}$ are unconfounded. (Here as usual $A_{1:T}$ denotes the actions of a behavioural agent; see Section 3 of the main text.) In other words, a twin that is correct in the sense of (20) will make accurate predictions at time t when every action taken before time t was unconfounded (as occurs for example when the twin is directly in control of the decision-making process), but in general not when certain taken actions before time t were chosen by a behavioural agent with access to more context than is available to the twin. However, should it be desirable, our framework could be extended to encompass the alternative condition in (21) by relabelling the observed history $(X_{0:t-1}(A_{1:t-1}), A_{1:t-1})$ as X_0 , and then assessing the correctness of the potential outcomes $\hat{X}_{t:T}^t(x_{0:t-1}, a_{1:T})$ in the sense of Definition 1 from the main text.

Overall, the “right” notion of correctness in this online setting is to some extent a design choice. We believe our causal approach to twin assessment provides a useful framework for formulating and reasoning about these possibilities, and consider the investigation of assessment strategies for additional usage regimes to be an interesting direction for future work.

Appendix D. Proof of Theorem 3 (Interventional Distributions Are Not Identifiable)

It is well-known in the causal inference literature that the interventional behaviour of the real-world process cannot be uniquely identified from observational data. For completeness, we now provide a self-contained proof of this result in our notation. Our statement here is lengthier than Theorem 3 in the main text in order to clarify what is meant by “uniquely identified”: intuitively, the idea is that there always exist distinct families of potential outcomes whose interventional behaviours differ and yet give rise to the same observational data.

Theorem 10 *Suppose we have $a_{1:T} \in \mathcal{A}_{1:T}$ such that $\mathbb{P}(A_{1:T} \neq a_{1:T}) > 0$. Then there exist potential outcomes $(\tilde{X}_{0:T}(a'_{1:T}) : a'_{1:T} \in \mathcal{A}_{1:T})$ such that*

$$(\tilde{X}_{0:T}(A_{1:T}), A_{1:T}) \stackrel{\text{a.s.}}{=} (X_{0:T}(A_{1:T}), A_{1:T}). \quad (22)$$

but for which $\text{Law}[\tilde{X}_{0:T}(a_{1:t})] \neq \text{Law}[X_{0:T}(a_{1:t})]$.

Proof Our assumption that $\mathbb{P}(A_{1:T} \neq a_{1:T}) > 0$ means there must exist some $t \in \{1, \dots, T\}$ such that $\mathbb{P}(A_{1:t} \neq a_{1:t}) > 0$. Since $\mathcal{X}_t = \mathbb{R}^{d_t}$, we may also choose some $x_t \in \mathcal{X}_t$ with $\mathbb{P}(X_t(a_{1:t}) = x_t \mid A_{1:t} \neq a_{1:t}) \neq 1$. Then, for each $s \in \{0, \dots, T\}$ and $a'_{1:s} \in \mathcal{A}_{1:s}$, define

$$\tilde{X}_s(a'_{1:s}) := \begin{cases} \mathbb{1}(A_{1:t} = a_{1:t}) X_t(a_{1:t}) + \mathbb{1}(A_{1:t} \neq a_{1:t}) x_t & \text{if } s = t \text{ and } a'_{1:s} = a_{1:t} \\ X_s(a'_{1:s}) & \text{otherwise,} \end{cases}$$

It is then easily checked that (22) holds, but

$$\begin{aligned} & \text{Law}[\tilde{X}_t(a_{1:t})] \\ &= \text{Law}[\tilde{X}_t(a_{1:t}) \mid A_{1:t} = a_{1:t}] \mathbb{P}(A_{1:t} = a_{1:t}) + \text{Law}[\tilde{X}_t(a_{1:t}) \mid A_{1:t} \neq a_{1:t}] \mathbb{P}(A_{1:t} \neq a_{1:t}) \\ &= \text{Law}[X_t(a_{1:t}) \mid A_{1:t} = a_{1:t}] \mathbb{P}(A_{1:t} = a_{1:t}) + \text{Dirac}(x_t) \mathbb{P}(A_{1:t} \neq a_{1:t}) \\ &\neq \text{Law}[X_t(a_{1:t}) \mid A_{1:t} = a_{1:t}] \mathbb{P}(A_{1:t} = a_{1:t}) + \text{Law}[X_t(a_{1:t}) \mid A_{1:t} \neq a_{1:t}] \mathbb{P}(A_{1:t} \neq a_{1:t}) \\ &= \text{Law}[X_t(a_{1:t})], \end{aligned}$$

from which the result follows. ■

Appendix E. Deterministic Potential Outcomes Are Unconfounded

In this section we expand on our earlier claim that, if the real-world process is deterministic, then the observational data is unconfounded. We first make this claim precise. By “deterministic”, we mean that there exist measurable functions g_t for $t \in \{1, \dots, T\}$ such that

$$X_t(a_{1:t}) \stackrel{\text{a.s.}}{=} g_t(X_{0:t-1}(a_{1:t-1}), a_{1:t}) \quad \text{for all } t \in \{1, \dots, T\} \text{ and } a_{1:t} \in \mathcal{A}_{1:t}. \quad (23)$$

By “unconfounded”, we mean that the *sequential randomisation assumption (SRA)* introduced by Robins (1986) holds, i.e.

$$(X_s(a_{1:s}) : s \in \{1, \dots, T\}, a_{1:s} \in \mathcal{A}_{1:s}) \perp\!\!\!\perp A_t \mid X_{0:t-1}(A_{1:t-1}), A_{1:t-1} \quad \text{for all } t \in \{1, \dots, T\}, \quad (24)$$

where $\perp\!\!\!\perp$ denotes conditional independence. Intuitively, this says that, apart from the historical observations $(X_{0:t-1}(A_{1:t-1}), A_{1:t-1})$, any additional factors that influence the agent’s choice of action A_t are independent of the behaviour of the real-world process. The SRA provides a standard formulation of the notion of unconfoundedness in longitudinal settings such as ours (see Tsiatis et al. 2019, Chapter 5 for a review).

It is now a standard exercise to show that (23) implies (24). We include a proof below for completeness. Key to this is the following straightforward Lemma.

Lemma 11 *Suppose U and V are random variables such that, for some measurable function g , it holds that $U \stackrel{\text{a.s.}}{=} g(V)$. Then, for any other random variable W , we have*

$$U \perp\!\!\!\perp W \mid V.$$

Proof By standard properties of conditional expectations, for any measurable sets S_1 and S_2 , we have almost surely

$$\begin{aligned} \mathbb{P}(U \in S_1, W \in S_2 \mid V) &= \mathbb{E}[\mathbb{1}(g(V) \in S_1) \mathbb{1}(W \in S_2) \mid V] \\ &= \mathbb{1}(g(V) \in S_1) \mathbb{E}[\mathbb{1}(W \in S_2) \mid V] \\ &= \mathbb{E}[\mathbb{1}(U \in S_1) \mid V] \mathbb{P}(W \in S_2 \mid V) \\ &= \mathbb{P}(U \in S_1 \mid V) \mathbb{P}(W \in S_2 \mid V), \end{aligned}$$

which gives the result. ■

It is now easy to see that (23) implies (24). Indeed, by recursive substitution, it is straightforward to show that there exist measurable functions $\tilde{g}_t$ for $t \in \{1, \dots, T\}$ such that

$$X_t(a_{1:t}) \stackrel{\text{a.s.}}{=} \tilde{g}_t(X_0, a_{1:t}) \quad \text{for all } t \in \{1, \dots, T\} \text{ and } a_{1:t} \in \mathcal{A}_{1:t},$$

and so

$$(X_s(a_{1:s}) : s \in \{1, \dots, T\}, a_{1:s} \in \mathcal{A}_{1:s}) = (\tilde{g}_t(X_0, a_{1:s}) : s \in \{1, \dots, T\}, a_{1:s} \in \mathcal{A}_{1:s}).$$

The right-hand side is now seen to be a measurable function of X_0 and hence certainly of $(X_{0:t-1}(A_{1:t-1}), A_{1:t-1})$, so that the result follows by Lemma 11.

Appendix F. Motivating Toy Example

We provide here a toy scenario that illustrates intuitively the pitfalls that may arise when assessing twins using observational data without properly accounting for unmeasured confounding. Suppose a digital twin has been designed for a particular make of car, e.g. to facilitate autonomous driving (Allamaa et al., 2022). The twin simulates how quantities

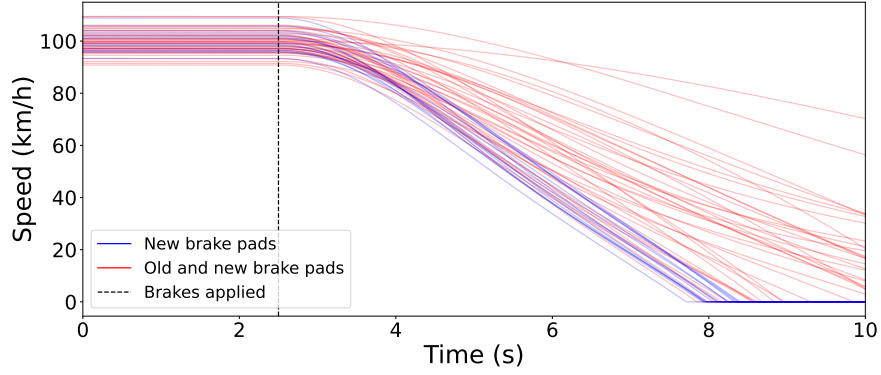


Figure 4: The discrepancy between observational data and interventional behavior. The data only show the effect of aggressive braking on cars with new brake pads (blue). This differs from what *would* be observed if aggressive braking were applied to the entire fleet of cars, encompassing both those with old and new brake pads (red).

such as the velocity and fuel consumption of the car respond as certain inputs are applied to it, such as braking, acceleration, steering, etc. We wish to assess the accuracy of this twin using a dataset obtained from a fleet of the same make. The braking performance of these vehicles is significantly affected by the age of their brake pads: if these are fairly new, then an aggressive braking strategy will stop the car, while if these are old, then the same aggressive strategy will send the car into a skid that will reduce braking efficacy. Brake pad age is not recorded in the data we have obtained, but *was* known to the drivers who operated these vehicles (e.g. perhaps they were aware of how recently their car was serviced), and so the drivers of cars with old brake pads tended to avoid braking aggressively out of safety concerns.

A naive approach to twin assessment in this situation would directly compare the outputs of the twin with the data and conclude the twin is accurate if these match closely. However, in this scenario, the data contains a spurious relationship between braking strategy and the performance of the car: since aggressive braking is only observed for cars with new brake pads, the data appears to show that aggressive braking is effective at stopping the car, while in fact this is not the case for cars with older brake pads. As such, the naive assessment approach would yield misleading information about the twin: a twin that captures only the behaviour of cars with newer brake pads would appear to be correct, while a twin that captures the full range of possibilities (i.e. regardless of brake pad age) would deviate from the observational data and appear therefore less accurate. Figure 4 illustrates this pictorially under a toy model for this scenario.

More generally, whenever an *unmeasured confounder* such as brake pad age is present, a potential discrepancy arises between how the real-world process was observed to behave in the dataset and how it *would* behave under certain interventions. An obvious approach towards mitigating this possibility is to measure additional quantities that may affect the outcome of interest. For example, if brake pad age were included in the data in the scenario above, then it would be possible to adjust for its effect on braking performance. However, in many cases, gathering additional data may be costly or impractical. Moreover, even

if this strategy is pursued, it is rarely possible to rule out the possibility of unmeasured confounding altogether, especially for complicated real-world problems (Tsiatis et al., 2019). For example, in the scenario above, it is very conceivable that some other factor such as weather conditions could play a similar confounding role as brake pad age, and so would need also to be included in the data, and so on. Analogous scenarios are also easily forthcoming for other application domains such as medicine and economics (Manski, 1995; Tsiatis et al., 2019; Hernán and Robins, 2020). As such, rather than attempting to sidestep the issue of unmeasured confounding, we instead propose a methodology for assessing twins using data that is robust to its presence.

Appendix G. Causal Bounds

G.1 Proof of Lemma 5

Proof We prove the lower bound, i.e. (8); the upper bound is analogous. To streamline notation, we will denote

$$\begin{aligned} p &:= \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) \\ q &:= \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}). \end{aligned}$$

Both quantities are always well-defined since we have

$$0 < \mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t}) \leq \mathbb{P}(X_{0:N}(A_{1:N}) \in B_{0:N}),$$

where the first inequality holds by assumption, and the second since $N \leq t$.

Now, from the definition of Y_{lo} , we can write the left-hand side of (8) as follows:

$$\begin{aligned} \mathbb{E}[Y_{lo} \mid X_{0:t}(a_{1:t}) \in B_{0:t}] &= p \mathbb{E}[Y(A_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t}] + (1 - p) y_{lo} \\ &= p \mathbb{E}[Y(A_{1:t}) \mid X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t}] + (1 - p) y_{lo}. \end{aligned}$$

Here the first step uses the law of total probability, and the second step uses the fact that $N = t$ on the event $\{A_{1:t} = a_{1:t}\}$, since $N = \max\{0 \leq s \leq t \mid A_{1:s} = a_{1:s}\}$. Similarly, the right-hand side of (8) can be written as

$$\mathbb{E}[Y_{lo} \mid X_{0:N}(A_{1:N}) \in B_{0:N}] = q \mathbb{E}[Y(A_{1:t}) \mid X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t}] + (1 - q) y_{lo},$$

again by the law of total probability.

Now, suppose we can show that $p \geq q$. It is then a general fact about convex combinations that for any real numbers $u \geq v$, we have

$$pu + (1 - p)v \geq qu + (1 - q)v.$$

Intuitively, moving from the left to the right-hand side shifts mass from the larger value u to the smaller value v , which can only decrease the value of the convex combination overall. Applying this with $u := \mathbb{E}[Y(A_{1:t}) \mid X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t}]$ and $v := y_{lo}$ then gives the result (8), since our assumption $\mathbb{P}(y_{lo} \leq Y(a_{1:t}) \leq y_{up} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) = 1$ straightforwardly implies that $u \geq v$.

We therefore wish to show that $q \leq p$ indeed holds. To this end, observe the following:

$$\begin{aligned}
 \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) &= \frac{\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t})}{\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t})} \\
 &\geq \frac{\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t})}{\mathbb{P}(X_{0:N}(a_{1:N}) \in B_{0:N})} \\
 &= \frac{\mathbb{P}(X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t})}{\mathbb{P}(X_{0:N}(a_{1:N}) \in B_{0:N})} \\
 &= \frac{\mathbb{P}(X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t})}{\mathbb{P}(X_{0:N}(A_{1:N}) \in B_{0:N})} \\
 &= \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}).
 \end{aligned}$$

Here the second step follows because $N \leq t$, and hence

$$\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t}) \leq \mathbb{P}(X_{0:N}(a_{1:N}) \in B_{0:N}).$$

The third step holds because on the event $\{A_{1:t} = a_{1:t}\}$, we have $N = t$ and hence $X_{0:t}(a_{1:t}) = X_{0:N}(A_{1:N})$. The fourth step then follows because by definition of $N = \max\{0 \leq s \leq t \mid A_{1:s} = a_{1:s}\}$, we have $a_{1:N} = A_{1:N}$ and hence $X_{0:N}(a_{1:N}) = X_{0:N}(A_{1:N})$. The result now follows. $\blacksquare$

G.2 Proof of Proposition 7

Proof From the definition of Y_{up} , we have straightforwardly

$$\begin{aligned}
 Q_{\text{up}} &= \mathbb{E}[Y(A_{1:t}) \mid X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t}] \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}) \\
 &\quad + y_{\text{up}} \mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}).
 \end{aligned}$$

A similar expression holds for Q_{lo} . Subtracting these two expressions yields

$$Q_{\text{up}} - Q_{\text{lo}} = (y_{\text{up}} - y_{\text{lo}}) (1 - \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})).$$

Similar manipulations show that

$$\mathbb{E}[Y_{\text{up}} \mid X_0 \in B_0] - \mathbb{E}[Y_{\text{lo}} \mid X_0 \in B_0] = (y_{\text{up}} - y_{\text{lo}}) (1 - \mathbb{P}(A_{1:t} = a_{1:t} \mid X_0 \in B_0)),$$

and the result now follows. $\blacksquare$

G.3 Proof of Proposition 8

Proof We consider the case of the lower bound; the case of the upper bound is analogous. Choose $x_{1:T} \in B_{1:T}$ arbitrarily. (Certainly some choice is always possible, since each B_s has positive measure and is therefore nonempty.) Define

$$\begin{aligned}
 \tilde{X}_0 &:= X_0 \\
 \tilde{X}_s(a'_{1:s}) &:= \mathbb{1}(A_{1:s} = a'_{1:s}) X_s(a'_{1:s}) + \mathbb{1}(A_{1:s} \neq a'_{1:s}) x_s
 \end{aligned}$$

for each $s \in \{0, \dots, T\}$ and $a'_{1:s} \in \mathcal{A}_{1:s}$, and similarly let

$$\tilde{Y}(a'_{1:t}) = \mathbb{1}(A_{1:t} = a'_{1:t}) Y(a'_{1:t}) + \mathbb{1}(A_{1:t} \neq a'_{1:t}) y_{\text{lo}} \quad \text{for all } a'_{1:t} \in \mathcal{A}_{1:t}.$$

It is easy to check that $(\tilde{X}_{0:T}(A_{1:T}), \tilde{Y}(A_{1:T}), A_{1:T}) \stackrel{\text{a.s.}}{=} (X_{0:T}(A_{1:T}), Y(A_{1:T}), A_{1:T})$. But now we have directly $\tilde{Y}(a_{1:t}) = Y_{\text{lo}}$. Moreover, it is easily checked from the definition of N and $\tilde{X}_{0:t}(a_{1:t})$ that

$$\tilde{X}_{0:t}(a_{1:t}) \stackrel{\text{a.s.}}{=} (X_{0:N}(A_{1:N}), x_{N+1:t}),$$

so that

$$\begin{aligned} \mathbb{1}(\tilde{X}_{0:t}(a_{1:t}) \in B_{0:t}) &\stackrel{\text{a.s.}}{=} \mathbb{1}(\tilde{X}_{0:N}(a_{1:N}) \in B_{0:N}, x_{N+1:t} \in B_{N+1:t}) \\ &\stackrel{\text{a.s.}}{=} \mathbb{1}(X_{0:N}(A_{1:N}) \in B_{0:N}) \end{aligned}$$

since each $x_s \in B_s$. Consequently,

$$\begin{aligned} \mathbb{E}[\tilde{Y}(a_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t}) \in B_{0:t}] &= \mathbb{E}[Y_{\text{lo}} \mid \tilde{X}_{0:t}(a_{1:t}) \in B_{0:t}] \\ &= \mathbb{E}[Y_{\text{lo}} \mid X_{0:N}(A_{1:N}) \in B_{0:N}], \end{aligned}$$

which gives the result. ■

G.4 Bounds on the Conditional Expectation Given Specific Covariate Values

Theorem 6 provides a bound on $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}]$, i.e. the conditional expectation given the *event* $\{X_{0:t}(a_{1:t}) \in B_{0:t}\}$, which is assumed to have positive probability. We consider here the prospect of obtaining bounds on $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})]$, i.e. the conditional expectation given the *value* of $X_{0:t}(a_{1:t})$. For falsification purposes, this would provide a means for determining that twin is incorrect when it outputs specific values of $\hat{X}_{0:t}(a_{1:t})$, rather than just that it is incorrect on average across all runs that output values $\hat{X}_{0:t}(a_{1:t}) \in B_{0:t}$.

When $X_{0:t}(a_{1:t})$ is discrete, Theorem 6 yields measurable functions $g_{\text{lo}}, g_{\text{up}} : \mathcal{X}_{0:t} \rightarrow \mathbb{R}$ such that

$$g_{\text{lo}}(X_{0:t}(a_{1:t})) \leq \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})] \leq g_{\text{up}}(X_{0:t}(a_{1:t})) \quad \text{almost surely.} \quad (25)$$

In particular, $g_{\text{lo}}(x_{0:t})$ is obtained as the value of $\mathbb{E}[Y_{\text{lo}} \mid X_{0:N}(A_{1:N}) \in B_{0:N}]$ for $B_{0:t} := \{x_{0:t}\}$, and similarly for $g_{\text{up}}(x_{0:t})$. Moreover, since the constants $y_{\text{lo}}, y_{\text{up}} \in \mathbb{R}$ in Theorem 6 were allowed to depend on $B_{0:t}$, and hence here on each choice of $x_{0:t} \in \mathcal{X}_{0:t}$, we may think of these now as measurable functions $y_{\text{lo}}, y_{\text{up}} : \mathcal{X}_{0:t} \rightarrow \mathbb{R}$ satisfying

$$y_{\text{lo}}(X_{0:t}(a_{1:t})) \leq Y(a_{1:t}) \leq y_{\text{up}}(X_{0:t}(a_{1:t})) \quad \text{almost surely.} \quad (26)$$

In other words, when $X_{0:t}(a_{1:t})$ is discrete, Theorem 6 provides bounds on the conditional expectation of $Y(a_{1:t})$ given the value of $X_{0:t}(a_{1:t})$ whenever we have y_{lo} and y_{up} such that (26) holds.

When $\mathbb{P}(X_{1:t}(a_{1:t}) \in B_{1:t}) > 0$, we can also obtain bounds of the following form:

$$\begin{aligned} \mathbb{E}[Y_{\text{lo}} \mid X_0, X_{1:N}(A_{1:N}) \in B_{1:N}] &\leq \mathbb{E}[Y(a_{1:t}) \mid X_0, X_{1:t}(a_{1:t}) \in B_{1:t}] \\ &\leq \mathbb{E}[Y_{\text{up}} \mid X_0, X_{1:N}(A_{1:N}) \in B_{1:N}] \quad \text{almost surely.} \end{aligned} \quad (27)$$

This can be seen by straightforwardly modifying the proof of Theorem 6 to make all relevant quantities conditional on X_0 (rather than $X_0 \in B_0$). In particular, this holds regardless of whether or not X_0 is discrete. In turn, if $X_{1:t}(a_{1:t})$ is discrete, then by a similar argument as was given in the previous subsection, this yields almost sure bounds on $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})]$ of the form in (25), provided (26) holds. Alternatively, by taking $B_{1:t} := \mathcal{X}_{1:t}$, (27) yields bounds of the form

$$\mathbb{E}[Y_{\text{lo}} \mid X_0] \leq \mathbb{E}[Y(a_{1:t}) \mid X_0] \leq \mathbb{E}[Y_{\text{up}} \mid X_0].$$

If the action sequence $a_{1:t}$ is thought of as a single choice of an action from the extended action space $\mathcal{A}_{1:t}$, then this recovers the bounds originally proposed by Manski (1990), which allowed conditioning on potentially continuous pre-treatment covariates corresponding to our X_0 .

G.5 Proof of Theorem 9 (and Discussion)

Proof Suppose we have a permissible g_{lo} . (The case of g_{up} is analogous). Choose $x_{1:T} \in \mathcal{X}_{1:T}$ arbitrarily, and define new potential outcomes

$$\begin{aligned} \tilde{X}_0 &:= X_0 \\ \tilde{X}_r(a'_{1:r}) &:= \mathbb{1}(A_{1:r} = a'_{1:r}) X_r(a'_{1:r}) + \mathbb{1}(A_{1:r} \neq a'_{1:r}) x_r \quad \text{for } r \in \{1, \dots, T\} \text{ and } a'_{1:r} \in \mathcal{A}_{1:r}. \end{aligned}$$

Similarly, define

$$\tilde{Y}(a'_{1:t}) := \mathbb{1}(A_{1:t} = a'_{1:t}) Y(a'_{1:t}) + \mathbb{1}(A_{1:t} \neq a'_{1:t}) y_{\text{lo}}(\tilde{X}_{0:t}(a'_{1:t})) \quad \text{for all } a'_{1:t} \in \mathcal{A}_{1:t}.$$

It immediately follows that

$$(\tilde{X}_{0:T}(A_{1:T}), \tilde{Y}(A_{1:t}), A_{1:T}) \stackrel{\text{a.s.}}{=} (X_{0:T}(A_{1:T}), Y(A_{1:t}), A_{1:T}).$$

Moreover, it is easily checked that

$$y_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t})) \leq \tilde{Y}(a_{1:t}) \leq y_{\text{up}}(\tilde{X}_{0:t}(a_{1:t})) \quad \text{almost surely.}$$

As such, since g_{lo} is permissible, we must have, almost surely,

$$\begin{aligned} g_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t})) &\leq \mathbb{E}[\tilde{Y}(a_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t})] \\ &= \mathbb{E}[\tilde{Y}(A_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t}), A_{1:t} = a_{1:t}] \mathbb{P}(A_{1:t} = a_{1:t} \mid \tilde{X}_{0:t}(a_{1:t})) \\ &\quad + \underbrace{\mathbb{E}[\tilde{Y}(a_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t}), A_{1:t} \neq a_{1:t}]}_{=y_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t}))} \mathbb{P}(A_{1:t} \neq a_{1:t} \mid \tilde{X}_{0:t}(a_{1:t})). \end{aligned} \quad (28)$$

Now, by our definition of $\tilde{X}_{0:t}(a_{1:t})$, we have almost surely

$$\begin{aligned}
 & \mathbb{1}(A_1 \neq a_1) \mathbb{P}(A_{1:t} = a_{1:t} \mid \tilde{X}_{0:t}(a_{1:t})) \\
 &= \mathbb{1}(A_1 \neq a_1, \tilde{X}_s(a_{1:s}) = x_s) \mathbb{P}(A_{1:t} = a_{1:t} \mid \tilde{X}_{0:t}(a_{1:t})) \\
 &= \mathbb{1}(A_1 \neq a_1) \mathbb{E}[\mathbb{1}(A_{1:t} = a_{1:t}, \tilde{X}_s(a_{1:s}) = x_s) \mid \tilde{X}_{0:t}(a_{1:t})] \\
 &= \mathbb{1}(A_1 \neq a_1) \mathbb{E}[\mathbb{1}(A_{1:t} = a_{1:t}, X_s(A_{1:s}) = x_s) \mid \tilde{X}_{0:t}(a_{1:t})] \\
 &= 0,
 \end{aligned}$$

where the last step follows by our assumption that $\mathbb{P}(X_s(A_{1:s}) = x_s) = 0$. Combining this with (28), we get, almost surely,

$$\begin{aligned}
 \mathbb{1}(A_1 \neq a_1) g_{\text{lo}}(X_0, x_{1:t}) &= \mathbb{1}(A_1 \neq a_1) g_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t})) \\
 &\leq \mathbb{1}(A_1 \neq a_1) y_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t})) \\
 &= \mathbb{1}(A_1 \neq a_1) y_{\text{lo}}(X_0, x_{1:t}).
 \end{aligned} \tag{29}$$

Now let $x_0 \in \mathcal{X}_0$ be the value such that $\mathbb{P}(X_0 = x_0) = 1$. Using our assumption that $\mathbb{P}(A_1 \neq a_1) > 0$ and the fact that $x_{1:t}$ was arbitrary, we obtain

$$g_{\text{lo}}(x_{0:t}) \leq y_{\text{lo}}(x_{0:t}) \quad \text{for all } x_{1:t} \in \mathcal{X}_{1:t}.$$

The result now follows. ■

To gain intuition for the phenomenon underlying Theorem 9, consider a simplified model consisting of $\mathcal{X}$ -valued potential outcomes $(X(a') : a' \in \mathcal{A})$, real-valued potential outcomes $(Y(a') : a' \in \mathcal{A})$, and an $\mathcal{A}$ -valued random variable A representing the choice of action. (This constitutes a special case of our setup with $T = 1$ and $\mathcal{X}_0$ taken to be a singleton set.) Suppose moreover that the following conditions hold:

$$\begin{aligned}
 \mathbb{P}(X(A) = x) &= 0 \quad \text{for all } x \in \mathcal{X} \\
 \mathbb{P}(A = a) &< 1.
 \end{aligned}$$

We then have

$$\begin{aligned}
 & \mathbb{E}[Y(a) \mid X(a)] \\
 & \stackrel{\text{a.s.}}{=} \mathbb{E}[Y(A) \mid X(A), A = a] \mathbb{P}(A = a \mid X(a)) + \mathbb{E}[Y(a) \mid X(a), A \neq a] \mathbb{P}(A \neq a \mid X(a)).
 \end{aligned} \tag{30}$$

But now, since the behaviour of $X(a)$ is only observed on $\{A = a\}$, for any given value of $x \in \mathcal{X}$, we cannot rule out the possibility that

$$X(a) = \mathbb{1}(A = a) X(A) + \mathbb{1}(A \neq a) x \quad \text{almost surely.}$$

In turn, since $\mathbb{P}(A = a) > 0$, this would imply $\mathbb{P}(X(a) = x) > 0$, and, since $\mathbb{P}(X(A) = x) = 0$, that $\mathbb{P}(A = a \mid X(a) = x) = 0$. From (30), this would yield

$$\mathbb{E}[Y(a) \mid X(a) = x] = \mathbb{E}[Y(a) \mid X(a) = x, A \neq a].$$

But now, since the behaviour of $Y(a)$ is unobserved on $\{A \neq a\}$, intuitively speaking, the observational distribution does not provide any information about the value of the right-hand side, and therefore about the behaviour of $\mathbb{E}[Y(a) \mid X(a)]$ more generally since $x \in \mathcal{X}$ was arbitrary.

Appendix H. Hypothesis Testing Methodology

H.1 Validity of Testing Procedure

We show here that our procedure for testing $\hat{Q} \geq Q_{\text{lo}}$ based on the one-sided confidence intervals R_{lo}^α and $\hat{R}^\alpha$ has the correct probability of type I error, provided R_{lo}^α and $\hat{R}^\alpha$ have the correct coverage probabilities. In particular, the result below (which applies a standard union bound argument) shows that if $\hat{Q} \geq Q_{\text{lo}}$, then our test rejects (i.e. $\hat{R}^\alpha < R_{\text{lo}}^\alpha$) with probability at most α . An analogous result is easily proven for testing $\hat{Q} \leq Q_{\text{up}}$ also, with R_{lo}^α replaced by a one-sided upper $(1 - \alpha/2)$ -confidence interval for Q_{up} , and $\hat{R}^\alpha$ replaced by a one-sided lower $(1 - \alpha/2)$ -confidence interval for $\hat{Q}$.

Proposition 12 *Suppose that for some $\alpha \in (0, 1)$ we have random variables $\hat{R}^\alpha$ and R_{lo}^α satisfying*

$$\mathbb{P}(Q_{\text{lo}} \geq R_{\text{lo}}^\alpha) \geq 1 - \frac{\alpha}{2} \quad (31)$$

$$\mathbb{P}(\hat{Q} \leq \hat{R}^\alpha) \geq 1 - \frac{\alpha}{2}. \quad (32)$$

If $\hat{Q} \geq Q_{\text{lo}}$, then $\mathbb{P}(\hat{R}^\alpha < R_{\text{lo}}^\alpha) \leq \alpha$.

Proof If $\hat{Q} \geq Q_{\text{lo}}$, then we have

$$\{\hat{R}^\alpha < R_{\text{lo}}^\alpha\} \subseteq \{\hat{Q} > \hat{R}^\alpha\} \cup \{Q_{\text{lo}} < R_{\text{lo}}^\alpha\}.$$

To see this, note that

$$\begin{aligned} (\{\hat{Q} > \hat{R}^\alpha\} \cup \{Q_{\text{lo}} < R_{\text{lo}}^\alpha\})^c &= \{\hat{Q} > \hat{R}^\alpha\}^c \cap \{Q_{\text{lo}} < R_{\text{lo}}^\alpha\}^c \\ &= \{\hat{Q} \leq \hat{R}^\alpha\} \cap \{Q_{\text{lo}} \geq R_{\text{lo}}^\alpha\} \\ &\subseteq \{R_{\text{lo}}^\alpha \leq \hat{R}^\alpha\}. \end{aligned}$$

As such,

$$\begin{aligned} \mathbb{P}(\hat{R}^\alpha < R_{\text{lo}}^\alpha) &\leq \mathbb{P}(\{\hat{Q} > \hat{R}^\alpha\} \cup \{Q_{\text{lo}} < R_{\text{lo}}^\alpha\}) \\ &\leq \mathbb{P}(\hat{Q} > \hat{R}^\alpha) + \mathbb{P}(Q_{\text{lo}} < R_{\text{lo}}^\alpha) \\ &\leq \alpha/2 + \alpha/2 \\ &= \alpha. \end{aligned}$$

■

H.2 Unbiased Sample Mean Estimates of Q_{lo} , $\hat{Q}$, and Q_{up}

We use our data to obtain one-sided confidence intervals R_{lo}^α and $\hat{R}^\alpha$ satisfying (31) and (32) as required by our procedure for testing $\hat{Q} \geq Q_{\text{lo}}$. We use an analogous procedure to obtain confidence intervals for testing $\hat{Q} \leq Q_{\text{up}}$. We tried two techniques for this:

an exact method based on Hoeffding's inequality, and an approximate method based on bootstrapping. Conceptually, both are based on obtaining unbiased sample mean estimates of Q_{lo} and $\hat{Q}$, which we describe now, before giving the particulars of each method in the next two subsections.

We begin with our sample mean estimator of Q_{lo} . Recall that we assume access to a dataset $\mathcal{D}$ consisting of i.i.d. copies of observational trajectories of the form

$$X_0, A_1, X_1(A_1), \dots, A_T, X_T(A_{1:T}).$$

Let $\mathcal{D}(a_{1:t}, B_{0:t})$ be the subset of trajectories in $\mathcal{D}$ for which $X_{0:N}(A_{1:N}) \in B_{0:N}$. Obtaining $\mathcal{D}(a_{1:t}, B_{0:t})$ is possible since the only random quantity that $N = \max\{0 \leq s \leq t \mid A_{1:s} = a_{1:s}\}$ depends on is $A_{1:t}$, which is included in the data. We denote the cardinality of $\mathcal{D}(a_{1:t}, B_{0:t})$ by $n := |\mathcal{D}(a_{1:t}, B_{0:t})|$. We then denote by $Y_{\text{lo}}^{(i)}$ for $i \in \{1, \dots, n\}$ the corresponding values of

$$Y_{\text{lo}} = \mathbb{1}(A_{1:t} = a_{1:t}) f(X_{0:t}(A_{1:t})) + \mathbb{1}(A_{1:t} \neq a_{1:t}) y_{\text{lo}}$$

obtained from each trajectory in $\mathcal{D}(a_{1:t}, B_{0:t})$. This is again possible since both terms only depend on the observational quantities $(X_{0:t}(A_{1:t}), A_{1:t})$. It is easily seen that the values of $Y_{\text{lo}}^{(i)}$ are i.i.d. and satisfy $\mathbb{E}[Y_{\text{lo}}^{(i)}] = Q_{\text{lo}}$. As a result, the sample mean

$$\mu_{\text{lo}} := \frac{1}{n} \sum_{i=1}^n Y_{\text{lo}}^{(i)} \quad (33)$$

is an unbiased estimator of Q_{lo} .

We obtain an unbiased sample mean estimate of $\hat{Q}$ in a similar fashion as for Q_{lo} . Recall that we assume access to a dataset $\hat{\mathcal{D}}(a_{1:t})$ consisting of i.i.d. copies of

$$X_0, \hat{X}_1(X_0, a_1), \dots, \hat{X}_t(X_0, a_{1:t}).$$

Let $\hat{\mathcal{D}}(a_{1:t}, B_{0:t})$ denote the subset of trajectories in $\hat{\mathcal{D}}(a_{1:t})$ with $(X_0, \hat{X}_t(X_0, a_{1:t})) \in B_{0:t}$, and denote its cardinality by $\hat{n} := |\hat{\mathcal{D}}(a_{1:t}, B_{0:t})|$. Then denote by $\hat{Y}^{(i)}$ for $i \in \{1, \dots, \hat{n}\}$ the corresponding values of

$$\hat{Y} = f(X_0, \hat{X}_{1:t}(X_0, a_{1:t}))$$

obtained from each trajectory in $\hat{\mathcal{D}}(a_{1:t}, B_{0:t})$. It is easily seen that the values $\hat{Y}^{(i)}$ are i.i.d. (since the entries of $\hat{\mathcal{D}}(a_{1:t})$ are) and satisfy $\mathbb{E}[\hat{Y}^{(i)}] = \hat{Q}$. As a result, the sample mean

$$\hat{\mu} := \frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \hat{Y}^{(i)}$$

is an unbiased estimator of $\hat{Q}$.

H.3 Exact Confidence Intervals via Hoeffding's Inequality

Recall that we assume in Section 5.1 that $Y(a_{1:t})$ has the form $Y(a_{1:t}) = f(X_{0:t}(a_{1:t}))$, and that moreover

$$y_{\text{lo}} \leq f(x_{0:t}) \leq y_{\text{up}} \quad \text{for all } x_{0:t} \in B_{0:t}. \quad (34)$$

This means $\widehat{Y}^{(i)}$ is almost surely bounded in $[y_{\text{lo}}, y_{\text{up}}]$, and so $\widehat{\mu}$ gives rise to one-sided confidence intervals via an application of Hoeffding's inequality. The exact form of these confidence intervals is as follows:

Proposition 13 *If (34) holds, then for each $\alpha \in (0, 1)$, letting*

$$\Delta := (y_{\text{up}} - y_{\text{lo}}) \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \quad \text{and} \quad \widehat{\Delta} := (y_{\text{up}} - y_{\text{lo}}) \sqrt{\frac{1}{2\widehat{n}} \log \frac{2}{\alpha}},$$

and similarly

$$R_{\text{lo}}^\alpha := \mu_{\text{lo}} - \Delta \quad \text{and} \quad \widehat{R}^\alpha := \widehat{\mu} + \widehat{\Delta},$$

it follows that

$$\mathbb{P}(Q_{\text{lo}} \geq R_{\text{lo}}^\alpha) \geq 1 - \frac{\alpha}{2} \quad \text{and} \quad \mathbb{P}(\widehat{Q} \leq \widehat{R}^\alpha) \geq 1 - \frac{\alpha}{2}.$$

Proof We only prove the result for R_{lo}^α ; the other statement can be proved analogously. Recall that μ_{lo} is the empirical mean of i.i.d. samples $Y_{\text{lo}}^{(i)}$ for $i \in \{1, \dots, n\}$ with $\mathbb{E}[Y_{\text{lo}}^{(i)}] = Q_{\text{lo}}$ (see Equation 33). Moreover, by (34), $Y_{\text{lo}}^{(i)}$ is almost surely bounded in $[y_{\text{lo}}, y_{\text{up}}]$. Hoeffding's inequality then implies that

$$\mathbb{P}(\mu_{\text{lo}} - Q_{\text{lo}} > \Delta) \leq \exp\left(-\frac{2n\Delta^2}{(y_{\text{up}} - y_{\text{lo}})^2}\right).$$

In turn, some basic manipulations yield

$$\begin{aligned} \mathbb{P}(Q_{\text{lo}} \geq R_{\text{lo}}^\alpha) &= 1 - \mathbb{P}(Q_{\text{lo}} < \mu_{\text{lo}} - \Delta) \\ &\geq 1 - \exp\left(-\frac{2n\Delta^2}{(y_{\text{up}} - y_{\text{lo}})^2}\right) \\ &= 1 - \frac{\alpha}{2}. \end{aligned}$$

■

H.4 Approximate Confidence Intervals via Bootstrapping

While Hoeffding's inequality yields the probability guarantees in (31) and (32) exactly, the confidence intervals obtained can be conservative. Consequently, our testing procedure may have lower probability of falsifying certain twins that in fact do not satisfy the causal bounds. To address this, we also consider an approximate approach based on bootstrapping (Efron, 1979) that can produce tighter confidence intervals. While other schemes are possible, bootstrapping provides a general-purpose approach that is straightforward to implement and works well in practice.

At a high level, our approach here is again to construct one-sided level $1 - \alpha/2$ confidence intervals via bootstrapping on Q_{lo} and $\widehat{Q}$. Many bootstrapping procedures for obtaining

confidence intervals have been proposed in the literature (Efron and Tibshirani, 1993; Davison and Hinkley, 1997; Hesterberg, 2015). Our results reported below were obtained via the *reverse percentile* bootstrap (see Hesterberg 2015 for an overview). (We also tried the *percentile* bootstrap method, which obtained nearly indistinguishable results.) In particular, this method takes

$$R_{\text{lo}}^\alpha := 2\mu_{\text{lo}} - \Delta_{\text{b}} \quad \hat{R}^\alpha := 2\hat{\mu} - \hat{\Delta}_{\text{b}},$$

where Δ_{b} and $\hat{\Delta}_{\text{b}}$ correspond to the approximate $1 - \alpha/2$ and $\alpha/2$ quantiles of the distributions of

$$\frac{1}{n} \sum_{i=1}^n Y_{\text{lo}}^{(i*)} \quad \text{and} \quad \frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \hat{Y}^{(i*)},$$

where each $Y_{\text{lo}}^{(i*)}$ and $Y^{(i*)}$ is obtained by sampling uniformly with replacement from among the values of $Y_{\text{lo}}^{(i)}$ and $Y^{(i)}$. In our case study, as is typically done in practice, we approximated Δ_{b} and $\hat{\Delta}_{\text{b}}$ via Monte Carlo sampling. It can be shown that the confidence intervals produced in this way obtain a coverage level that approaches the desired level of $1 - \alpha/2$ as n and $\hat{n}$ grow to infinity under mild assumptions (Hall, 1988).

Appendix I. Experimental Details

I.1 MIMIC Preprocessing

For data extraction and preprocessing, we used the same procedure as Komorowski et al. (2018) with minor modifications. For completeness, we describe the pre-processing steps applied in Komorowski et al. (2018) and subsequently outline our modifications to these. First, we extracted adult patients fulfilling the sepsis-3 criteria (Singer et al., 2016). Sepsis was defined as a suspected infection (as indicated by prescription of antibiotics and sampling of bodily fluids for microbiological culture) combined with evidence of organ dysfunction, defined by a SOFA score ≥ 2 (Singer et al., 2016; Seymour et al., 2016). Then we excluded patients for whom any of the following was true: their age was less than 18 years old at the time of ICU admission; their mortality not documented; their IV fluid/vasopressors intake was not documented; their treatment was withdrawn.

We made the following modifications to the preprocessing code of Komorowski et al. (2018) for our experiment. First, instead of extracting physiological quantities (e.g. heart rate) every 4 hours, we extracted these every hour. Additionally, we excluded patients with any missing hourly vitals during the first 4 hours of their ICU stay.

We then extracted a total of 19 quantities of interest listed in Table 2. Of these, 17 were physiological quantities associated with the patient, including static demographic quantities (e.g. age), patient vital signs (e.g. heart rate), and patient lab values (e.g. potassium blood concentration). All of these were continuous values, apart from sex. These were chosen since they were the subset of physiological quantities extracted from MIMIC by Komorowski et al. (2018) that are also modelled by Pulse, and were used to define our observation spaces $\mathcal{X}_t$ as described next. The remaining 2 quantities (intravenous fluids and vasopressor doses) were chosen since they correspond to treatments that the patient received, and were used to define our action spaces $\mathcal{A}_t$ as described below.

Category	Physiological quantity
Demographic	Age Sex Weight
Vital Signs	Heart rate (HR) Systolic blood pressure (SysBP) Diastolic blood pressure (DiaBP) Mean blood pressure (MeanBP) Respiratory Rate (RR) Skin Temperature (Temp)
Lab Values	Potassium Blood Concentration (Potassium) Sodium Blood Concentration (Sodium) Chloride Blood Concentration (Chloride) Glucose Blood Concentration (Glucose) Calcium Blood Concentration (Calcium) Bicarbonate Blood Concentration (HCO_3) Arterial O_2 Pressure (PaO_2) Arterial CO_2 Pressure (PaCO_2)
Treatments	Intravenous fluid (IV) dose Vasopressor dose

Table 2: Physiological quantities and treatments extracted from MIMIC

I.2 Sample Splitting

Before proceeding further, we randomly selected 5% of our extracted trajectories (583 trajectories, denoted as $\mathcal{D}_0$) to use for preliminary tasks such as choosing the parameters of our hypotheses. We reserved the remaining 95% (11,094 trajectories, denoted as $\mathcal{D}$) for the actual testing. By a standard sample splitting argument (Cox, 1975), the statistical guarantees of our testing procedure established above continue to apply even when our hypotheses are defined in this data-dependent way.

I.3 Observation Spaces

Our $\mathcal{X}_0$ consisted of the following features: age, sex, weight, heart rate, systolic blood pressure, diastolic blood pressure and respiration rate. We chose $\mathcal{X}_0$ in this way because, out of the 17 physiological quantities we extracted from MIMIC, these were the quantities that can be initialised to user-provided values before starting a simulation in version 4 of Pulse, which we used. (In contrast, Pulse initialises the other 10 features to default values.) For the remaining observation spaces, we used the full collection of the 17 physiological quantities we extracted to define $\mathcal{X}_1 = \dots = \mathcal{X}_4$. We encoded all features in $\mathcal{X}_t$ numerically, i.e. $\mathcal{X}_0 = \mathbb{R}^7$, and $\mathcal{X}_t = \mathbb{R}^{17}$ for $t \in \{1, 2, 3, 4\}$.

I.4 Action Spaces

Following Komorowski et al. (2018), we constructed our action space using 2 features obtained from MIMIC, namely intravenous fluid (IV) and vasopressor doses. To obtain discrete action spaces suitable for our framework, we used the same discretization procedure

		Vasopressor dose ($\mu\text{g/kg/min}$)				
		0	0.0 - 0.061	0.061 - 0.15	0.15 - 0.313	> 0.313
IV dose (mL/h)	0	16659	329	256	152	145
	0 - 20	5840	428	351	244	145
	20 - 75	6330	297	378	383	309
	75 - 214	6232	176	175	197	273
	> 214	5283	347	488	544	747

Table 3: Action space with frequency of occurrence in observational data

for these quantities as was used by Komorowski et al. (2018). Specifically, we divided the hourly doses of intravenous fluids and vasopressors into 5 bins each, with the first bin corresponding to zero drug dosage, and the remaining 4 bins based on the quartiles of the non-zero drug dosages in our held-out observational dataset $\mathcal{D}_0$. From this we obtained action spaces $\mathcal{A}_1 = \dots = \mathcal{A}_4$ with $5 \times 5 = 25$ elements. Table 3 shows the dosage bins constructed in this way, as well as the frequency of each bin’s occurrence in the observational data.

I.5 Hypothesis Parameters

We used our held-out observational dataset $\mathcal{D}_0$ to obtain a collection of hypothesis parameters $(t, f, a_{1:t}, B_{0:t})$. Specifically, for each physiological quantity of interest (e.g. heart rate) in the list of Vital Signs and Lab Values given in Table 2, we did the following. First, for each $t \in \{0, \dots, 4\}$, we obtained 16 choices of B_t by discretizing the patient space $\mathcal{X}_t$ into 16 subsets based on the values of certain features as follows: 2 bins corresponding to sex; 4 bins corresponding to the quartiles of the ages of patients in $\mathcal{D}_0$; 2 bins corresponding to whether or not the value of the chosen physiological quantity of interest at time t was above or below its median value in $\mathcal{D}_0$.

Next, for each $t \in \{1, \dots, 4\}$, $a_{1:t} \in \mathcal{A}_{1:t}$, and sequence $B_{0:t}$ with each $B_{t'}$ as defined in the previous step, let $\mathcal{D}_0(t, a_{1:t}, B_{0:t})$ denote the subset of $\mathcal{D}_0$ corresponding to $(t, a_{1:t}, B_{0:t})$, i.e.

$$\begin{aligned} \mathcal{D}_0(t, a_{1:t}, B_{0:t}) \\ := \{X_{0:t}(A_{1:t}) \mid (X_{0:T}(A_{1:T}), A_{1:T}) \in \mathcal{D}_0 \text{ with } A_{1:t} = a_{1:t} \text{ and } X_{0:t}(A_{1:T}) \in B_{0:t}\}. \end{aligned}$$

We then selected the set of all triples $(t, a_{1:t}, B_{0:t})$ such that $\mathcal{D}_0(t, a_{1:t}, B_{0:t})$ contained at least one trajectory. This meant the number of combinations of hypotheses parameters that we considered was limited to a tractable quantity, which had benefits both computationally, and also by ensuring that we did not sacrifice too much power when adjusting for multiple testing.

Finally, for each selected triple $(t, a_{1:t}, B_{0:t})$, we chose a corresponding f as follows. First, we let $i \in \{1, \dots, d_t\}$ denote the index of the physiological quantity of interest in $\mathcal{X}_t = \mathbb{R}^{d_t}$. We then set $y_{\text{lo}}, y_{\text{up}}$ to be the .2 and the .8 quantiles of the values in

$$\{(X_t(A_{1:t}))_i \mid X_{0:t}(A_{1:t}) \in \mathcal{D}_0(t, a_{1:t}, B_{0:t})\}.$$

We then obtained $f : \mathcal{X}_{0:t} \rightarrow \mathbb{R}$ as the function that extracts the physiological quantity of interest from $\mathcal{X}_t$ and clips its value to lie between the values y_{lo} and y_{up} . Explicitly, $f(x_{0:t}) := \min(\max((x_t)_i, y_{\text{lo}}), y_{\text{up}})$.

Overall, accounting for all physiological quantities of interest, we obtained 721 distinct choices of $(t, f, a_{1:t}, B_{0:t})$ in this way. Figure 5 shows the amount of non-held out observational and twin data that we subsequently used for testing each hypothesis, i.e. the values of n and $\hat{n}$ as defined in Section H.2 above. We describe how we generated our dataset of twin trajectories next.

I.6 Generating Twin Trajectories Using the Pulse Physiology Engine

The Pulse Physiology Engine is a comprehensive open-source human physiology simulator that has been used in medical education, research, and training. Pulse allows users to initialize patient trajectories with given age, sex, weight, heart rate, systolic blood pressure, diastolic blood pressure and respiration rate and medical conditions such as sepsis, COPD, ARDS, etc. Once initialised, users have the ability to advance patient trajectories by a given time step (one hour in our case), and administer actions (e.g. administer a given dose of IV fluids or vasopressors). Detailed documentation is available at: pulse.kitware.com.

In Algorithm 1 we describe how we generated the twin data to test the chosen hypotheses. Note that we sampled X_0 without replacement as it ensures that each X_0 is chosen at most once and consequently twin trajectories in $\hat{\mathcal{D}}(a_{1:t})$ are i.i.d. Additionally, Algorithm 1 can be easily parallelised to improve efficiency. Figure 5 shows histograms of the number of twin trajectories $\hat{n}$ (as defined in Section H.2 above) obtained in this way across all hypotheses.

Algorithm 1: Generating Twin data $\hat{\mathcal{D}}(a_{1:t})$.

Inputs: Action sequence $a_{1:t}$; Observational dataset $\mathcal{D}$.

Output: Twin data $\hat{\mathcal{D}}(a_{1:t})$ of size m .

for $i = 1, \dots, m$ **do**

 Sample X_0 without replacement from $\mathcal{D}$;

 Initialize the Pulse trajectory with the information of X_0 ;

for $s = 1, \dots, t$ **do**

 Administer the median doses of IV fluids and vasopressors in action bin $a_{t'}$;

if $s \equiv 0 \pmod{3}$ **then**

 Virtual patient in Pulse consumes nutrients and water, and urinates;

end

 Advance the twin trajectory by one hour;

end

 Add the trajectory to $\hat{\mathcal{D}}(a_{1:t})$;

end

Return $\hat{\mathcal{D}}(a_{1:t})$

Physiological quantity	Ours		Manski	
	Rejs.	Hyps.	Rejs.	Hyps.
Chloride Blood Concentration (Chloride)	24	94	1	46
Sodium Blood Concentration (Sodium)	21	94	9	46
Potassium Blood Concentration (Potassium)	13	94	0	46
Skin Temperature (Temp)	10	86	9	46
Calcium Blood Concentration (Calcium)	5	88	0	46
Glucose Blood Concentration (Glucose)	5	96	1	46
Arterial CO ₂ Pressure (paCO ₂)	3	70	0	46
Bicarbonate Blood Concentration (HCO ₃)	2	90	1	46
Systolic Arterial Pressure (SysBP)	2	154	0	46
Arterial O ₂ Pressure (paO ₂)	0	78	1	46
Arterial pH (Arterial_pH)	0	80	0	46
Diastolic Arterial Pressure (DiaBP)	0	72	0	46
Mean Arterial Pressure (MeanBP)	0	92	0	46
Respiration Rate (RR)	0	172	0	46
Heart Rate (HR)	0	162	0	46

Table 4: Total hypotheses (Hyps.) and rejections (Rejs.) per physiological quantity obtained using Hoeffding’s inequality

Physiological quantity	Ours		Manski	
	Rejs.	Hyps.	Rejs.	Hyps.
Chloride Blood Concentration (Chloride)	47	94	1	46
Sodium Blood Concentration (Sodium)	46	94	12	46
Potassium Blood Concentration (Potassium)	33	94	0	46
Skin Temperature (Temp)	43	86	13	46
Calcium Blood Concentration (Calcium)	44	88	0	46
Glucose Blood Concentration (Glucose)	19	96	0	46
Arterial CO ₂ Pressure (paCO ₂)	13	70	0	46
Bicarbonate Blood Concentration (HCO ₃)	8	90	0	46
Systolic Arterial Pressure (SysBP)	8	154	0	46
Arterial O ₂ Pressure (paO ₂)	4	78	1	46
Arterial pH (Arterial_pH)	0	80	0	46
Diastolic Arterial Pressure (DiaBP)	0	72	0	46
Mean Arterial Pressure (MeanBP)	3	92	0	46
Respiration Rate (RR)	12	172	0	46
Heart Rate (HR)	1	162	0	46

Table 5: Total hypotheses (Hyps.) and rejections (Rejs.) per physiological quantity obtained using the reverse percentile bootstrap

I.7 Bootstrapping Details

In addition to Hoeffding’s inequality, we also used reverse percentile bootstrap method (see e.g. Hesterberg 2015) to obtain our confidence intervals on Q_{lo} and Q_{up} as described in Section H.2. We used 100 bootstrap samples for each confidence interval. To avoid bootstrapping on small numbers of data points, we did not reject any hypothesis where either the number of observational trajectories n or twin trajectories $\hat{n}$ was less than 100, and returned a p -value of 1 in each such case.

Table 5 shows the number of rejected hypotheses for each physiological quantity using this approach. We observed a similar trend as in our results obtained using Hoeffding’s inequality (Table 4). For example, we obtained high number of rejections for Sodium, Chloride and Potassium blood concentrations but few rejections for Arterial Pressure and Heart Rate. Overall, bootstrapping increased the number of rejected hypotheses by a factor of roughly 3.3 compared with Hoeffding’s inequality (281 vs. 85 rejections in total). Like we described for Hoeffding’s inequality in the main text, we also ran this analysis using the unconditional bounds of Manski (1990), and again obtained substantially fewer rejections compared with our approach based on Theorem 6.

I.8 Tightness of Bounds and Number of Data Points per Hypothesis

In this section, we show empirically how both the tightness of the bounds $[Q_{\text{lo}}, Q_{\text{up}}]$ and the number of data points per hypothesis relate to the number of falsifications obtained in our case study. Recall that the tightness of $[Q_{\text{lo}}, Q_{\text{up}}]$ is determined by the value of $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$, since we have

$$\frac{Q_{\text{up}} - Q_{\text{lo}}}{y_{\text{up}} - y_{\text{lo}}} = 1 - \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}). \quad (35)$$

Here the left-hand side is a number in $[0, 1]$ that quantifies the tightness of the bounds $[Q_{\text{lo}}, Q_{\text{up}}]$ relative to the trivial worst-case bounds $[y_{\text{lo}}, y_{\text{up}}]$, with smaller values meaning tighter bounds. The equation above shows that the higher the value of $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$, the tighter the bounds are.

Figure 8 shows the bounds are often informative in practice, with $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$ being reasonably large (and hence the bounds tight, by (35) above) for a significant number of hypotheses we consider. However, rejections still occur even when the bounds are reasonably loose (e.g. $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}) \approx 0.3$), which shows our method can still yield useful information even in this case. We moreover observe rejections across a range of different numbers of observational data points used to test each hypothesis, which shows that our method is not strongly dependent on the size of the dataset obtained.

I.9 Sensitivity to y_{lo} and y_{up}

We investigated the sensitivity of our methodology with respect to our choices of the values y_{lo} and y_{up} . Specifically, we repeated our procedure with the intervals $[y_{\text{lo}}, y_{\text{up}}]$ replaced with $[y_{\text{lo}}(1 - \Delta/2), y_{\text{up}}(1 + \Delta/2)]$ for a range of different values of $\Delta \in \mathbb{R}$. Figure 9 plots the number of rejections for different values of Δ . We observe that for significantly larger $[y_{\text{lo}}, y_{\text{up}}]$ intervals, we do obtain fewer rejections, although this is to be expected since the widths of the bounds $[Q_{\text{lo}}, Q_{\text{up}}]$ and our confidence intervals R_{lo}^α and R_{up}^α obtained using Hoeffding’s inequality (see Proposition 13) both grow increasingly large as the width of $[y_{\text{lo}}, y_{\text{up}}]$ grows. However, we observe that the number of rejections per outcome is stable for a moderate range of widths of $[y_{\text{lo}}, y_{\text{up}}]$, which indicates that our method is reasonably robust to the choice of $y_{\text{lo}}, y_{\text{up}}$ parameters.

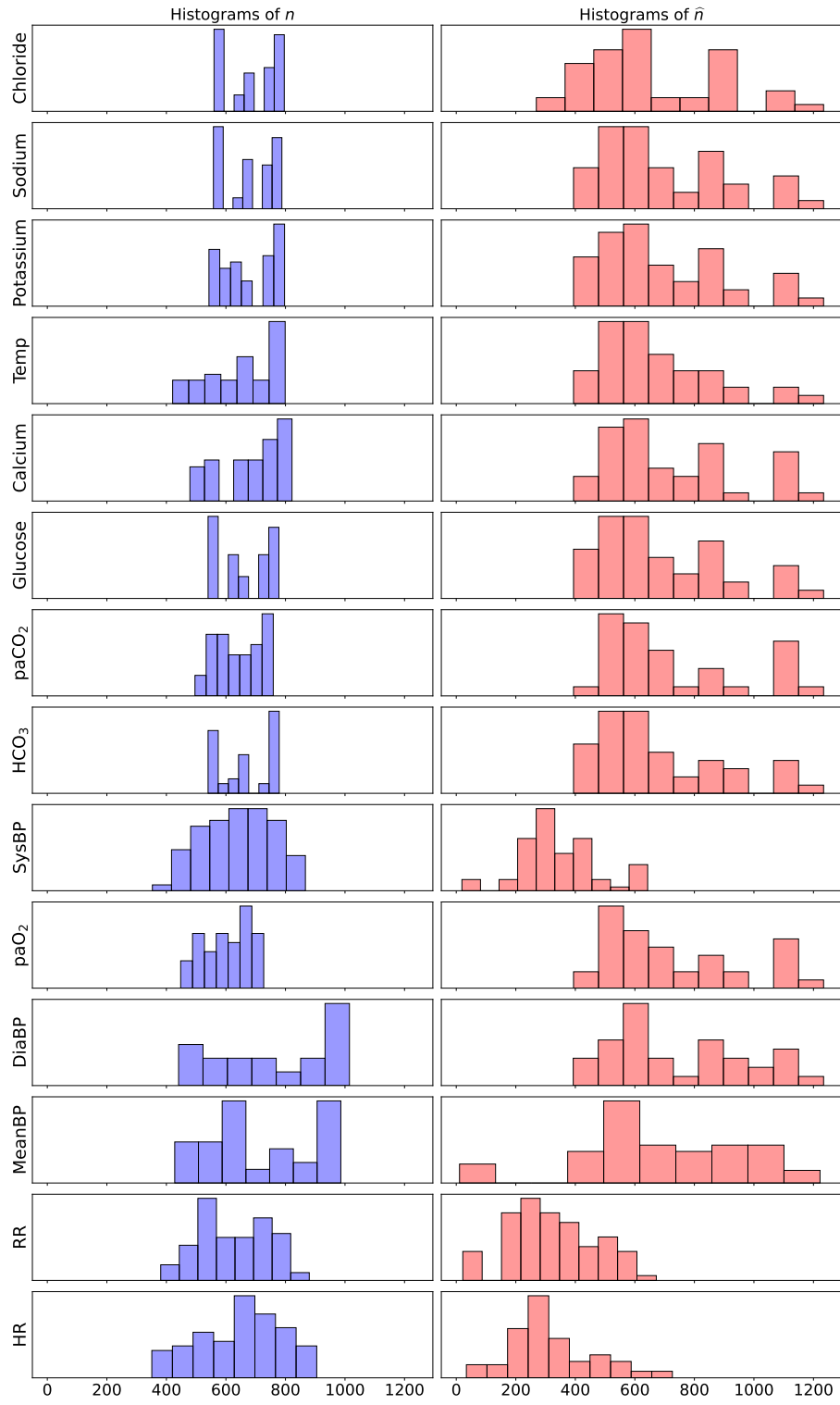


Figure 5: Histograms of n and $\hat{n}$ (as defined in Section H.2) across all hypothesis parameters corresponding to each physiological quantity of interest.

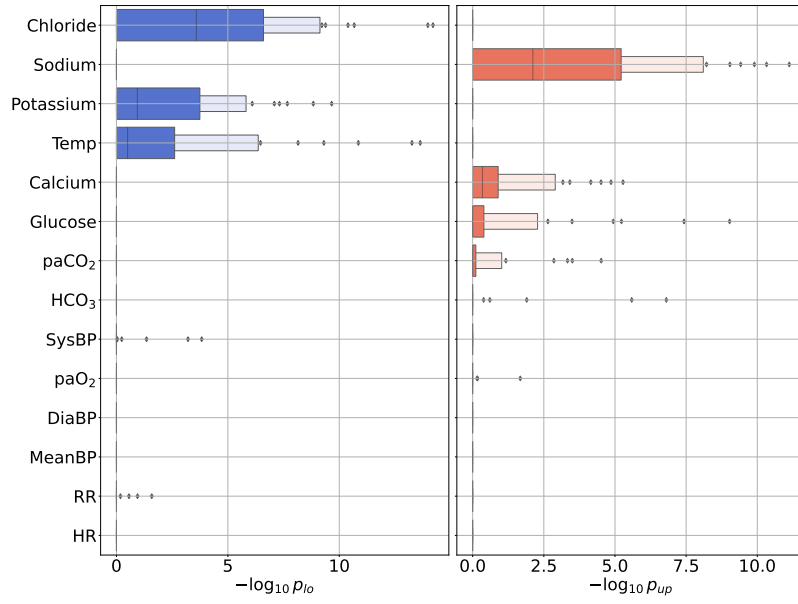


Figure 6: Boxenplots showing distributions of $-\log_{10} p_{lo}$ and $-\log_{10} p_{up}$ for different physiological quantities obtained via Hoeffding's inequality. Higher values indicate greater evidence in favour of rejection.

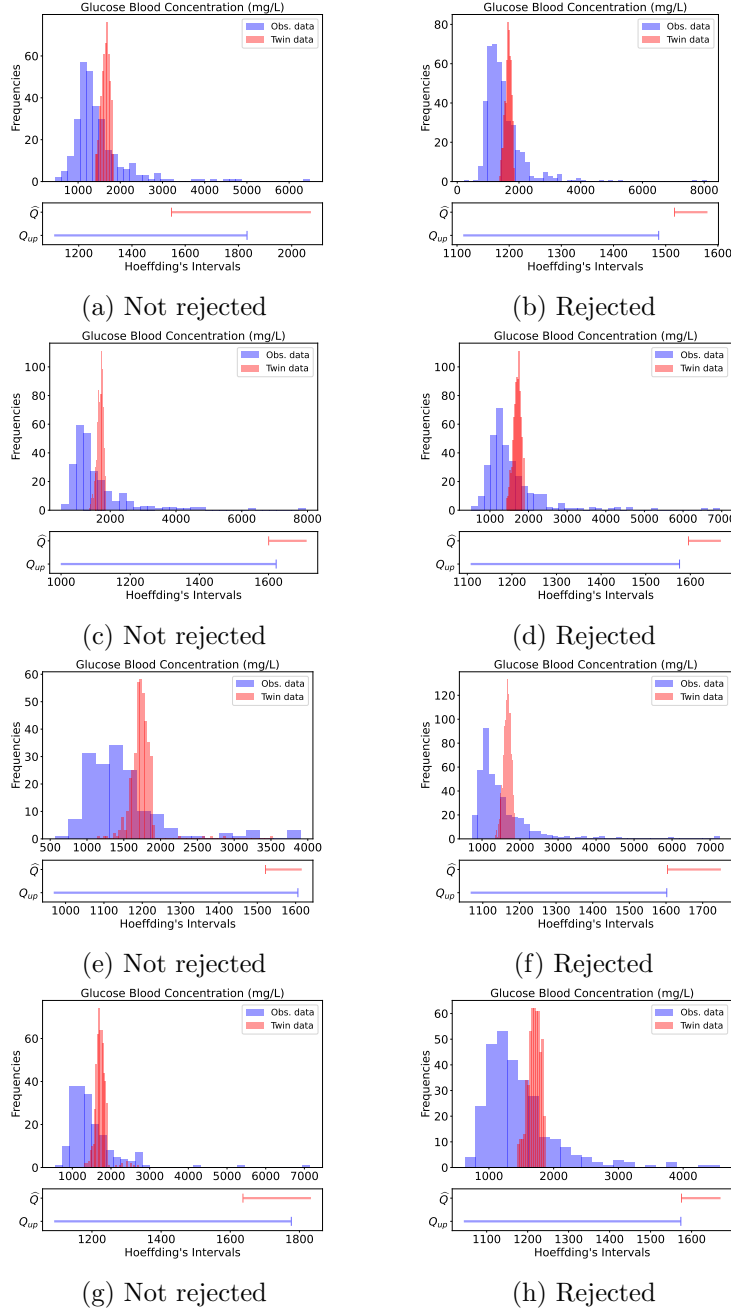


Figure 7: Raw observational data values conditional on $A_{1:t} = a_{1:t}$ and $X_{0:t}(A_{1:t}) \in B_{0:t}$, and from the output of the twin conditional on $\hat{X}_{0:t}(a_{1:t}) \in B_{0:t}$. Each row shows two distinct choices of $(B_{0:t}, a_{1:t})$. Below each figure are shown 95% Hoeffding confidence intervals for $\hat{Q}$ and Q_{up} . Unlike Figure 3 from the main text, the horizontal axes of the histograms are not truncated, and the first row is in particular an untruncated version of Figure 3 from the main text. Note however that the scales of the horizontal axes of the confidence intervals differ from those of the histograms, since it is visually more difficult to determine whether or not the confidence intervals overlap when fully zoomed out.

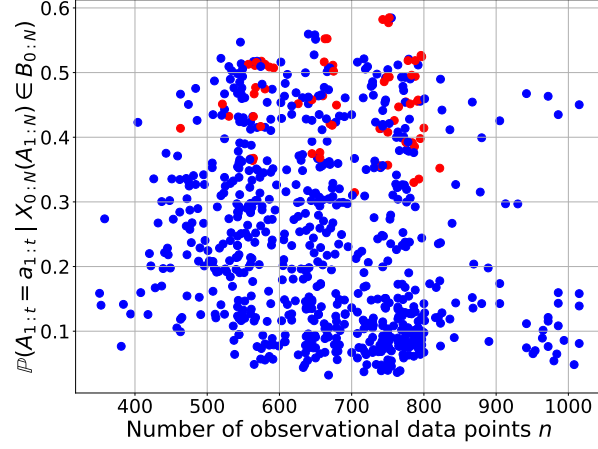


Figure 8: Sample mean estimate of $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$ for each pair of hypotheses $(\mathcal{H}_{lo}, \mathcal{H}_{up})$ corresponding to the same set of parameters $(t, f, a_{1:t}, B_{0:t})$ that we tested, along with the corresponding number of observational data points used to test each hypothesis. Red points indicate that either $\mathcal{H}_{lo}$ or $\mathcal{H}_{up}$ were rejected, while blue points indicate that both $\mathcal{H}_{lo}$ and $\mathcal{H}_{up}$ were not rejected.

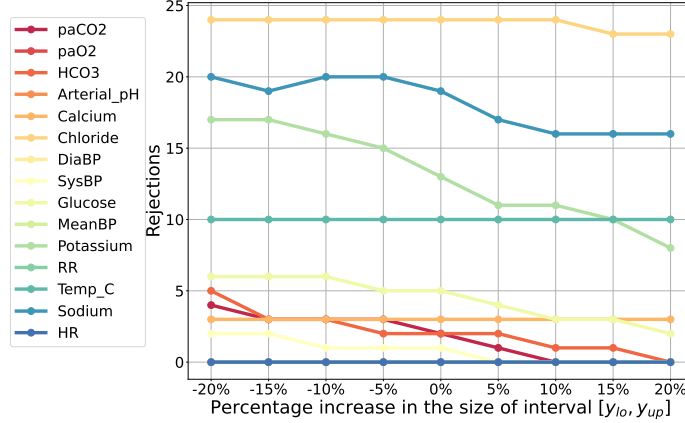


Figure 9: Rejections obtained as the width of the $[y_{lo}, y_{up}]$ interval changes. Here, the interval is increased (or decreased) symmetrically on each side.

References

- Jean Pierre Allamaa, Panagiotis Patrinos, Herman Van der Auweraer, and Tong Duy Son. Sim2real for autonomous vehicle control using executable digital twin. *IFAC-PapersOnLine*, 55(24):385–391, 2022. ISSN 2405-8963. doi: <https://doi.org/10.1016/j.ifacol.2022.10.314>. URL <https://www.sciencedirect.com/science/article/pii/S2405896322023461>. 10th IFAC Symposium on Advances in Automotive Control AAC 2022.
- Donald W. K. Andrews and Gustavo Soares. Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157, 2010. doi: <https://doi.org/10.3982/ECTA7502>. URL <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA7502>.
- ASME. *Assessing Credibility of Computational Modeling through Verification and Validation: Application to Medical Devices*. American Society of Mechanical Engineers, New York, NY, 2018. ASME V&V 40-2018.
- Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997. doi: 10.1080/01621459.1997.10474074. URL <https://doi.org/10.1080/01621459.1997.10474074>.
- Barbara Rita Barricelli, Elena Casiraghi, and Daniela Fogli. A survey on digital twin: Definitions, characteristics, applications, and design implications. *IEEE Access*, 7:167653–167671, 2019.
- Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998. URL <https://doi.org/10.1214/aos/1013699998>.
- Aaron Bray, Jeffrey B. Webb, Andinet Enquobahrie, Jared Vicory, Jerry Heneghan, Robert Hubal, Stephanie TerMaath, Philip Asare, and Rachel B. Clipp. Pulse Physiology Engine: An open-source software platform for computational modeling of human medical simulation. *SN Comprehensive Clinical Medicine*, 1(5):362–377, 2019. doi: 10.1007/s42399-019-00053-w. URL <https://doi.org/10.1007/s42399-019-00053-w>.
- Federico A. Bugni, Ivan A. Canay, and Xiaoxia Shi. Specification tests for partially identified models defined by moment inequalities. *Journal of Econometrics*, 185(1):259–282, 2015. ISSN 0304-4076. doi: <https://doi.org/10.1016/j.jeconom.2014.10.013>. URL <https://www.sciencedirect.com/science/article/pii/S0304407614002577>.
- Ivan A. Canay, Gastón Illanes, and Amilcar Velez. A user’s guide for inference in models defined by moment inequalities. *Journal of Econometrics*, page 105558, 2023. ISSN 0304-4076. doi: <https://doi.org/10.1016/j.jeconom.2023.105558>. URL <https://www.sciencedirect.com/science/article/pii/S0304407623002749>.
- Victor Chernozhukov, Han Hong, and Elie Tamer. Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284, 2007. ISSN 00129682, 14680262. URL <http://www.jstor.org/stable/4502031>.

- Genevieve Coorey, Gemma A. Figtree, David F. Fletcher, Victoria J. Snelson, Stephen Thomas Vernon, David Winlaw, Stuart M. Grieve, Alistair McEwan, Jean Yee Hwa Yang, Pierre Qian, Kieran O'Brien, Jessica Orchard, Jinman Kim, Sanjay Patel, and Julie Redfern. The health digital twin to tackle cardiovascular disease—a review of an emerging interdisciplinary field. *npj Digital Medicine*, 5(1):126, 2022.
- Jorge Corral-Acero, Francesca Margara, Maciej Marciniak, Cristobal Roderio, Filip Loncaric, Yingjing Feng, Andrew Gilbert, Joao F Fernandes, Hassaan A Bukhari, Ali Wajdan, Manuel Villegas Martinez, Mariana Sousa Santos, Mehrdad Shamohammdi, Hongxing Luo, Philip Westphal, Paul Leeson, Paolo DiAchille, Viatcheslav Gurev, Manuel Mayr, Liesbet Geris, Pras Pathmanathan, Tina Morrison, Richard Cornelussen, Frits Prinzen, Tammo Delhaas, Ada Doltra, Marta Sitges, Edward J Vigmond, Ernesto Zaccur, Vicente Grau, Blanca Rodriguez, Espen W Remme, Steven Niederer, Peter Mortier, Kristin McLeod, Mark Potse, Esther Pueyo, Alfonso Bueno-Orovio, and Pablo Lamata. The ‘Digital Twin’ to enable the vision of precision cardiology. *European Heart Journal*, 41(48):4556–4564, 2020. ISSN 0195-668X. doi: 10.1093/eurheartj/ehaa159. URL <https://doi.org/10.1093/eurheartj/ehaa159>.
- David R. Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.
- Ulrich Richard Dahmen, Tobias Osterloh, and Heinz-Jürgen Roßmann. Verification and validation of digital twins and virtual testbeds. *International Journal of Advances in Applied Sciences*, 11(1):47–64, 2022. doi: 10.11591/ijaas.v11.i1.pp47-64. URL <https://publications.rwth-aachen.de/record/843535>.
- Anthony C. Davison and David V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997. doi: 10.1017/CBO9780511802843.
- Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552. URL <https://doi.org/10.1214/aos/1176344552>.
- Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, New York, 1993.
- Suran Galappaththige, Richard A. Gray, Caroline Mendonca Costa, Steven Niederer, and Pras Pathmanathan. Credibility assessment of patient-specific computational modeling using patient-specific cardiac modeling as an exemplar. *PLoS Computational Biology*, 18(10):e1010541, 2022.
- Michael Grieves and John Vickers. Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In *Transdisciplinary Perspectives on Complex Systems*, pages 85–113. Springer, 2017.

- Peter Hall. Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, 16(3):927–953, 1988. doi: 10.1214/aos/1176350933. URL <https://doi.org/10.1214/aos/1176350933>.
- André Hemmler, Brigitta Lutz, Günay Kalender, Christian Reeps, and Michael W. Gee. Patient-specific in silico endovascular repair of abdominal aortic aneurysms: Application and validation. *Biomechanics and Modeling in Mechanobiology*, 18(4):983–1004, 2019.
- Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman and Hall/CRC, Boca Raton, 2020.
- Tim C. Hesterberg. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4):371–386, 2015. doi: 10.1080/00031305.2015.1089789. URL <https://doi.org/10.1080/00031305.2015.1089789>. PMID: 27019512.
- Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 01621459. URL <http://www.jstor.org/stable/2289064>.
- Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469. URL <http://www.jstor.org/stable/4615733>.
- Zeshan M. Hussain, Michael Oberst, Ming-Chieh Shih, and David Sontag. Falsification before extrapolation in causal effect estimation. In *Advances in Neural Information Processing Systems*, volume 35, pages 6161–6174, 2022.
- Guido W. Imbens and Charles F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.
- Guido W. Imbens and Jeffrey M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, 2009. ISSN 00220515. URL <http://www.jstor.org/stable/27647134>.
- Melanie Jans-Singh, Kathryn Leeming, Ruchi Choudhary, and Mark Girolami. Digital twin of an urban-integrated hydroponic farm. *Data-Centric Engineering*, 1:e20, 2020. doi: 10.1017/dce.2020.21.
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. doi: 10.1038/sdata.2016.35. URL <https://doi.org/10.1038/sdata.2016.35>.
- David Jones, Chris Snider, Aydin Nassehi, Jason Yon, and Ben Hicks. Characterising the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29:36–52, 2020.

- Michael G. Kapteyn, Jacob V. R. Pretorius, and Karen E. Willcox. A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nature Computational Science*, 1(5):337–347, 2021.
- Adnan Khan, Martin Dahl, Petter Falkman, and Martin Fabian. Digital twin for legacy systems: Simulation model testing and validation. In *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, pages 421–426. IEEE, 2018.
- Toru Kitagawa. A test for instrument validity. *Econometrica*, 83(5):2043–2063, 2015. doi: <https://doi.org/10.3982/ECTA11974>. URL <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA11974>.
- Brendan Kochunas and Xun Huan. Digital twin concepts with uncertainty for nuclear power applications. *Energies*, 14(14):4235, 2021.
- Matthieu Komorowski, Leo Anthony Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. doi: 10.1038/s41591-018-0213-5. URL <https://doi.org/10.1038/s41591-018-0213-5>.
- Amos Lal, Guangxi Li, Edin Cubro, Sarah Chalmers, Heyi Li, Vitaly Herasevich, Yue Dong, Brian Pickering, Oguz Kilickaya, and Ognjen Gajic. Development and verification of a digital twin patient model to predict treatment response in sepsis. *Critical Care Medicine*, 49(1):611, 2021. doi: 10.1097/01.ccm.0000730744.82258.38.
- Ignacio Larrabide, Minsuok Kim, Luca Augsburg, Maria Cruz Villa-Uriol, Daniel Rüfenacht, and Alejandro F. Frangi. Fast virtual deployment of self-expandable stents: Method and in vitro evaluation for intracranial aneurysmal stenting. *Medical Image Analysis*, 16(3):721–730, 2012. ISSN 1361-8415. doi: 10.1016/j.media.2010.04.009. URL <https://doi.org/10.1016/j.media.2010.04.009>.
- Philip W. Lavori and Ree Dawson. Dynamic treatment regimes: Practical design considerations. *Clinical Trials*, 1(1):9–20, 2004.
- Yuqian Lu, Chao Liu, Kevin I-Kai Wang, Huiyue Huang, and Xun Xu. Digital twin-driven smart manufacturing: Connotation, reference model, applications and research issues. *Robotics and Computer-Integrated Manufacturing*, 61:101837, 2020.
- Charles F. Manski. Anatomy of the selection problem. *The Journal of Human Resources*, 24(3):343–360, 1989. ISSN 0022166X. URL <http://www.jstor.org/stable/145818>.
- Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990. ISSN 00028282. URL <http://www.jstor.org/stable/2006592>.
- Charles F. Manski. *Identification Problems in the Social Sciences*. Harvard University Press, 1995.
- Charles F. Manski. Monotone treatment response. *Econometrica*, 65(6):1311–1334, 1997. ISSN 00129682, 14680262. URL <http://www.jstor.org/stable/2171738>.

- Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.
- Charles F. Manski. *Identification for Prediction and Decision*. Harvard University Press, 2009.
- Charles F. Manski and John V. Pepper. Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, 68(4):997–1010, 2000. ISSN 00129682, 14680262. URL <http://www.jstor.org/stable/2999533>.
- J. Masison, J. Beezley, Y. Mei, H. A. L. Ribeiro, A. C. Knapp, L. Sordo Vieira, B. Adhikari, Y. Scindia, M. Grauer, B. Helba, W. Schroeder, B. Mehrad, and R. Laubenbacher. A modular computational framework for medical digital twins. *Proceedings of the National Academy of Sciences*, 118(20):e2024287118, 2021.
- Matthew McDaniel and Austin Baird. A full-body model of burn pathophysiology and treatment using the BioGears Engine. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 261–264, 2019. doi: 10.1109/EMBC.2019.8857686.
- Matthew McDaniel, Jonathan M. Keller, Steven White, and Austin Baird. A whole-body mathematical model of sepsis progression and treatment designed in the BioGears Physiology Engine. *Frontiers in Physiology*, 10:1321, 2019. ISSN 1664-042X. doi: 10.3389/fphys.2019.01321. URL <https://www.frontiersin.org/article/10.3389/fphys.2019.01321>.
- Francesca Molinari. Microeconometrics with partial identification. In Steven N. Durlauf, Lars Peter Hansen, James J. Heckman, and Rosa L. Matzkin, editors, *Handbook of Econometrics, Volume 7A*, volume 7 of *Handbook of Econometrics*, pages 355–486. Elsevier, 2020. doi: <https://doi.org/10.1016/bs.hoe.2020.05.002>. URL <https://www.sciencedirect.com/science/article/pii/S1573441220300027>.
- Ismael Mourifié and Yuanyuan Wan. Testing local average treatment effect assumptions. *The Review of Economics and Statistics*, 99(2):305–313, 2017. ISSN 0034-6535.
- Susan A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.
- Susan A. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1481, 2005. doi: <https://doi.org/10.1002/sim.2022>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2022>.
- Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In *Advances in Neural Information Processing Systems*, volume 33, pages 18819–18831, 2020. URL <https://proceedings.neurips.cc/paper/2020/file/da21bae82c02d1e2b8168d57cd3fbab7-Paper.pdf>.
- Steven A. Niederer, Michael S. Sacks, Mark Girolami, and Karen Willcox. Scaling digital twins from the artisanal to the industrial. *Nature Computational Science*, 1(5):313–320, 2021.

- Judea Pearl. *Causality*. Cambridge University Press, second edition, 2009. doi: 10.1017/CBO9780511803161.
- Karl Popper. *The Logic of Scientific Discovery*. Routledge, 2005.
- James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9–12):1393–1512, 1986.
- Joseph P. Romano and Azeem M. Shaikh. Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, 138(9): 2786–2807, 2008. ISSN 0378-3758. doi: <https://doi.org/10.1016/j.jspi.2008.03.015>. URL <https://www.sciencedirect.com/science/article/pii/S0378375808000694>. Special Issue in Honor of Theodore Wilbur Anderson, Jr. on the Occasion of his 90th Birthday.
- Joseph P. Romano and Azeem M. Shaikh. Inference for the identified set in partially identified econometric models. *Econometrica*, 78(1):169–211, 2010. doi: <https://doi.org/10.3982/ECTA6706>. URL <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA6706>.
- Joseph P. Romano, Azeem M. Shaikh, and Michael Wolf. A practical two-step method for testing moment inequalities. *Econometrica*, 82(5):1979–2002, 2014. ISSN 00129682, 14680262. URL <http://www.jstor.org/stable/24029299>.
- Adam M. Rosen. Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. *Journal of Econometrics*, 146(1):107–117, 2008. ISSN 0304-4076. doi: <https://doi.org/10.1016/j.jeconom.2008.08.001>. URL <https://www.sciencedirect.com/science/article/pii/S0304407608000791>.
- Paul R. Rosenbaum. *Observational Studies*. Springer, New York, NY, 2002.
- Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. ISSN 00063444. URL <http://www.jstor.org/stable/2335942>.
- Christopher J. Roy and William L. Oberkampf. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25):2131–2144, 2011. ISSN 0045-7825. doi: <https://doi.org/10.1016/j.cma.2011.03.016>. URL <https://www.sciencedirect.com/science/article/pii/S0045782511001290>.
- Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974. doi: <https://doi.org/10.1037/h0037350>.
- Donald B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005. doi: 10.1198/016214504000001880. URL <https://doi.org/10.1198/016214504000001880>.

- Rafael Sacks, Ioannis Brilakis, Ergo Pikas, Haiyan Sally Xie, and Mark Girolami. Construction with digital twin information systems. *Data-Centric Engineering*, 1:e14, 2020.
- Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S. Deutschman, Gabriel J. Escobar, and Derek C. Angus. Assessment of clinical criteria for sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):762–774, 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.0288. URL <https://doi.org/10.1001/jama.2016.0288>.
- Mervyn Singer, Clifford S. Deutschman, Christopher W. Seymour, Manu Shankar-Hari, Djilali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):801–810, 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.0287. URL <https://doi.org/10.1001/jama.2016.0287>.
- Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.
- Anastasios A. Tsiatis, Marie Davidian, Shannon T. Holloway, and Eric B. Laber. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. Chapman and Hall/CRC, 2019.
- Eric J. Tuegel, Anthony R. Ingraffea, Thomas G. Eason, and S. Michael Spottswood. Reengineering aircraft structural life prediction using a digital twin. *International Journal of Aerospace Engineering*, 2011:154798, 2011. doi: 10.1155/2011/154798. URL <https://doi.org/10.1155/2011/154798>.
- Steve Yadowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *The Annals of Statistics*, 50(5):2587–2615, 2022. doi: 10.1214/22-AOS2195. URL <https://doi.org/10.1214/22-AOS2195>.
- Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL <https://proceedings.neurips.cc/paper/2019/file/8252831b9fce7a49421e622c14ce0f65-Paper.pdf>.
- Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26548–26558. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/zhang22ab.html>.