Causal Falsification of Digital Twins

Rob Cornish*[1], Muhammad Faaiz Taufiq*[1], Arnaud Doucet[1], and Chris Holmes[1,2]

[1]Department of Statistics, University of Oxford, UK
[2]The Alan Turing Institute, UK

Abstract

Digital twins are virtual systems 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 causal inference problem, which leads to a precise definition of what it means for a twin to be "correct" appropriate for many applications. Unfortunately, fundamental results from causal inference mean observational data cannot be used to certify that a twin is correct 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, unmeasured confounding, hypothesis testing

*Both authors contributed equally to this work.

1 Introduction

A *digital twin* is a virtual system that runs alongside some physical process of interest and mimics its behaviour over time in response to user-provided interventions. This modelling approach promises to transform planning and decision-making across a breadth of fields, since it allows actions to be chosen with a fuller understanding of the range of possible outcomes that may result. Early research on digital twins has considered use-cases including aviation [5], manufacturing [32], healthcare [9, 8], civil engineering [50], climate modelling [4], and agriculture [22]. For a comprehensive overview of this emerging technology, we refer the reader to [3, 24, 41].

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 [41]. 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.

A principal design objective for digital twins is to improve decision-making, with much emphasis placed on their ability to reveal choices of actions that are expected to produce some desirable outcome in the real world [3, 24, 41]. We therefore adopt the position that, to be considered "accurate", a twin must capture the range of outcomes that *would* occur when certain actions or interventions are applied to the real-world process of interest in a controlled way. Accordingly, we formulate twin assessment as a problem of *causal inference* [48, 49, 42, 17], which offers a particularly suitable framework for reasoning about interventional behaviour of this kind.

The primary reason for assessing the accuracy of a twin is to determine its reliability and robustness. It is therefore desirable for any assessment procedure itself to be reliable and robust, and for any conclusions it draws 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 highly general, and so may be applied to a wide variety of twins across application domains.

Contribution Our contribution has the following components:

- We show that if interventional information is sought from a twin, then it is not possible to use observational data to *certify* that the twin is 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 advocate for an assessment paradigm instead based on *falsification*: we aim to find specific examples of where the twin is *not* accurate, rather than trying to quantify its overall accuracy in a holistic sense.

- We propose a general-purpose statistical testing procedure for falsifying a twin that relies on only the assumption of an independent and identically distributed (i.i.d.) dataset of observational trajectories. In particular, the procedure 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. Key to our approach is a novel longitudinal form of Manski's classical bounds [33] from the causal inference literature that may also be of interest in other contexts (see Section 4).

- We demonstrate the effectiveness of our procedure through a large-scale, real-world case study in which we use the MIMIC-III ICU dataset [23] to assess the Pulse Physiology Engine [7], a high-fidelity, open-source model for human physiology simulation.

Related work Various high-level guidelines and workflows have been proposed for the assessment of digital twins in the literature to-date [14, 26, 9, 27, 41, 11], as well as in the field of computational science more broadly [47, 41]. In some cases, these guidelines have been codified as standards: for example, the ASME V&V40 Standard [2] 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 [13]. 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 [9, 41, 25, 36]. 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 [7], which we consider in our empirical case study, as well as the related BioGears [38, 37] were both assessed by comparing their outputs with ad hoc values based either on available medical literature or 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 [30, 16, 29, 22, 13]. 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, our paper is also the first to identify the need for a causal approach to twin assessment and the pitfalls that arise when causal considerations are not properly accounted for.

Outline In Section 2 we introduce precise causal models for the twin and the real-world process. In Section 3, we distinguish between certification and falsification assessment procedures, showing that certification is in general unsound and advocating for falsification

as a more robust strategy. In Section 4, we introduce and analyse a novel causal bound that in Section 5 provides the basis for a hypothesis testing procedure for twin falsification with exact, finite-sample guarantees of validity under only the assumption of i.i.d. observational data. Section 6 contains the results of our empirical case study.

2 Causal formulation

Here we provide causal models for the twin and the corresponding real-world process that the twin is designed to simulate. We do so in the language of *potential outcomes* [48, 49], although we note that we could have used the alternative framework of directed acyclic graphs and structural causal models [42] (see also [20] for a comparison of the two).

2.1 The real-world process

We assume the real-world process operates over a fixed time horizon $T \in \{1, 2, \ldots\}$. This simplifies our presentation in what follows, and it is straightforward to generalise our methodology to variable length time horizons if needed. For each $t \in \{0, \ldots, T\}$, we assume the process gives rise to a *observation* at time t, which takes values in some real-valued space $\mathcal{X}_t \coloneqq \mathbb{R}^{d_t}$. We also assume that the process can be influenced by some *action* taken at each time $t \in \{1, \ldots, T\}$. 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 $\mathcal{A}_1 \times \cdots \times \mathcal{A}_t$, and $a_{1:t} \in \mathcal{A}_{1:t}$ is a choice of $a_1 \in \mathcal{A}_1, \ldots, a_t \in \mathcal{A}_t$.

We model the dynamics of the real-world process via the longitudinal potential outcomes framework proposed by Robins [44], 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* $X_0, X_1(a_1), \ldots, X_T(a_{1:T})$, 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. 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 both to specific actions chosen from $\mathcal{A}_{1:T}$ as well as additional external 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 in turn be used to guide planning and decision-making downstream.

Precisely, we model the output 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 (possibly vector-valued) random variable. We will denote $\widehat{X}_t(x_0, a_{1:t}) \coloneqq h_t(x_0, a_{1:t}, U_{1:t})$, which we also refer to as a potential outcome. A full twin trajectory therefore consists of $\widehat{X}_1(x_0, a_1), \dots, \widehat{X}_T(x_0, a_{1:T})$, which we write more compactly by $\widehat{X}_{1:T}(x_0, a_{1:T})$. 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 random numbers $U_{1:T}$ and the real-world outcomes $(X_{0:T}(a_{1:T}) : a_{1:T} \in \mathcal{A}_{1:T})$ 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.

Correctness We assume the twin has been designed to be correct in the following sense:

Definition 1. *The twin is* interventionally correct *if, for* $\mathrm{Law}[X_0]$*-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*[1] *to the conditional distribution of* $X_{1:T}(a_{1:T})$ *given* $X_0 = x_0$*.*

Operationally, if this holds, 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

[1] More generally we could allow for inequality up to some specified tolerance level. However, our stricter condition here is simpler, and suffices to convey the main ideas in what follows.

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 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 Supplement 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* $X_{0:T}(a_{1:T})$*.*

Online prediction Our model here represents a twin at time $t = 0$ making predictions about all future timesteps $t \in \{1, \ldots, T\}$ 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 [14, 41]. Various notions of correctness can be devised for this online setting. We describe two possibilities in Section C of the Supplement, 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

We now consider how to assess the accuracy of a twin. There are many conceivable methods for doing this, 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 [2, 41, 13]. However, in this paper, we focus on what we will call *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 data obtained from the real-world process and trajectories simulated by the twin. We show in this section that, without further assumptions, it is not possible to obtain a data-driven assessment procedure that can *certify* that a twin is interventionally correct. We instead propose a strategy based on *falsifying* the twin, which we develop into a concrete statistical testing procedure in later sections.

3.1 Certification is unsound in general

Data model We 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 $t \in \{1, \ldots, T\}$ 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:

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

This corresponds to the standard *consistency* assumption in causal inference [17], and intuitively means that 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).

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

$$\text{If } \mathcal{H} \text{ is true, then the twin is interventionally correct.}^2 \tag{2}$$

Then data is 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* [18]) means that data-driven certification procedures of this kind are in general unsound, as we explain next. For completeness, Section D of the Supplement 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 $(X_0, \widehat{X}_T(X_0, a_{1:T}))$. 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 $(X_0, \widehat{X}_{1:T}(X_0, a_{1:T}))$ 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

[2] More generally, and in practice more typically, we could consider a family of hypotheses that each imply correctness up to some tolerance level. However, (2) suffices for our exposition here.

by the behavioural agent solely on the basis of the information available at time t that is actually recorded in the dataset, namely $X_0, A_1, X_1(A_1), \ldots, A_{t-1}, X_{t-1}(A_{1:t-1})$, as well as possibly some additional randomness that is independent of the real-world process, such as the outcome of a coin toss.[3] 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, in certain situations 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 [31, 39], constitute a widespread example of this approach. Likewise, it is possible to show[4] 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. However, for stochastic phenomena and for typical datasets, it is widely acknowledged that the assumption of no unmeasured confounding will rarely hold, and so assessment procedures based on this assumption may yield unreliable results in practice [40, 53]. Section F of the Supplement 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 [43], 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), namely:

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

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

[3] This can be made precise via the *sequential randomisation assumption* (also known as *sequential ignorability*) introduced by Robins [44]. Chapter 5 of [53] provides an overview of this.

[4] See Section E of the Supplement for a proof.

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 Causal bounds

We now describe a novel longitudinal variant of Manski's classical bounds [33] from the causal inference literature. Intuitively, this result bounds the expected outcome output by any interventionally correct twin. In the next section, we use this result to define hypotheses $\mathcal{H}$ with the desired property (3), which then yields a procedure for falsifying twins via hypothesis testing techniques. We state the result before providing intuition below. A proof is given in Section G of the Supplement.

Theorem 4. *Suppose* $(Y(a_{1:t}) : a_{1:t} \in \mathcal{A}_{1:t})$ *are real-valued potential outcomes, and 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*

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

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

Then it holds that

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

where we define $N \coloneqq \max\{0 \leq s \leq t \mid A_{1:s} = a_{1:s}\}$, *and similarly*

$$Y_{\text{lo}} \coloneqq \mathbb{1}(A_{1:t} = a_{1:t})\, Y(A_{1:t}) + \mathbb{1}(A_{1:t} \neq a_{1:t})\, y_{\text{lo}}, \quad Y_{\text{up}} \coloneqq \mathbb{1}(A_{1:t} = a_{1:t})\, Y(A_{1:t}) + \mathbb{1}(A_{1:t} \neq a_{1:t})\, y_{\text{up}}.$$

We will write the terms in (6) as Q_{lo}, Q, and Q_{up} respectively, so we have $Q_{\text{lo}} \leq Q \leq Q_{\text{up}}$.

Intuition Here $Y(a_{1:t})$ 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 $Y(a_{1:t}) \coloneqq f(X_{0:t}(a_{1:t}))$, where $f : \mathcal{X}_{0:t} \to \mathbb{R}$ is some scalar function. The value Q is then simply the (conditional) average behaviour of this outcome. By Theorem 3, Q is in general not identified by the data since it depends on $X_{0:t}(a_{1:t})$. On the other hand, both Q_{lo} and Q_{up} *are* identifiable, since the relevant random variables Y_{lo}, Y_{up}, N, and $X_{0:N}(A_{1:N})$ 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 when $A_{1:t} = a_{1:t}$, 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. We emphasise that Theorem 4 does not require any additional causal assumptions, and in particular remains true under arbitrary unmeasured confounding.

Related results Theorem 4 is as a longitudinal generalisation of Manski's classical bounds from [33].[5] In the structural causal modelling framework [42], a related bound was recently given by Zhang & Bareinboim as Corollary 1 in [55]. 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 the ratio term. In contrast, our proposed causal 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.

Tightness of bounds Straightforward manipulations yield the following:

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

Here the left-hand side quantifies the *tightness* of the bounds $[Q_{\text{lo}}, Q_{\text{up}}]$ relative to the worst-case bounds $[y_{\text{lo}}, y_{\text{up}}]$ that are trivially implied by (5), with smaller values meaning tighter bounds. In this way, the tightness of the 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 closely related to the classical *propensity score* in the causal inference literature [46]. Intuitively, as $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$ grows larger, $Y(a_{1:t}) = Y(A_{1:t})$ holds on the event $\{X_{0:N}(A_{1:N}) \in B_{0:N}\}$ with higher probability, so that the effect of unmeasured confounding on the overall expectation $Q = \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}]$ is reduced, and Q may be bounded more tightly.

Sharpness of bounds The following result shows that Theorem 4 cannot be improved without further assumptions. Intuitively, the result states that there always exists *some* potential outcomes family which leads to the same observational data as our model, but which attains the worst-case bounds Q_{lo} or Q_{up}. Therefore, we cannot rule out the possibility that our model achieves Q_{lo} or Q_{up} from the observational data alone.

Proposition 5. *Under the same setup as in Theorem 4, 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* (5) *(mutatis mutandis) with* $(\tilde{X}_{0:T}(A_{1:T}), \tilde{Y}(A_{1:t}), A_{1:T}) \overset{\text{a.s.}}{=} (X_{0:T}(A_{1:T}), Y(A_{1:t}), A_{1:T})$ *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}.

Pointwise conditional bounds In some cases, it may be desirable to obtain bounds on the alternative quantity $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})]$, i.e. conditional 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 is natural to generalise our assumption (5) by supposing we have measurable functions $y_{\text{lo}}, y_{\text{up}} : \mathcal{X}_{0:t} \to \mathbb{R}$ such that

$$y_{\text{lo}}(X_{0:t}(a_{1:t})) \leq Y(a_{1:t}) \leq y_{\text{up}}(X_{0:t}(a_{1:t})) \qquad \text{almost surely.} \tag{7}$$

When this holds, almost sure bounds on $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})]$ can be obtained directly from Theorem 4 if $X_{0:t}(a_{1:t})$ is discrete, or by a simple modification of the proof of Theorem 4 if $X_{1:t}(a_{1:t})$ is discrete (but X_0 is possibly continuous), which recovers Manski's original bounds from [33] as a special case. We discuss this further in Section G.3 of the Supplement.

[5] We make this relationship precise in Section G.3.2 of the Supplement.

However, somewhat surprisingly, if the distribution of $X_{1:t}(a_{1:t})$ is not discrete, then we cannot obtain nontrivial bounds of this kind without further assumptions. To make this precise, we will say that candidate bounds $g_{\text{lo}}, g_{\text{up}} : \mathcal{X}_{0:t} \to \mathbb{R}$ are *permissible* if

$$g_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t})) \leq \mathbb{E}[\tilde{Y}(a_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t})] \leq g_{\text{up}}(\tilde{X}_{0:t}(a_{1:t})) \qquad \text{almost surely} \tag{8}$$

for all potential outcomes $(\tilde{X}_{0:T}(a'_{1:T}), \tilde{Y}(a'_{1:t}), \tilde{A}_{1:T} : a'_{1:T} \in \mathcal{A}_{1:T})$ that also satisfy (7) (mutatis mutandis) with $\text{Law}[\tilde{X}_{0:T}(\tilde{A}_{1:T}), \tilde{A}_{1:T}, \tilde{Y}(\tilde{A}_{1:t})] = \text{Law}[X_{0:T}(A_{1:T}), A_{1:T}, Y(A_{1:t})]$. Intuitively, this means that g_{lo} and g_{up} depend only on the information we have available, i.e. our assumed worst-case values and the observational distribution. We then have the following:

Theorem 6. *Suppose X_0 is almost surely constant, $\mathbb{P}(A_1 \neq a_1) > 0$, and for some $s \in \{1, \ldots, t\}$ we have $\mathbb{P}(X_s(A_{1:s}) = x_s) = 0$ for all $x_s \in \mathcal{X}_s$. Then $g_{\text{lo}}, g_{\text{up}} : \mathcal{X}_{0:t} \to \mathbb{R}$ are permissible bounds only if they are trivial, i.e. $g_{\text{lo}}(X_{0:t}(a_{1:t})) \leq y_{\text{lo}}(X_{0:t}(a_{1:t}))$ and $g_{\text{up}}(X_{0:t}(a_{1:t})) \geq y_{\text{up}}(X_{0:t}(a_{1:t}))$ almost surely.*

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 6 holds more generally, provided the other assumptions are made conditional on X_0 accordingly. 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 4 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

We now use Theorem 4 to define a hypothesis testing methodology that can be used to falsify the twin. We first define the hypotheses $\mathcal{H}$ satisfying (3) that we will consider, and then give a statistical procedure for testing these while controlling for type I error. Our overall procedure does not rely on any further assumptions than we have already provided.

5.1 Hypotheses derived from causal bounds

Each hypothesis we test will depend on a specific choice of the following parameters:

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

To streamline notation, in this section, we will consider these parameters to be fixed. However, we emphasise 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 $Y(a'_{1:t}) \coloneqq f(X_{0:t}(a'_{1:t}))$.

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.

Hypothesis definition The hypotheses we consider are based on the corresponding outcome produced by the twin when initialised at X_0, which we define for $a'_{1:t} \in \mathcal{A}_{1:t}$ as $\widehat{Y}(a'_{1:t}) \coloneqq f(X_0, \widehat{X}_{1:t}(X_0, a'_{1:t}))$. Supposing it holds that

$$\mathbb{P}(\widehat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}) > 0, \tag{9}$$

we may then define $\widehat{Q} \coloneqq \mathbb{E}[\widehat{Y}(a_{1:t}) \mid X_0 \in B_0, \widehat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}]$, i.e. the analogue of Q for the twin. By Proposition 2, if the twin is interventionally correct, then $\widehat{Q} = Q$. Theorem 4 therefore implies that the following hypotheses have our desired property (3):

$$\mathcal{H}_{\text{lo}}\text{: If (4), (5), and (9) hold, then } \widehat{Q} \geq Q_{\text{lo}}$$
$$\mathcal{H}_{\text{up}}\text{: If (4), (5), and (9) hold, then } \widehat{Q} \leq Q_{\text{up}}.$$

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 $\widehat{Q}$ can be estimated by generating trajectories from the twin.

Interpreting a falsification From a practical perspective, falsifying either $\mathcal{H}_{\text{lo}}$ or $\mathcal{H}_{\text{up}}$ leads to knowledge about a specific failure mode of the twin with various potential implications downstream. For example, if $\mathcal{H}_{\text{lo}}$ is false (i.e. if $\widehat{Q} < Q_{\text{lo}}$), it follows that, among those trajectories for which $(X_0, \widehat{X}_{1:t}(X_0, a_{1:t})) \in B_{0:t}$, the corresponding value of $\widehat{Y}(a_{1:t})$ is on average too small. 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. We illustrate this concretely through our case study in Section 6.

5.2 Exact testing procedure

We now describe a procedure for testing $\mathcal{H}_{\text{lo}}$ and $\mathcal{H}_{\text{up}}$ using a finite dataset that obtains exact control over type I error without relying on additional assumptions or asymptotic approximations. We show in our case study below that this procedure is nevertheless powerful enough to obtain useful information about a twin in practice.

We focus here on obtaining 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 Holm-Bonferroni [19] method (which we employ in our case study) or the Benjamini–Yekutieli [6] procedure, both of which are valid without additional assumptions.

Datasets 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 $\widehat{\mathcal{D}}(a_{1:t})$ of i.i.d. copies of

$$X_0, \widehat{X}_1(X_0, a_1), \ldots, \widehat{X}_t(X_0, a_{1:t}). \tag{10}$$

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 initialise the twin at most once, then the resulting trajectories in $\widehat{\mathcal{D}}(a_{1:t})$ are guaranteed to be i.i.d., since we assumed in Section 2 that the potential outcomes $\widehat{X}_t(x_0, a_{1:t})$ 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 (9) all hold. We account for this in our testing procedure as follows. First, (5) immediately follows if

$$y_{\text{lo}} \leq f(x_{0:t}) \leq y_{\text{up}} \qquad \text{for all } x_{0:t} \in B_{0:t}. \tag{11}$$

This holds automatically in certain cases, such as when dealing with a binary outcome (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 scheme for choosing f in this way in Section 6.

To account for (4) and (9), our first step in our testing process is simply to check whether there is some trajectory in $\mathcal{D}$ with $A_{1:t} = a_{1:t}$ and $X_{0:t}(A_{1:t}) \in B_{0:t}$, and some trajectory in $\widehat{\mathcal{D}}(a_{1:t})$ with $(X_0, \widehat{X}_{1:t}(X_0, a_{1:t})) \in B_{0:t}$. If there are not, then we refuse to reject $\mathcal{H}_{\text{lo}}$ at any significance level; otherwise, we proceed to test $\widehat{Q} \geq Q_{\text{lo}}$ as described next. It easily follows that there is zero probability we will reject $\mathcal{H}_{\text{lo}}$ if (4) and (9) do not in fact hold, and so our overall type I error is controlled at the desired level.

Testing procedure To test $\widehat{Q} \geq Q_{\text{lo}}$, we begin by constructing a one-sided lower confidence interval for Q_{lo}, and a one-sided upper confidence interval for $\widehat{Q}$. In detail, for each significance level $\alpha \in (0, 1)$, we obtain R^{α}_{lo} and $\widehat{R}^{\alpha}$ as functions of $\mathcal{D}$ and $\widehat{\mathcal{D}}(a_{1:t})$ such that

$$\mathbb{P}(Q_{\text{lo}} \geq R^{\alpha}_{\text{lo}}) \geq 1 - \frac{\alpha}{2} \qquad \mathbb{P}(\widehat{Q} \leq \widehat{R}^{\alpha}) \geq 1 - \frac{\alpha}{2}. \tag{12}$$

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

From these confidence intervals, we obtain a test for the hypothesis $\widehat{Q} \geq Q_{\text{lo}}$ that rejects when $\widehat{R}^{\alpha} < R^{\alpha}_{\text{lo}}$. A straightforward argument given in Section H.1 of the Supplement 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. $p_{\text{lo}} \coloneqq \inf\{\alpha \in (0, 1) \mid \widehat{R}^{\alpha} < R^{\alpha}_{\text{lo}}\}$, or 1 if the test does not reject at any level.

Confidence intervals We now consider how to obtain confidence intervals for Q_{lo} and $\widehat{Q}$ 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 $\widehat{\mathcal{D}}(a_{1:t})$.

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 $\widehat{\mathcal{D}}(a_{1:t})$ with $(X_0, \widehat{X}_t(X_0, a_{1:t})) \in B_{0:t}$. For each such trajectory, we obtain a corresponding value of $\widehat{Y}(a_{1:t})$ that is i.i.d. and has expectation $\widehat{Q}$. 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 $\widehat{Y}(a_{1:t})$ and Y_{lo} are bounded in $[y_{\text{lo}}, y_{\text{up}}]$, which gives rise to R^{α}_{lo} and $\widehat{R}^{\alpha}$ via an application of Hoeffding's inequality. This approach has the appealing property that (12) holds exactly, although often at the expense of conservativeness. In practice, this could be mitigated by instead obtaining confidence intervals via (for example) the bootstrap [12], although at the expense of requiring (often mild) asymptotic assumptions. Section H.2 of the Supplement describes both methods in greater detail. Our empirical results reported in the next section all use Hoeffding's inequality are hence exact, but we also provide some additional results obtained using bootstrapping in Section I.6 of the Supplement.

6 Case Study: Pulse Physiology Engine

We applied our assessment methodology to the Pulse Physiology Engine [7], 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 the Section I of the Supplement.

6.1 Experimental setup

Real-world data We utilized the MIMIC-III dataset [23], 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 [51], following the methodology of [28] 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 [10], with the remaining 95% (11,094 trajectories, denoted as $\mathcal{D}$) reserved for the actual testing.

Observation and action spaces We considered hourly observations of each patient over the 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 [28], the actions we considered were the administration of intravenous fluids and vasopressors, which both play a primary role in the treatment of sepsis in clinical practice. Since these are

Physiological quantity	# Rejections	# Hypotheses
Chloride Blood Concentration (Chloride)	24	94
Sodium Blood Concentration (Sodium)	21	94
Potassium Blood Concentration (Potassium)	13	94
Skin Temperature (Temp)	10	86
Calcium Blood Concentration (Calcium)	5	88
Glucose Blood Concentration (Glucose)	5	96
Arterial CO_2 Pressure ($paCO_2$)	3	70
Bicarbonate Blood Concentration (HCO_3)	2	90
Systolic Arterial Pressure (SysBP)	2	154

Table 1: Total hypotheses and rejections per physiological quantity

recorded in MIMIC as continuous doses, we discretised their values via the same procedure as [28], obtaining finite action spaces $\mathcal{A}_1 = \cdots = \mathcal{A}_4$, each with 25 distinct 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 Supplement for a complete list). For each of these, we selected combinations of t, $a_{1:t}$, and $B_{0:t}$ 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[6] of the same physiological quantity as 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 (11) holds. We describe this procedure in full in Section I.4 of the Supplement. We obtained 721 unique parameter choices, leading to 1,442 total hypotheses.

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 initialise a twin trajectory. Then, at each hour $t' \in \{1, \ldots, 4\}$ 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.5 in the Supplement. This produced 26,115 total simulated trajectories.

6.2 Results

We tested these hypotheses 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 $\widehat{Q}$. We also tried confidence intervals obtained via bootstrapping, and obtained similar if less conservative results (see Section I of the Supplement). We used the Holm-Bonferroni method to adjust for multiple testing, with a family-wise error rate of 0.05.

[6]We also investigated the sensitivity of our procedure to the choice of these quantiles and found it to be relatively stable (see Section I.8 of the Supplement).

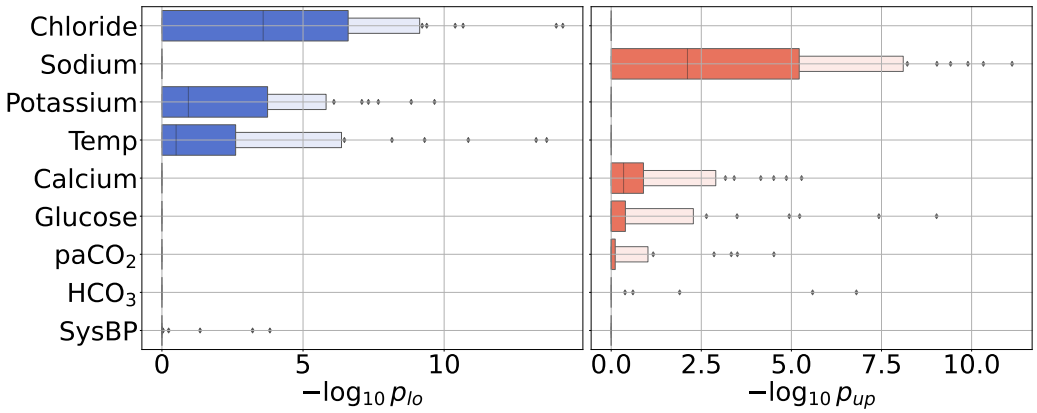


Figure 1: Distributions of $-\log_{10} p_{\text{lo}}$ and $-\log_{10} p_{\text{up}}$ across hypotheses, grouped by physiological quantity. Higher values indicate greater evidence in favour of rejection.

Hypothesis rejections We obtained rejections for hypotheses corresponding to 10 different physiological quantities shown in Table 1. (Table 4 in the Supplement 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 to inform their decision-making may wish to rely less on these outputs as a result.

***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, 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 was rejected. (The remaining p-values are shown in Figure 7 in the Supplement.) 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 $\mathcal{H}_{\text{lo}}$, or very little in favour of rejecting any $\mathcal{H}_{\text{up}}$. 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.

6.3 Pitfalls of naive assessment

A naive approach to twin assessment involves simply comparing the output of the twin with 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 $\widehat{Q}_t$ and

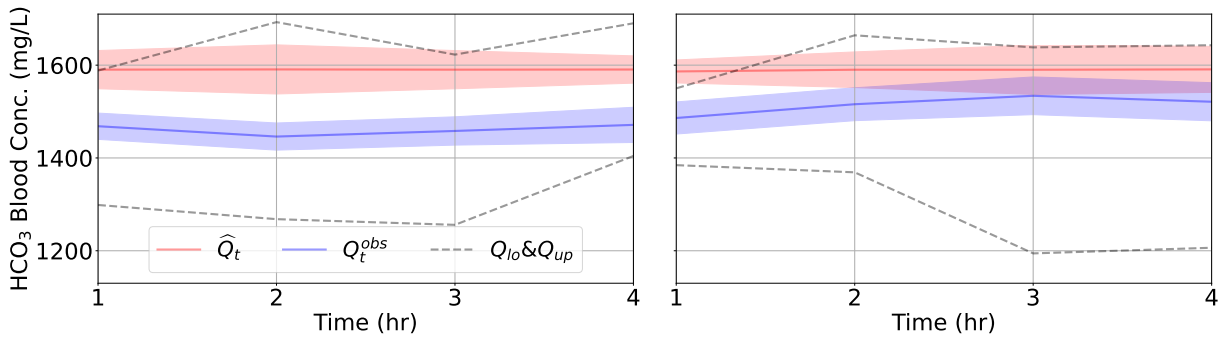


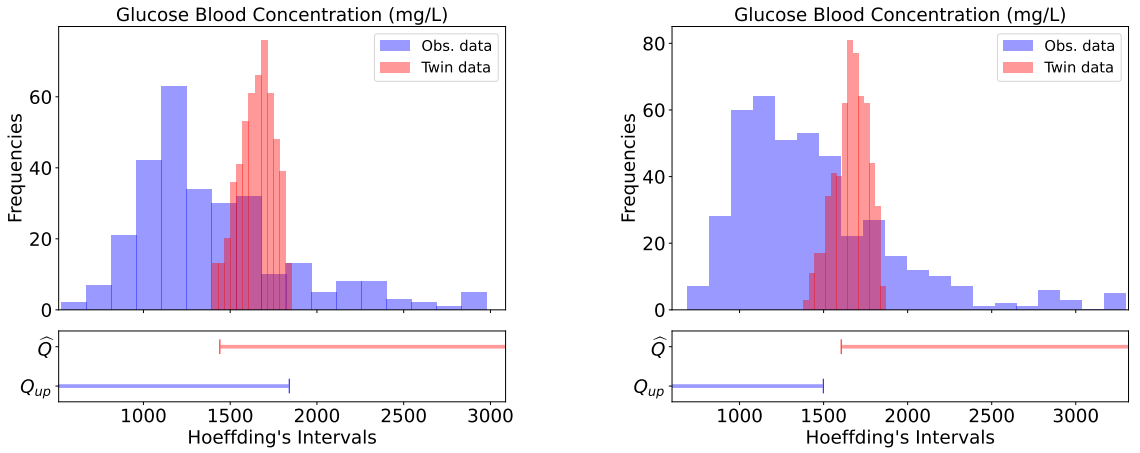
Figure 2: Estimates and 95% confidence intervals for $\widehat{Q}_t$ and Q_t^{obs} at each $1 \leq t \leq 4$ for two choices of $(B_{0:4}, a_{1:4})$, where $\widehat{Y}(a_{1:t})$ and $Y(a_{1:t})$ correspond to HCO_3 blood concentration. The dashed lines indicate lower and upper 95% confidence intervals for $Q_{\text{lo}}, Q_{\text{up}}$ respectively.

Q_t^{obs} for $t \in \{1, \ldots, 4\}$, where

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

Here $\widehat{Q}_t$ is just $\widehat{Q}$ 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 $t \in \{1, \ldots, 4\}$ obtained from Hoeffding's inequality. Directly comparing the estimates of $\widehat{Q}_t$ 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 $\widehat{Q}_1$ 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 of this kind.

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 $\widehat{X}_{0:t}(a_{1:t}) \in B_{0:t}$. (Note that these raw values differ slightly from $Y(A_{1:t})$ and $\widehat{Y}(a_{1:t})$ since they are not clipped to lie between y_{lo} and y_{up}.) Below each histogram we also show 95% confidence intervals for Q_{up} and $\widehat{Q}$ 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 9 in the Supplement. This demonstrates that our methodology does not simply rely on direct comparison of observed and simulated trajectories, but also accounts for the possibility of confounding in the data.



(a) Not rejected

(b) Rejected

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 $\widehat{Q}$ 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 9 of the Supplement.

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 the interventional distribution of the real-world process does not differ between when our dataset was obtained and when the twin is deployed to production. 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 $B_{0:t}$ 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 appear possible. For example, it is possible to replace our one-sided confidence intervals for Q_{lo}, Q_{up}, and $\widehat{Q}$ with two-sided ones, and thereby to obtain a procedure that may yield more precise information about the twin than we obtain by rejecting one of $\mathcal{H}_{\text{lo}}$ or $\mathcal{H}_{\text{up}}$. We outline this at a high level in Section H.3 of the Supplement. It also seems possible to leverage ideas from the literature on partial identification [35] to obtain greater statistical efficiency, for example by building on the line of work initiated by [21] for obtaining more informative confidence intervals. Beyond this, it may be useful in some contexts 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 [45, 52, 54]. This can yield tighter bounds on Q than are implied by Theorem 4, albeit at the expense of less robustness if these assumptions are violated.

Funding

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.

References

[1] 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. 10th IFAC Symposium on Advances in Automotive Control AAC 2022.

[2] AMSE. *Assessing Credibility of Computational Modeling through Verification and Validation: Application to Medical Devices*. AMSE, 2018.

[3] 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.

[4] Peter Bauer, Bjorn Stevens, and Wilco Hazeleger. A digital twin of earth for the green transition. *Nature Climate Change*, 11(2):80–83, 2021.

[5] Nicholas Bellinger, 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.

[6] 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.

[7] 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.

[8] 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, et al. The health digital twin to tackle cardiovascular disease—a review of an emerging interdisciplinary field. *NPJ Digital Medicine*, 2022.

[9] Jorge Corral-Acero, Francesca Margara, Maciej Marciniak, Cristobal Rodero, Filip Loncaric, Yingjing Feng, Andrew Gilbert, Joao F Fernandes, Hassaan A Bukhari, Ali Wajdan, et al. The 'digital twin' to enable the vision of precision cardiology. *European Heart Journal*, 41(48):4556–4564, 2020.

[10] David R Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.

[11] Ulrich Richard Dahmen, Tobias Osterloh, and Heinz-Jürgen Roßmann. Verification and validation of digital twins and virtual testbeds. *International journal of advances in engineering sciences and applied mathematics*, 11(1):47–64, 2022.

[12] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979.

[13] 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.

[14] 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.

[15] Peter Hall. Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, 16(3):927 – 953, 1988.

[16] 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.

[17] Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman and Hall/CRC, Boca Raton, 2020.

[18] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[19] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[20] Guido W Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79, 2020.

[21] Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

[22] Melanie Jans-Singh, Kathryn Leeming, Ruchi Choudhary, and Mark Girolami. Digital twin of an urban-integrated hydroponic farm. *Data-Centric Engineering*, 1:e20, 2020.

[23] 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.

[24] 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.

[25] Michael G Kapteyn, Jacob VR 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.

[26] 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)*. IEEE, 2018.

[27] Brendan Kochunas and Xun Huan. Digital twin concepts with uncertainty for nuclear power applications. *Energies*, 14(14):4235, 2021.

[28] Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.

[29] Amos Lal, Guangxi Li, Edin Cubro, Sarah Chalmers, Heyi Li, Vitaly Herasevich, Yue Dong, Brian Pickering, Kilickaya Oguz, and Ognjen Gajic. Development and verification of a digital twin patient model to predict treatment response in sepsis. *Critical Care Medicine*, 49:611–611, 01 2021.

[30] Ignacio Larrabide, Minsuok Kim, Luca Augsburger, 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, April 2012.

[31] Philip W Lavori and Ree Dawson. Dynamic treatment regimes: practical design considerations. *Clinical trials*, 1(1):9–20, 2004.

[32] Yuqian Lu, Chao Liu, I Kevin, 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.

[33] Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

[34] Charles F Manski. *Identification Problems in the Social Sciences*. Harvard University Press, 1995.

[35] Charles F Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

[36] Joseph Masison, Jonathan Beezley, Yu Mei, Henrique Assis Lopes Ribeiro, Adam C Knapp, L Sordo Vieira, Bandita Adhikari, Yogesh Scindia, Michael Grauer, Brian Helba, et al. A modular computational framework for medical digital twins. *Proceedings of the National Academy of Sciences*, 118(20):e2024287118, 2021.

[37] 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.

[38] 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.

[39] S. A. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1481, 2005.

[40] Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

[41] 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.

[42] Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

[43] Karl Popper. *The Logic of Scientific Discovery*. Routledge, 2005.

[44] 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.

[45] Paul R. Rosenbaum. *Observational Studies*. Springer, New York, NY, 2002.

[46] 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.

[47] 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.

[48] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[49] Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[50] Rafael Sacks, Ioannis Brilakis, Ergo Pikas, Haiyan Xie, and Mark Girolami. Construction with digital twin information systems. *Data-Centric Engineering*, 1, 2020.

[51] Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali 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, 02 2016.

[52] Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

[53] 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.

[54] Steve Yadlowsky, 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.

[55] Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Supplementary Material

Causal Falsification of Digital Twins

Rob Cornish*[1], Muhammad Faaiz Taufiq*[1], Arnaud Doucet[1], and Chris Holmes[1,2]

[1]Department of Statistics, University of Oxford, UK
[2]The Alan Turing Institute, UK

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 $\mathcal{Z}_t \times \cdots \times \mathcal{Z}_{t'}$ (or the empty set 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 $t > t'$)
$\mathrm{Law}[Z]$	The distribution of the random variable Z
$\mathrm{Law}[Z \mid M]$	The conditional distribution of Z given M, where M is either an event or a random variable
$Z \overset{\text{a.s.}}{=} Z'$	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

B Proof of Proposition 2 (unconditional form of interventional correctness)

In this section we prove Proposition 2 from the main text. To account for measure-theoretic technicalities, we first clarify slightly our earlier definition of interventional correctness. Since each $\mathcal{X}_t = \mathbb{R}^{d_t}$ is real-valued, $X_{0:T}(a_{1:T})$ admits a regular conditional distribution given X_0 [58, Theorem 6.3]. By interventional correctness, we then mean that, for all $a_{1:T} \in \mathcal{A}_{1:T}$, the map $(x_0, B_{1:T}) \mapsto \mathrm{Law}[\widehat{X}_{1:T}(x_0, a_{1:T})](B_{1:T})$ is a version of this conditional distribution, i.e. it is a Markov kernel such that

$$\mathrm{Law}[\widehat{X}_{1:T}(x_0, a_{1:T})] = \mathrm{Law}[X_{1:T}(a_{1:T}) \mid X_0 = x_0] \qquad \text{for } \mathrm{Law}[X_0]\text{-almost all } x_0 \in \mathcal{X}_0. \tag{13}$$

*Both authors contributed equally to this work.

The result is then as follows.

Proposition 2. *The twin is interventionally correct if and only if, for all $a_{1:T} \in \mathcal{A}_{1:T}$, it holds that*

$$\mathrm{Law}[X_0, \widehat{X}_{1:T}(X_0, a_{1:T})] = \mathrm{Law}[X_{0:T}(a_{1:T})] \tag{14}$$

Proof. Fix any choice of $a_{1:T} \in \mathcal{A}_{1:T}$. Since both $\mathrm{Law}[X_0, \widehat{X}_{1:T}(X_0, a_{1:T})]$ and $\mathrm{Law}[X_{0:T}(a_{1:T})]$ have the same $\mathcal{X}_0$-marginal, namely $\mathrm{Law}[X_0]$, (14) holds if and only if

$$\mathrm{Law}[\widehat{X}_{1:T}(X_0, a_{1:T}) \mid X_0 = x_0] = \mathrm{Law}[X_{1:T}(a_{1:T}) \mid X_0 = x_0] \qquad \text{for } \mathrm{Law}[X_0]\text{-almost all } x_0 \in \mathcal{X}_0. \tag{15}$$

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

$$\widehat{X}_{1:T}(X_0, a_{1:T}) = \boldsymbol{h}(X_0, a_{1:T}, U_{1:T}),$$

where $\boldsymbol{h}(x_0, a_{1:T}, u_{1:T}) \coloneqq (h_1(x_0, a_1, u_1), \ldots, h_T(x_0, a_{1:T}, u_{1:T}))$. For all $x_0 \in \mathcal{X}_0$ and measurable $B_{1:T} \subseteq \mathcal{X}_{1:T}$, we then have

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

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

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

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

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

Substituting this into (15), we see that (14) holds if and only if (13) does. The result now follows since $a_{1:T}$ was arbitrary. □

C Online prediction

A distinguishing feature of many digital twins is their ability to integrate real-time information obtained from sensors in their environment [3]. 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 formalise 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 $(\widehat{X}^1_{1:T}(x_0, a_{1:T}) : a_{1:T} \in \mathcal{A}_{1:T})$, 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 $(\widehat{X}^2_{2:T}(x_{0:1}, a_{1:T}) : a_{2:T} \in \mathcal{A}_{2:T})$. More generally, for $t \in \{1, \ldots, T\}$, 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.

C.1 Correctness in the online setting

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

$$\text{Law}[\widehat{X}^t_{t:T}(x_{0:t-1}, a_{1:T})] = \text{Law}[X_{t:T}(a_{1:T}) \mid X_{0:t-1}(a_{1:t-1}) = x_{0:t-1}] \tag{16}$$

for all $t \in \{1, \ldots, T\}$, $a_{1:T} \in \mathcal{A}_{1:T}$, and $\text{Law}[X_{0:t-1}(a_{1:t-1})]$-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 (16) holds if and only if we have

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

for all $t \in \{1, \ldots, T\}$, $a_{1:T} \in \mathcal{A}_{1:T}$, $\text{Law}[X_0]$-almost all $x_0 \in \mathcal{X}_0$, and $\text{Law}[\widehat{X}^1_{1:t-1}(x_0, a_{1:t-1})]$-almost all $x_{1:t-1} \in \mathcal{X}_{1:t-1}$. The first condition here says that $\widehat{X}^1_{1:T}(x_0, a_{1:T})$ 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 $\widehat{X}^1_{1:T}(x_0, a_{1:T})$ 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 (16) satisfies

$$\text{Law}[\widehat{X}^t_{t:T}(x_{0:t-1}, a_{1:T})] = \text{Law}[X_{t:T}(a_{1:T}) \mid X_{0:t-1}(a_{1:t-1}) = x_{0:t-1}, A_{1:t-1} = a_{1:t-1}] \tag{17}$$

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

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

for all $a_{1:T} \in \mathcal{A}_{1:T}$ and $\mathrm{Law}[X_{0:t-1}(a_{1:t-1}) \mid A_{1:t-1} = a_{1:t-1}]$-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 (16) 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 (as may occur for example when the twin is used as a decision-support tool). However, should it be desirable, our framework could be extended to encompass the alternative condition in (17) 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 $\widehat{X}^t_{t:T}(x_{0:t-1}, a_{1:T})$ 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.

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 3. *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*

$$(\tilde{X}_{0:T}(A_{1:T}), A_{1:T}) \overset{\text{a.s.}}{=} (X_{0:T}(A_{1:T}), A_{1:T}). \tag{18}$$

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

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

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

It is then easily checked that (18) holds, but

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

from which the result follows. □

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 $t \in \{1, \ldots, T\}$ such that

$$X_t(a_{1:t}) \overset{\text{a.s.}}{=} g_t(X_{0:t-1}(a_{1:t-1}), a_{1:t}) \qquad \text{for all } t \in \{1, \ldots, T\} \text{ and } a_{1:t} \in \mathcal{A}_{1:t}. \tag{19}$$

By "unconfounded", we mean that the *sequential randomisation assumption (SRA)* introduced by Robins [44] holds, i.e.

$$(X_s(a_{1:s}) : s \in \{1, \ldots, T\}, a_{1:s} \in \mathcal{A}_{1:s}) \perp\!\!\!\perp A_t \mid X_{0:t-1}(A_{1:t-1}), A_{1:t-1} \qquad \text{for all } t \in \{1, \ldots, T\}, \tag{20}$$

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 [53, Chapter 5] for a review).

It is now a standard exercise to show that (19) implies (20). We include a proof below for completeness. Key to this is the following straightforward Lemma.

Lemma 7. *Suppose U and V are random variables such that, for some measurable function g, it holds that $U \overset{\text{a.s.}}{=} g(V)$. 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 (19) implies (20). Indeed, by recursive substitution, it is straightforward to show that there exist measurable functions $\tilde{g}_t$ for $t \in \{1, \ldots, T\}$ such that

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

and so

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

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 7.

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 causal considerations (including unmeasured confounding in particular). Suppose a digital twin has been designed for a particular make of car, e.g. to facilitate autonomous driving [1]. The twin simulates how quantities 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.

In the causal inference literature, any unmeasured quantity (e.g. brake pad age) that affects both some choice of action taken in the data (e.g. aggressive braking) and the resulting observation (e.g. speed) is referred to as an *unmeasured confounder*. In general, whenever an unmeasured confounder 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

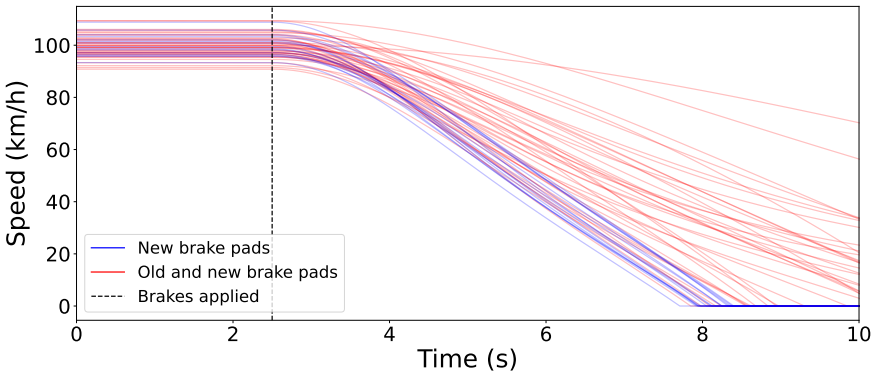


Figure 4: The discrepancy between observational data and interventional behavior. The data only show the effect of aggressive braking on cars with new brake pads (blue). This differs from what *would* be observed if aggressive braking were applied to the entire fleet of cars, encompassing both those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

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 [53]. 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 [34, 53, 17]. 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.

G Causal bounds

G.1 Proof of Theorem 4

Theorem 4. *Suppose* $(Y(a_{1:t}) : a_{1:t} \in \mathcal{A}_{1:t})$ *are real-valued potential outcomes, and 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*

$$\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t}) > 0 \tag{21}$$

$$\mathbb{P}(y_{\text{lo}} \leq Y(a_{1:t}) \leq y_{\text{up}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) = 1. \tag{22}$$

Then it holds that

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

where we define the following random variables:

$$\begin{aligned} N &\coloneqq \max\{0 \le s \le t \mid A_{1:s} = a_{1:s}\} \\ Y_{\text{lo}} &\coloneqq \mathbb{1}(A_{1:t} = a_{1:t})\, Y(A_{1:t}) + \mathbb{1}(A_{1:t} \neq a_{1:t})\, y_{\text{lo}} \\ Y_{\text{up}} &\coloneqq \mathbb{1}(A_{1:t} = a_{1:t})\, Y(A_{1:t}) + \mathbb{1}(A_{1:t} \neq a_{1:t})\, y_{\text{up}}. \end{aligned}$$

Proof. We prove the lower bound; the upper bound is analogous. It is easily checked that

$$\begin{aligned} \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}] = \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t}]\, \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) \\ + \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} \neq a_{1:t}]\, \mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}). \quad (23) \end{aligned}$$

If $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) > 0$, then

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

where the second step follows because $\mathbb{P}(N = t \mid A_{1:t} = a_{1:t}) = 1$. Similarly, if $\mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) > 0$, then (22) implies

$$\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} \neq a_{1:t}] \ge y_{\text{lo}}.$$

Substituting these results into (23), we obtain

$$\begin{aligned} \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}] \ge \mathbb{E}[Y(A_{1:t}) \mid X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t}]\, \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) \\ + y_{\text{lo}}\, \mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}). \quad (24) \end{aligned}$$

Now observe that the right-hand side of (24) is a convex combination with mixture weights $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t})$ and $\mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t})$. We can bound

$$\begin{aligned} \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) &= \frac{\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t})}{\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t})} \\ &\ge \frac{\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t})}{\mathbb{P}(X_{0:N}(a_{1:N}) \in B_{0:N})} \\ &= \frac{\mathbb{P}(X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t})}{\mathbb{P}(X_{0:N}(A_{1:N}) \in B_{0:N})} \\ &= \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}), \quad (25) \end{aligned}$$

where the inequality holds because $t \ge N$ almost surely, and the second equality holds because the definition of N means

$$X_{0:N}(a_{1:N}) \overset{\text{a.s.}}{=} X_{0:N}(A_{1:N}).$$

As such, we can bound the convex combination in (24) from below by replacing its mixture weights with $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$ and $\mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:N}(a_{1:N}) \in B_{0:N})$,

which shifts weight from the $\mathbb{E}[Y(A_{1:t}) \mid A_{1:t} = a_{1:t}, X_{0:N}(A_{1:N}) \in B_{0:N}]$ term onto the y_{lo} term. This yields

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

□

G.2 Proof of Proposition 5 (sharpness of bounds)

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

$$\mathbb{P}(y_{\text{lo}} \leq \tilde{Y}(a_{1:t}) \leq y_{\text{lo}} \mid \tilde{X}_{0:t} \in B_{0:t}) = 1$$

and moreover

$$(\tilde{X}_{0:T}(A_{1:T}), \tilde{Y}(A_{1:t}), A_{1:T}) \overset{\text{a.s.}}{=} (X_{0:T}(A_{1:T}), Y(A_{1:t}), A_{1:T}), \tag{26}$$

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}.

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

$$\begin{aligned}
\tilde{X}_0 &\coloneqq X_0 \\
\tilde{X}_s(a'_{1:s}) &\coloneqq \mathbb{1}(A_{1:s} = a'_{1:s})\, X_s(a'_{1:s}) + \mathbb{1}(A_{1:s} \neq a'_{1:s})\, x_s \qquad \text{for each } s \in \{0, \ldots, T\} \text{ and } a'_{1:s} \in \mathcal{A}_{1:s},
\end{aligned}$$

and similarly let

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

It is easy to check that (26) holds. 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

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

so that

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

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.2.1 Our case

Proposition 5 considered $(Y(a'_{1:t}) : a'_{1:t} \in \mathcal{A}_{1:t})$ to be arbitrary potential outcomes defined jointly on the same probability space as $(X_{0:t}(a'_{1:t}) : a'_{1:t} \in \mathcal{A}_{1:t})$. In contrast, our falsification methodology (Section 5 of the main text) assumes the particular form

$$Y(a'_{1:t}) \overset{\text{a.s.}}{=} f(X_{0:t}(a'_{1:t})),$$

for some known measurable function $f : \mathcal{X}_{0:t} \to \mathbb{R}$. Proposition 5 does not directly imply that the bounds in Theorem 4 are sharp under this additional assumption, but this nevertheless remains true for many cases of interest in practice. In particular, if f depends only on the final observation space $\mathcal{X}_t$ (which is true for example throughout our case study), i.e. we have

$$Y(a'_{1:t}) \overset{\text{a.s.}}{=} f(X_t(a'_{1:t})),$$

then it still holds that the bounds can be achieved, provided the worst-case values are chosen sensibly as

$$y_{\text{lo}} = \min_{x_t \in B_t} f(x_t) \qquad y_{\text{up}} = \max_{x_t \in B_t} f(x_t).$$

This follows straightforwardly by modifying the proof of Proposition 5 to define x_t as either the minimiser or maximiser of f on B_t.

G.3 Bounds on the conditional expectation given specific covariate values

Theorem 4 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 $\widehat{X}_{0:t}(a_{1:t})$, rather than just that it is incorrect on average across all runs that output values $\widehat{X}_{0:t}(a_{1:t}) \in B_{0:t}$.

G.3.1 Discrete covariates

When $X_{0:t}(a_{1:t})$ is discrete, Theorem 4 yields measurable functions $g_{\text{lo}}, g_{\text{up}} : \mathcal{X}_{0:t} \to \mathbb{R}$ such that

$$g_{\text{lo}}(X_{0:t}(a_{1:t})) \leq \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})] \leq g_{\text{up}}(X_{0:t}(a_{1:t})) \qquad \text{almost surely.} \tag{27}$$

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 $B_{0:t} \coloneqq \{x_{0:t}\}$, and similarly for $g_{\text{up}}(x_{0:t})$. Moreover, since the constants $y_{\text{lo}}, y_{\text{up}} \in \mathbb{R}$ in Theorem 4 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 $y_{\text{lo}}, y_{\text{up}} : \mathcal{X}_{0:t} \to \mathbb{R}$ satisfying

$$y_{\text{lo}}(X_{0:t}(a_{1:t})) \leq Y(a_{1:t}) \leq y_{\text{up}}(X_{0:t}(a_{1:t})) \qquad \text{almost surely.} \tag{28}$$

In other words, when $X_{0:t}(a_{1:t})$ is discrete, Theorem 4 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 (28) holds.

G.3.2 Continuous initial covariates

When $\mathbb{P}(X_{1:t}(a_{1:t}) \in B_{1:t}) > 0$, a fairly straightforward modification of the proof of Theorem 4 yields bounds of the following form:

$$\begin{aligned}\mathbb{E}[Y_{\text{lo}} \mid X_0, X_{1:N}(A_{1:N}) \in B_{1:N}] &\leq \mathbb{E}[Y(a_{1:t}) \mid X_0, X_{1:t}(a_{1:t}) \in B_{1:t}] \\ &\leq \mathbb{E}[Y_{\text{up}} \mid X_0, X_{1:N}(A_{1:N}) \in B_{1:N}] \qquad \text{almost surely.}\end{aligned} \tag{29}$$

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 (27), provided (28) holds. Alternatively, by taking $B_{1:t} \coloneqq \mathcal{X}_{1:t}$, (29) 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 [33], which allowed conditioning on potentially continuous pre-treatment covariates corresponding to our X_0.

G.3.3 Proof of Theorem 6 (no bounds for continuous subsequent covariates)

We now give a proof of Theorem 6 from the main text, which shows that, unlike in the examples just given, bounds on $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})]$ analogous to Theorem 4 cannot be obtained without further assumptions. We refer the reader to the main text for a full explanation, including a definition of "permissible".

Theorem 6. *Suppose X_0 is almost surely constant, $\mathbb{P}(A_1 \neq a_1) > 0$, and for some $s \in \{1, \ldots, t\}$ we have*

$$\mathbb{P}(X_s(A_{1:s}) = x_s) = 0 \text{ for all } x_s \in \mathcal{X}_s. \tag{30}$$

Then $g_{\text{lo}}, g_{\text{up}} : \mathcal{X}_{0:t} \to \mathbb{R}$ are permissible bounds only if they are trivial, i.e. we have

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

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

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

Similarly, define

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

It immediately follows that

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

Moreover, it is easily checked that

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

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

$$\begin{aligned} g_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t})) &\leq \mathbb{E}[\tilde{Y}(a_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t})] \\ &= \mathbb{E}[\tilde{Y}(A_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t}), A_{1:t} = a_{1:t}]\, \mathbb{P}(A_{1:t} = a_{1:t} \mid \tilde{X}_{0:t}(a_{1:t})) \\ &\quad + \underbrace{\mathbb{E}[\tilde{Y}(a_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t}), A_{1:t} \neq a_{1:t}]}_{= y_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t}))}\, \mathbb{P}(A_{1:t} \neq a_{1:t} \mid \tilde{X}_{0:t}(a_{1:t})). \end{aligned} \tag{31}$$

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 in (30). Combining this with (31), we get, almost surely,

$$\begin{aligned} \mathbb{1}(A_1 \neq a_1)\, g_{\text{lo}}(X_0, x_{1:t}) &= \mathbb{1}(A_1 \neq a_1)\, g_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t})) \\ &\leq \mathbb{1}(A_1 \neq a_1)\, y_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t})) \\ &= \mathbb{1}(A_1 \neq a_1)\, y_{\text{lo}}(X_0, x_{1:t}). \end{aligned} \tag{32}$$

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

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

The result now follows. □

Illustrative example To gain intuition for the phenomenon underlying Theorem 6, consider a simplified model consisting of $\mathcal{X}$-valued potential outcomes $(X(a') : a \in \mathcal{A})$, $\mathbb{R}$-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:

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

We then have

$$\mathbb{E}[Y(a) \mid X(a)] \overset{\text{a.s.}}{=} \mathbb{E}[Y(A) \mid X(A), A = a]\,\mathbb{P}(A = a \mid X(a)) + \mathbb{E}[Y(a) \mid X(a), A \neq a]\,\mathbb{P}(A \neq a \mid X(a)). \tag{33}$$

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

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

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 (33), 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.

Our case The discussion in this subsection considered $(Y(a'_{1:t}) : a'_{1:t} \in \mathcal{A}_{1:t})$ to be arbitrary potential outcomes defined jointly on the same probability space as $(X_{0:t}(a'_{1:t}) : a'_{1:t} \in \mathcal{A}_{1:t})$. In contrast, our falsification methodology (Section 5 of the main text) assumes the particular form

$$Y(a_{1:t}) \overset{\text{a.s.}}{=} f(X_{0:t}(a_{1:t})),$$

which means $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})] \overset{\text{a.s.}}{=} f(X_{0:t}(a_{1:t}))$ is known trivially. In this context, an alternative quantity to consider is $\mathbb{E}[Y(a_{1:t}) \mid X_{0:r}(a_{1:r})]$ with $r \in \{0, \ldots, t-1\}$, which in general will be unknown and therefore still interesting to bound. In the discrete case, Theorem 4 yields a bound on this quantity obtained by taking $B_{0:r} = \{x_{0:r}\}$ with $\mathbb{P}(X_{0:r}(a_{1:r}) = x_{0:r}) > 0$ and then $B_{r+1:t} \coloneqq \mathcal{X}_{r+1:t}$, and with y_{lo} and y_{up} in (28) now replaced by

$$\min_{x_t \in \mathcal{X}_t} f(x_t) \qquad \text{and} \qquad \max_{x_t \in \mathcal{X}_t} f(x_t)$$

respectively. However, in the continuous case, the same issues described above continue to apply in many cases of interest. For example, when f is a function of $\mathcal{X}_t$ only (which holds for example throughout our case study), then under the assumptions of the previous result, the most informative almost sure lower bound on $\mathbb{E}[Y(a_{1:t}) \mid X_{0:r}(a_{1:r})]$ is

$$\min_{x_t \in \mathcal{X}_t} f(x_t),$$

which is already known trivially. Roughly, this follows by modifying the proof of Theorem 6 so that $\tilde{X}_t(a'_{1:t})$ becomes

$$\tilde{X}_t(a'_{1:t}) \coloneqq \mathbb{1}(A_{1:t} = a'_{1:t})\, X_t(a'_{1:t}) + \mathbb{1}(A_{1:t} \neq a'_{1:t}) \operatorname*{arg\,min}_{x_t \in \mathcal{X}_t} f(x_t),$$

and $\tilde{Y}(a'_{1:t})$ becomes

$$\tilde{Y}(a'_{1:t}) \coloneqq f(\tilde{X}_t(a_{1:t})).$$

The remainder of the argument is then unchanged. An analogous result holds for the upper bound also.

H Hypothesis testing methodology

H.1 Validity of testing procedure

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

Proposition 8. *Suppose that for some $\alpha \in (0,1)$ we have random variables $\widehat{R}^\alpha$ and R^α_{lo} satisfying*

$$\mathbb{P}(Q_{\text{lo}} \geq R^\alpha_{\text{lo}}) \geq 1 - \frac{\alpha}{2} \tag{34}$$

$$\mathbb{P}(\widehat{Q} \leq \widehat{R}^\alpha) \geq 1 - \frac{\alpha}{2}. \tag{35}$$

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

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

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

To see this, note that

$$(\{\widehat{Q} > \widehat{R}^\alpha\} \cup \{Q_{\text{lo}} < R^\alpha_{\text{lo}}\})^c = \{\widehat{Q} > \widehat{R}^\alpha\}^c \cap \{Q_{\text{lo}} < R^\alpha_{\text{lo}}\}^c = \{\widehat{Q} \leq \widehat{R}^\alpha\} \cap \{Q_{\text{lo}} \geq R^\alpha_{\text{lo}}\} \subseteq \{R^\alpha_{\text{lo}} \leq \widehat{R}^\alpha\}.$$

As such,

$$\mathbb{P}(\widehat{R}^\alpha < R^\alpha_{\text{lo}}) \leq \mathbb{P}(\{\widehat{Q} > \widehat{R}^\alpha\} \cup \{Q_{\text{lo}} < R^\alpha_{\text{lo}}\}) \leq \mathbb{P}(\widehat{Q} > \widehat{R}^\alpha) + \mathbb{P}(Q_{\text{lo}} < R^\alpha_{\text{lo}}) \leq \alpha/2 + \alpha/2 = \alpha.$$

□

H.2 Methodology for obtaining confidence intervals

In this section, we describe concretely how we use our data to obtain one-sided confidence intervals R^{α}_{lo} and $\widehat{R}^{\alpha}$ satisfying (34) and (35) as required by our procedure for testing $\widehat{Q} \geq Q_{\mathrm{lo}}$. We use an analogous procedure to obtain confidence intervals for testing $\widehat{Q} \leq Q_{\mathrm{up}}$. We consider two techniques: 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 $\widehat{Q}$. We therefore describe this first, and then describe the particulars of each technique separately.

Unbiased sample mean estimate of Q_{lo}

First, we describe how to obtain an unbiased sample mean estimate 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

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

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 $n \coloneqq |\mathcal{D}(a_{1:t}, B_{0:t})|$. We then denote by $Y^{(i)}_{\mathrm{lo}}$ for $i \in \{1, \ldots, n\}$ the corresponding values of

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

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

$$\mu_{\mathrm{lo}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} Y^{(i)}_{\mathrm{lo}} \tag{36}$$

is an unbiased estimator of Q_{lo}.

Unbiased sample mean estimate of $\widehat{Q}$

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

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

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

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

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

$$\widehat{\mu} \coloneqq \frac{1}{\widehat{n}} \sum_{i=1}^{\widehat{n}} \widehat{Y}^{(i)}$$

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

Exact confidence intervals via Hoeffding's inequality

Recall that we assume in our methodology that $Y(a_{1:t})$ has the form

$$Y(a_{1:t}) = f(X_{0:t}(a_{1:t})),$$

and that moreover

$$y_{\mathrm{lo}} \leq f(x_{0:t}) \leq y_{\mathrm{up}} \qquad \text{for all } x_{0:t} \in B_{0:t}. \tag{37}$$

This means that $\widehat{Y}^{(i)}$ is almost surely bounded in $[y_{\mathrm{lo}}, y_{\mathrm{up}}]$, 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 9. *If* (37) *holds, then for each* $\alpha \in (0,1)$*, letting*

$$\Delta \coloneqq (y_{\mathrm{up}} - y_{\mathrm{lo}}) \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \qquad \textit{and} \qquad \widehat{\Delta} \coloneqq (y_{\mathrm{up}} - y_{\mathrm{lo}}) \sqrt{\frac{1}{2\widehat{n}} \log \frac{2}{\alpha}},$$

and similarly

$$R^{\alpha}_{\mathrm{lo}} \coloneqq \mu_{\mathrm{lo}} - \Delta \qquad \textit{and} \qquad \widehat{R}^{\alpha} \coloneqq \widehat{\mu} + \widehat{\Delta},$$

it follows that

$$\mathbb{P}(Q_{\mathrm{lo}} \geq R^{\alpha}_{\mathrm{lo}}) \geq 1 - \frac{\alpha}{2} \qquad \textit{and} \qquad \mathbb{P}(\widehat{Q} \leq \widehat{R}^{\alpha}) \geq 1 - \frac{\alpha}{2}.$$

Proof. We only prove the result for R^{α}_{lo}; the other statement can be proved analogously. Recall that μ_{lo} is the empirical mean of i.i.d. samples $Y^{(i)}_{\mathrm{lo}}$ for $i \in \{1, \ldots, n\}$ with $\mathbb{E}[Y^{(i)}_{\mathrm{lo}}] = Q_{\mathrm{lo}}$ (see (36)). Moreover, by (37), $Y^{(i)}_{\mathrm{lo}}$ is almost surely bounded in $[y_{\mathrm{lo}}, y_{\mathrm{up}}]$. Hoeffding's inequality then implies that

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

In turn, some basic manipulations yield

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

□

H.2.1 Approximate confidence intervals via bootstrapping

While Hoeffding's inequality yields the probability guarantees in (34) and (35) 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 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 [12] on Q_{lo} and $\widehat{Q}$. Many bootstrapping procedures for obtaining confidence intervals have been proposed in the literature [61, 56, 57]. Our results reported below were obtained via the *reverse percentile* bootstrap (see [57] for an overview). (We also tried the *percentile* bootstrap method, which obtained nearly indistinguishable results.) In particular, this method takes

$$R_{\text{lo}}^{\alpha} \coloneqq 2\mu_{\text{lo}} - \Delta \qquad \widehat{R}^{\alpha} \coloneqq 2\widehat{\mu} - \widehat{\Delta},$$

where Δ and $\widehat{\Delta}$ correspond to the approximate $1-\alpha/2$ and $\alpha/2$ quantiles of the distributions of

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

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 and $\widehat{\Delta}$ 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 $\widehat{n}$ grow to infinity under mild assumptions [15].

H.3 Testing with two-sided confidence intervals

Although we do not consider it in our case study, it is possible to replace the one-sided confidence intervals for Q_{lo} and $\widehat{Q}$ that we use with two-sided intervals. This would allow us to define a procedure that does the following:

1. If the interval for Q_{lo} lies completely below the interval for $\widehat{Q}$ (without overlap), then infer that $\widehat{Q} \geq Q_{\text{lo}}$;
2. If the interval for Q_{lo} lies completely above the interval for $\widehat{Q}$ (without overlap), then infer that $\widehat{Q} < Q_{\text{lo}}$;
3. Otherwise, draw no inference.

In particular, notice that this procedure is now able to infer that $\widehat{Q} \geq Q_{\text{lo}}$ is true (if the first case occurs), as well as to infer that $\widehat{Q} \geq Q_{\text{lo}}$ is false as previously. By a closed testing

argument [59], this procedure can be shown to have at most probability α of drawing a false inference about the twin. A similar approach can be used for the hypothesis $\widehat{Q} \leq Q_{\text{up}}$ by obtaining two-sided intervals for Q_{up} and $\widehat{Q}$.

Since each of Q_{lo}, Q_{up}, and $\widehat{Q}$ are identifiable from the observational distribution, it is straightforward to obtain two-sided confidence intervals whose widths will shrink to zero as the size of the observational dataset grows large. (For instance, this holds for confidence intervals obtained from Hoeffding's inequality and from bootstrapping.) Consequently, with a sufficiently large dataset, with high probability, only one of the first two cases above will be observed to occur. As such, when it does occur, the third case would indicate that insufficient data has been collected to draw an appropriate conclusion about the twin, which could be useful information for practitioners. On the other hand, this procedure comes at an expense as it leads to a more conservative upper bound for Q_{lo} and lower bound for $\widehat{Q}$ when testing the hypothesis $\widehat{Q} \geq Q_{\text{lo}}$ (and analogously for $\widehat{Q} \leq Q_{\text{up}}$), and so may result in fewer falsifications than the method we consider based on one-sided confidence intervals instead.

I Experimental Details

In this section, we provide additional experimental details relating to our case study.

I.1 MIMIC preprocessing

For data extraction and preprocessing, we re-used the same procedure as [28] with minor modifications. For completeness, we describe the pre-processing steps applied in [28] and subsequently outline our modifications to these.

Patient cohorts

Following [28], we extracted adult patients fulfilling the sepsis-3 criteria [51]. 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 [51] [60].

Exclusion criteria

Following [28], we excluded patients for whom any of the following was true:

- Age < 18 years old at the time of ICU admission
- Mortality not documented
- IV fluid/vasopressors intake not documented

- Withdrawal of treatment

Our modifications

We made the following modifications to the pre-processing code of [28] for our experiment:

- Instead of extracting physiological quantities (e.g. heart rate) every 4 hours, we extracted these every hour.
- 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). These were chosen as the subset of physiological quantities extracted from MIMIC by [28] 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.

Sample splitting Before proceeding further, we randomly selected 5% of the extracted our 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 [10], the statistical guarantees of our testing procedure established above continue to apply even when our hypotheses are defined in this data-dependent way.

I.2 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 the version of Pulse we considered (4.x). (In contrast, Pulse initialises the other 10 features to default values.) For the remaining observation spaces, we used the full collection of 17 physiological quantities we extracted (i.e. those listed in Categories 1-3 of Table 2) to define $\mathcal{X}_1 = \cdots = \mathcal{X}_4$. 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.3 Action spaces

Following [28], we constructed our action space using 2 features obtained from MIMIC, namely intravenous fluid (IV) and vasopressor doses. To obtain discrete action spaces

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

Table 2: Physiological quantities extracted from MIMIC

		Vasopressor dose (μg/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

suitable for our framework, we used the same discretization procedure for these quantities as was used by [28]. 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 $\mathcal{A}_1 = \cdots = \mathcal{A}_4$ 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.4 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:

- For each $t \in \{0, \ldots, 4\}$, 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 $\mathcal{D}_0$.

- For each $t \in \{1, \ldots, 4\}$, $a_{1:t} \in \mathcal{A}_{1:t}$, 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.

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

  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.

- For each selected triple $(t, a_{1:t}, B_{0:t})$, we chose a corresponding f as follows:

  - Let $i \in \{1, \ldots, d_t\}$ denote the index of the physiological quantity of interest in $\mathcal{X}_t = \mathbb{R}^{d_t}$. We 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 defined $f : \mathcal{X}_{0:t} \to \mathbb{R}$ as the function that extracts the physiological quantity of interest from $\mathcal{X}_t$ and clips its value to between y_{lo} and y_{up}, i.e.

    $$f(x_{0:t}) \coloneqq \text{clip}((x_t)_i, y_{\text{lo}}, y_{\text{up}}). \tag{38}$$

    where $\text{clip}(z, a, b) \coloneqq \min(\max(z, a), b)$.

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 $\widehat{n}$ as defined in Section H.2 above. (We describe how we generated our dataset of twin trajectories in Section I.5.)

I.4.1 Implications of a falsification for clipped and unclipped outcomes

For a given index $i \in \{1, \ldots, d_t\}$ of some physiological quantity in $\mathcal{X}_t = \mathbb{R}^{d_t}$, denote the corresponding *unclipped* potential outcomes by

$$Z(a_{1:t}) \coloneqq (X_t(a_{1:t}))_i$$
$$\widehat{Z}(a_{1:t}) \coloneqq (\widehat{X}_t(X_0, a_{1:t}))_i,$$

so that our choice of f in (38) gives

$$Q = \mathbb{E}[\mathrm{clip}(Z(a_{1:t}), y_{\mathrm{lo}}, y_{\mathrm{up}}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}]$$
$$\widehat{Q} = \mathbb{E}[\mathrm{clip}(\widehat{Z}(a_{1:t}), y_{\mathrm{lo}}, y_{\mathrm{up}}) \mid X_0 \in B_0, \widehat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}].$$

Falsifying $\mathcal{H}_{\mathrm{up}}$ immediately yields the following inference about the twin:

$$\mathrm{clip}(\widehat{Z}(a_{1:t}), y_{\mathrm{lo}}, y_{\mathrm{up}}) \text{ is on average too large, conditional on } \{X_0 \in B_0, \widehat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}\}. \tag{39}$$

When $\mathcal{H}_{\mathrm{lo}}$ is false, we may similarly infer that $\mathrm{clip}(\widehat{Z}(a_{1:t}), y_{\mathrm{lo}}, y_{\mathrm{up}})$ is on average too small, conditional on the same event. However, these statements are expressed in terms of the clipped outcomes $\mathrm{clip}(\widehat{Z}(a_{1:t}), y_{\mathrm{lo}}, y_{\mathrm{up}})$. In practice, it may be of interest to draw some inference about the twin that is expressed more directly in terms of the unclipped outcomes $Z(a_{1:t})$ and $\widehat{Z}(a_{1:t})$. This is indeed possible as we explain now.

For notational simplicity, we will consider the case where $B_{0:t} = \mathcal{X}_{0:t}$ is trivial, which means Q and $\widehat{Q}$ simplify as

$$Q = \mathbb{E}[\mathrm{clip}(Z(a_{1:t}), y_{\mathrm{lo}}, y_{\mathrm{up}})]$$
$$\widehat{Q} = \mathbb{E}[\mathrm{clip}(\widehat{Z}(a_{1:t}), y_{\mathrm{lo}}, y_{\mathrm{up}})].$$

However, our considerations here can easily be adapted to the more general case. We then have the following result for $\mathcal{H}_{\mathrm{up}}$ (an analogous result holds for $\mathcal{H}_{\mathrm{lo}}$):

Proposition 10. *If $\mathcal{H}_{\mathrm{up}}$ is false (i.e. $\widehat{Q} > Q_{\mathrm{up}}$), then one of the following must hold:*

$$\mathbb{P}(\widehat{Z}(a_{1:t}) \geq y_{\mathrm{up}}) > \mathbb{P}(Z(a_{1:t}) \geq y_{\mathrm{up}}) \tag{40}$$
$$\mathbb{P}(\widehat{Z}(a_{1:t}) > y_{\mathrm{lo}}) > \mathbb{P}(Z(a_{1:t}) > y_{\mathrm{lo}}) \tag{41}$$
$$\mathbb{E}[\widehat{Z}(a_{1:t}) \mid y_{\mathrm{lo}} < \widehat{Z}(a_{1:t}) < y_{\mathrm{up}}] > \mathbb{E}[Z(a_{1:t}) \mid y_{\mathrm{lo}} < Z(a_{1:t}) < y_{\mathrm{up}}]. \tag{42}$$

Proof. We will use the following straightforward fact: given real numbers $a \leq b \leq c$ and discrete probability vectors (p, q, r) and (p', q', r') (i.e. $p, q, r \in [0, 1]$ with $p+q+r = 1$, and similarly for (p', q', r')), if it holds that $r \leq r'$ and $p \geq p'$, then

$$\begin{aligned} p\,a + q\,b + r\,c &\leq p'\,a + (q + p - p')\,b + r\,c \\ &\leq p'\,a + (q + p - p' - (r' - r))\,b + r'\,c \\ &= p'\,a + q'\,b + r'\,c. \end{aligned} \tag{43}$$

Intuitively, we first move $p' - p$ units of mass from a to b, and then $r' - r$ units from b to c, and both steps can only increase the expected value since $a \leq b \leq c$.

Now assume that (40), (41), and (42) do *not* hold. We will show that in this case $\mathcal{H}_{\mathrm{up}}$ must be *true*, i.e. $\widehat{Q} \leq Q_{\mathrm{up}}$. Indeed, we have

$$\begin{aligned} \widehat{Q} &= \mathbb{P}(\widehat{Z}(a_{1:t}) \leq y_{\mathrm{lo}})\, y_{\mathrm{lo}} + \mathbb{P}(y_{\mathrm{lo}} < \widehat{Z}(a_{1:t}) < y_{\mathrm{up}})\, \mathbb{E}[\widehat{Z}(a_{1:t}) \mid y_{\mathrm{lo}} < \widehat{Z}(a_{1:t}) < y_{\mathrm{up}}] + \mathbb{P}(\widehat{Z}(a_{1:t}) \geq y_{\mathrm{up}})\, y_{\mathrm{up}} \\ &\leq \mathbb{P}(\widehat{Z}(a_{1:t}) \leq y_{\mathrm{lo}})\, y_{\mathrm{lo}} + \mathbb{P}(y_{\mathrm{lo}} < \widehat{Z}(a_{1:t}) < y_{\mathrm{up}})\, \mathbb{E}[Z(a_{1:t}) \mid y_{\mathrm{lo}} < Z(a_{1:t}) < y_{\mathrm{up}}] + \mathbb{P}(\widehat{Z}(a_{1:t}) \geq y_{\mathrm{up}})\, y_{\mathrm{up}} \\ &\leq \mathbb{P}(Z(a_{1:t}) \leq y_{\mathrm{lo}})\, y_{\mathrm{lo}} + \mathbb{P}(y_{\mathrm{lo}} < Z(a_{1:t}) < y_{\mathrm{up}})\, \mathbb{E}[Z(a_{1:t}) \mid y_{\mathrm{lo}} < Z(a_{1:t}) < y_{\mathrm{up}}] + \mathbb{P}(Z(a_{1:t}) \geq y_{\mathrm{up}})\, y_{\mathrm{up}} \\ &= Q, \end{aligned}$$

where the first inequality follows since we assumed (42) is false, and the second from (43), where we note that (41) is false precisely when $\mathbb{P}(\widehat{Z}(a_{1:t}) \leq y_{\text{lo}}) \leq \mathbb{P}(Z(a_{1:t}) \leq y_{\text{lo}})$. □

As a result, when $\mathcal{H}_{\text{up}}$ is falsified, Proposition 10 yields the following inference about the behaviour of these outcomes:

$$\text{One of (40), (41), or (42) is false.} \tag{44}$$

Unlike (39), this inference is expressed directly in terms of the unclipped outcomes $Z(a_{1:t})$ and $\widehat{Z}(a_{1:t})$, since each of (40), (41), and (42) are. Intuitively, we may still interpret (44) as saying that the output $\widehat{Z}(a_{1:t})$ produced by the twin is typically too large: either the twin places too much probability mass in one of the upper tails (y_{lo}, ∞) or $[y_{\text{up}}, \infty)$ (i.e. (40) or (41)), or else it is on average too large in the region $(y_{\text{lo}}, y_{\text{up}})$ (i.e. (42)). The situation is analogous when $\mathcal{H}_{\text{lo}}$ is false, and can be interpreted as saying that the output of the twin $\widehat{Z}(a_{1:t})$ is typically too small.

In some situations, it may be desirable to obtain a more granular conclusion about the outcomes of the twin by identifying *which* case of (40), (41), and (42) holds. This can be achieved by an appropriate choice of f or $B_{0:t}$. For instance, when $f(x_{0:t}) = \mathbb{1}((x_t)_i \geq y_{\text{up}})$, if $\mathcal{H}_{\text{up}}$ is false, then (40) must hold. Similarly, when $B_t = \{x_t \in \mathcal{X}_t \mid (x_t)_i \in (y_{\text{lo}}, y_{\text{up}})\}$, if $\mathcal{H}_{\text{up}}$ is false, then (42) must hold.

I.5 Generating twin trajectories using the Pulse Physiology Engine

The Pulse Physiology Engine is an open source comprehensive human physiology simulator that has been used in medical education, research, and training. The core engine of Pulse is C++ based with APIs available in different languages, including python. Detailed documentation is available at: pulse.kitware.com. Pulse allows users to initialise 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).

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 $\widehat{\mathcal{D}}(a_{1:t})$ are i.i.d. Additionally, Algorithm 1 can be easily parallelised to improve efficiency. Figure 5 shows histograms of the number of twin trajectories $\widehat{n}$ (as defined in Section H.2 above) obtained in this way across all hypotheses.

I.6 Bootstrapping details

In addition to Hoeffding's inequality, we also used reverse percentile bootstrap method (see e.g. [57]) to test the chosen hypotheses as described in Section H.1. We used 100 bootstrap

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

Inputs: Action sequence $a_{1:t}$; Observational dataset $\mathcal{D}$.
Output: Twin data $\widehat{\mathcal{D}}(a_{1:t})$ of size m.
for $i = 1, \ldots, m$ **do**
  Sample X_0 without replacement from $\mathcal{D}$;
  $\widehat{X}_0 \leftarrow X_0$ i.e., initialise the Pulse trajectory with the information of X_0;
  for $t' = 1, \ldots, t$ **do**
    Administer the median doses of IV fluids and vasopressors in action bin $a_{t'}$;
    if $t' \equiv 0 \pmod 3$ **then**
      Virtual patient in Pulse consumes nutrients and water, and urinates;
    end
    Advance the twin trajectory by one hour;
  end
  Add the trajectory $\widehat{X}_{0:t}(a_{1:t})$ to $\widehat{\mathcal{D}}(a_{1:t})$;
end
Return $\widehat{\mathcal{D}}(a_{1:t})$

Physiological quantity	# Rejections	# Hypotheses
Chloride Blood Concentration (Chloride)	24	94
Sodium Blood Concentration (Sodium)	21	94
Potassium Blood Concentration (Potassium)	13	94
Skin Temperature (Temp)	10	86
Calcium Blood Concentration (Calcium)	5	88
Glucose Blood Concentration (Glucose)	5	96
Arterial CO_2 Pressure ($paCO_2$)	3	70
Bicarbonate Blood Concentration (HCO_3)	2	90
Systolic Arterial Pressure (SysBP)	2	154
Arterial O_2 Pressure (paO_2)	0	78
Arterial pH (Arterial_pH)	0	80
Diastolic Arterial Pressure (DiaBP)	0	72
Mean Arterial Pressure (MeanBP)	0	92
Respiration Rate (RR)	0	172
Heart Rate (HR)	0	162

Table 4: Total hypotheses and rejections per physiological quantity obtained using Hoeffding's inequality.

Physiological quantity	# Rejections	# Hypotheses
Chloride Blood Concentration (Chloride)	47	94
Sodium Blood Concentration (Sodium)	46	94
Potassium Blood Concentration (Potassium)	33	94
Skin Temperature (Temp)	43	86
Calcium Blood Concentration (Calcium)	44	88
Glucose Blood Concentration (Glucose)	19	96
Arterial CO_2 Pressure ($paCO_2$)	13	70
Bicarbonate Blood Concentration (HCO_3)	8	90
Systolic Arterial Pressure (SysBP)	8	154
Arterial O_2 Pressure (paO_2)	4	78
Arterial pH (Arterial_pH)	0	80
Diastolic Arterial Pressure (DiaBP)	0	72
Mean Arterial Pressure (MeanBP)	3	92
Respiration Rate (RR)	12	172
Heart Rate (HR)	1	162

Table 5: The total number of hypotheses per outcome, along with rejections obtained using the reverse percentile bootstrap.

samples to construct the confidence intervals on Q_{lo}, $\widehat{Q}$ and Q_{up} for each hypothesis. 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 $\widehat{n}$ was less than 100, and returned a p-value of 1 in each such case.

Hypothesis rejections

Table 5 shows the number of rejected hypotheses for each physiological quantity. We observed similar trends between results obtained using Hoeffding's inequality (Table 4) and bootstrapping (Table 5). 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).

p-value plots

Figure 6 shows the distributions of $-\log_{10} p_{\text{lo}}$ and $-\log_{10} p_{\text{up}}$ for different physiological quantities, with higher values indicating greater evidence in favour of rejection. We again saw the same trends between p-value plots for bootstrapping (Figure 6) and Hoeffding's inequality (Figure 7). Specifically, we can see that the p-values for Sodium, Chloride, Potassium blood concentrations and Skin Temperature are often low, suggesting that the twin simulation of these quantities is not accurate. Additionally, we again observed that for each physiological quantity, we either obtain low values for p_{lo}, or low values of p_{up}, but not both. Moreover, for each quantity, whether we obtain low values for p_{lo} or low values for p_{up} remained consistent between Figures 6 and 7. For example, both plots suggest that Calcium and Sodium blood concentrations are over-estimated by the twin whereas Skin Temperature and Chloride blood concentrations are under-estimated.

Overall, it can be seen that we obtained lower p-values and consequently more rejections when using bootstrapping, as compared to Hoeffding's inequality. This happens because the finite sample guarantee in Hoeffding's inequality comes at the cost of more conservative intervals. We confirmed this empirically in Figure 8, which plots the histograms of the ratios

$$\frac{\text{Hoeffding's interval length}}{\text{Bootstrapping interval length}},$$

for each physiological quantity. We observed that bootstrapping yielded confidence intervals that were typically between 7.5 to 15 times smaller than those produced by Hoeffding's inequality.

I.7 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

$$\frac{Q_{\text{up}} - Q_{\text{lo}}}{y_{\text{up}} - y_{\text{lo}}} = 1 - \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}). \tag{45}$$

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 10 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 (45) 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.8 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 11 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 our both the bounds $[Q_{\text{lo}}, Q_{\text{up}}]$ and our confidence intervals R^{α}_{lo} and R^{α}_{up} obtained using Hoeffding's inequality (see Proposition 9) 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.

References

[56] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[57] Tim C. Hesterberg. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4):371–386, 2015. PMID: 27019512.

[58] Olav Kallenberg and Olav Kallenberg. *Foundations of Modern Probability*, volume 2. Springer, 1997.

[59] Ruth Marcus, Eric Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

[60] 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, 02 2016.

[61] Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436, 1993.

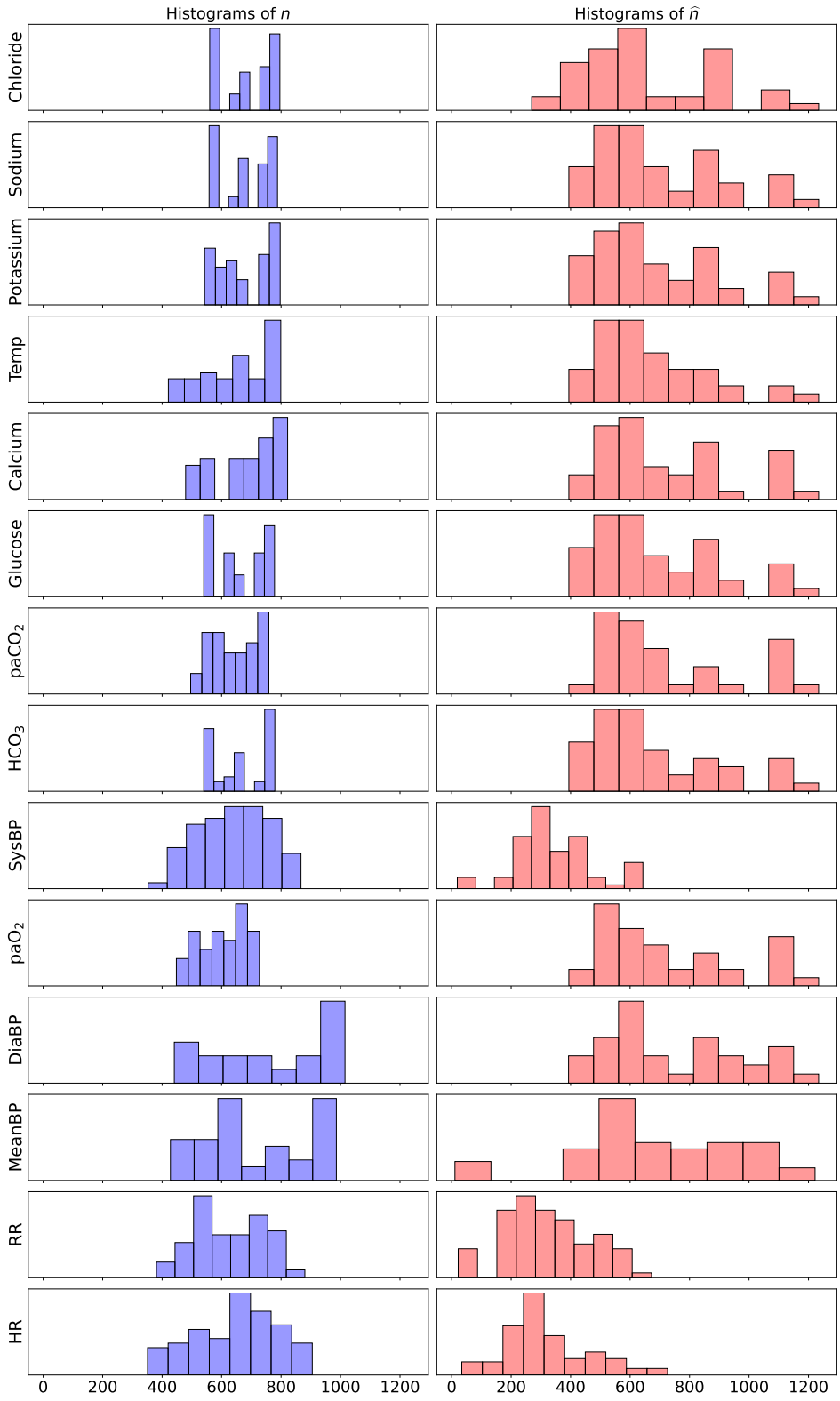


Figure 5: Histograms of n and $\widehat{n}$ (as defined in Section H.2) across all hypothesis parameters corresponding to each physiological quantity of interest.

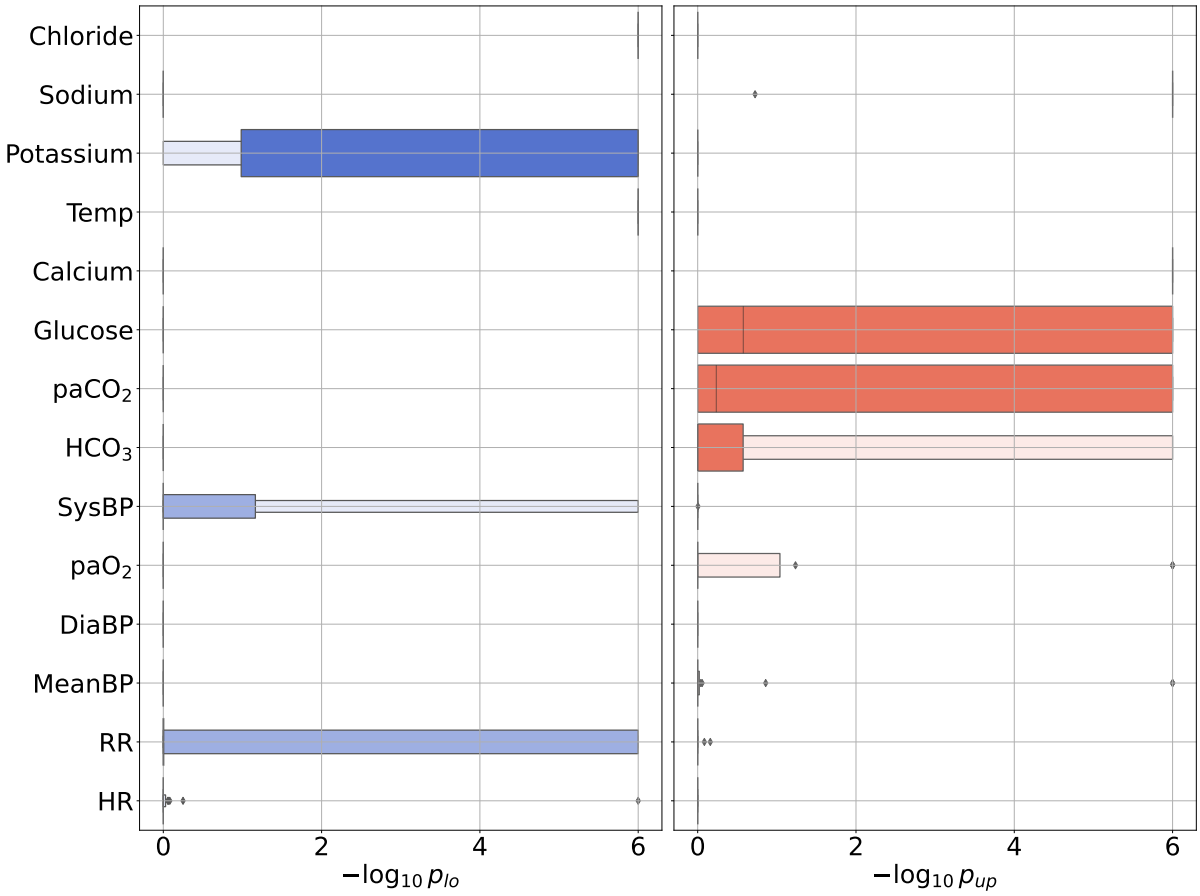


Figure 6: Boxenplots showing distributions of $-\log_{10} p_{\text{lo}}$ and $-\log_{10} p_{\text{up}}$ for different physiological quantities obtained via the reverse percentile method. Higher values of $-\log_{10} p$ indicate greater evidence in favour of rejection. Note that we computed these p-values numerically by determining the lowest level at which each hypothesis was rejected over a grid of values in $(0, 1)$, with the smallest such level being 10^{-6}. In cases where a hypothesis was rejected at every level tested, we defined the p-value to be 10^{-6}, and so the horizontal axis here is truncated to between 0 and 6. In some cases, e.g. $\mathcal{H}_{\text{lo}}$ for Chloride, every hypothesis obtained a p-value of 10^{-6} in this way.

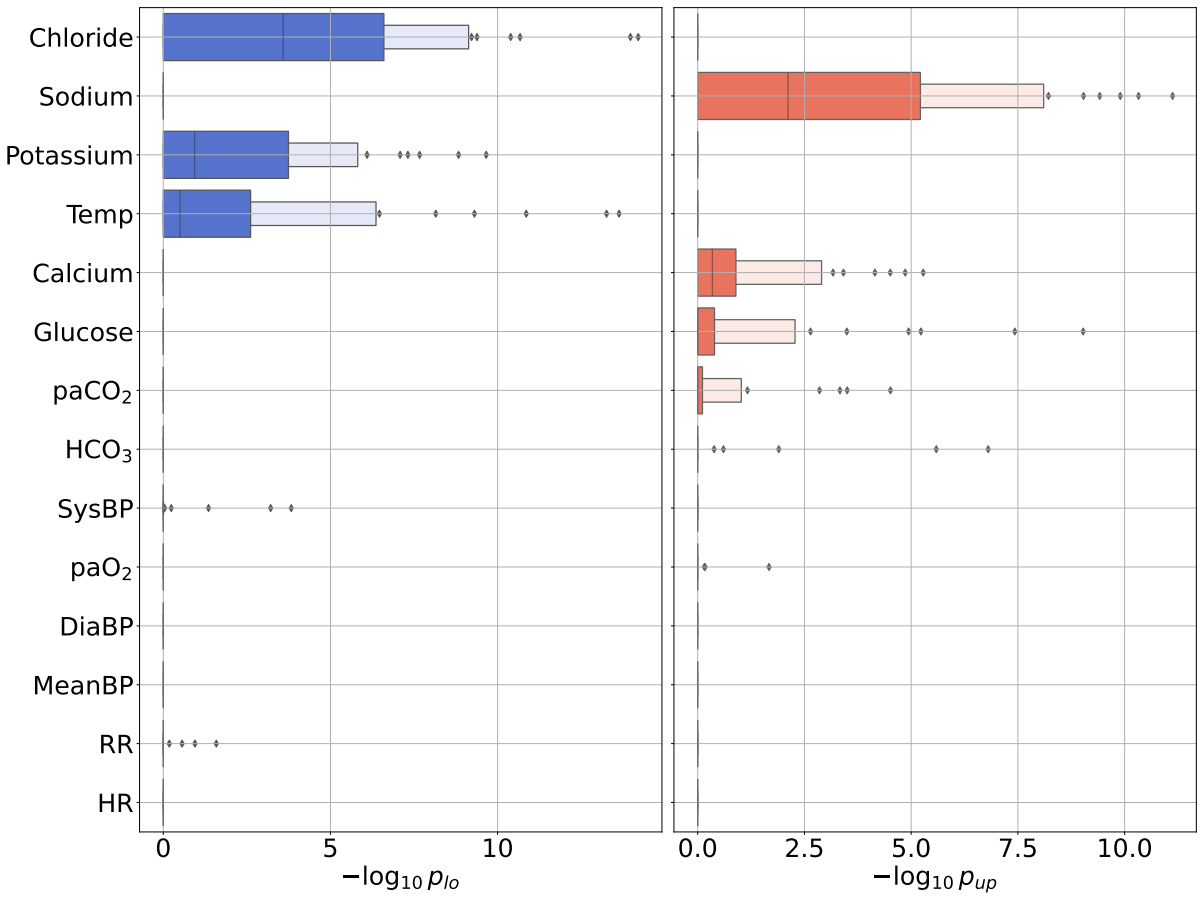


Figure 7: Boxenplots showing distributions of $-\log_{10} p_{\text{lo}}$ and $-\log_{10} p_{\text{up}}$ for different physiological quantities obtained via Hoeffding's inequality. Higher values indicate greater evidence in favour of rejection.

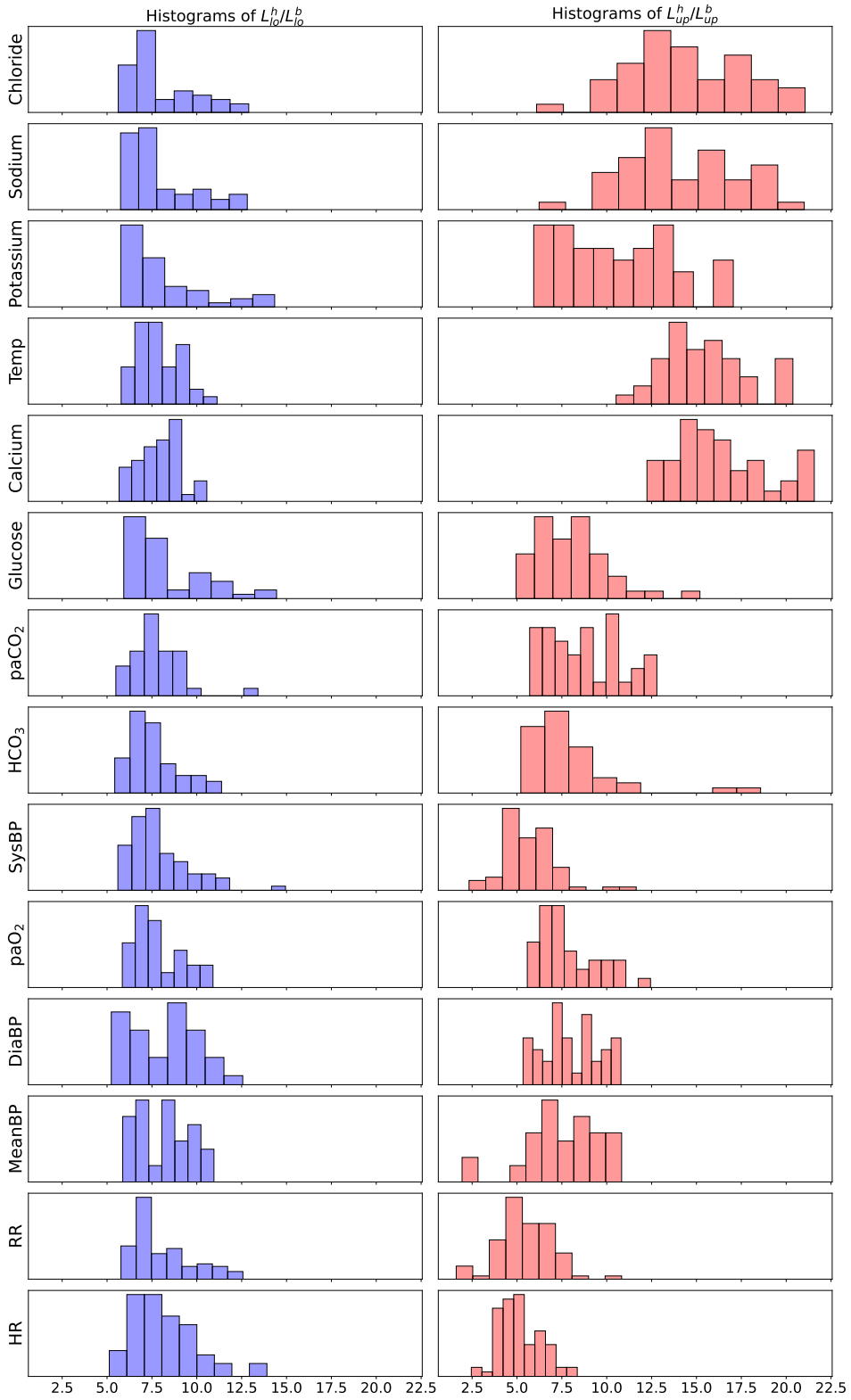
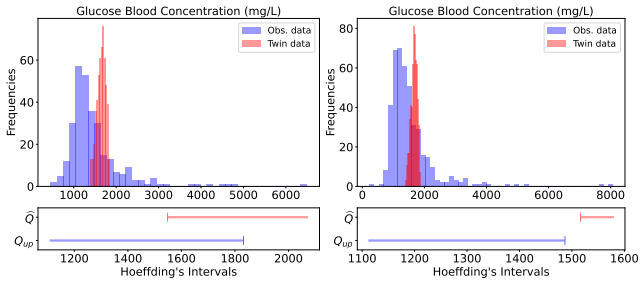
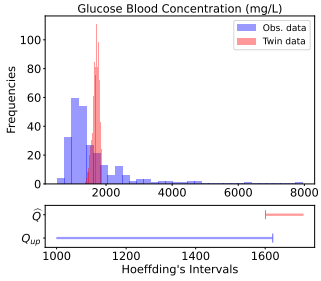


Figure 8: Histograms of $L^{\mathrm{h}}_{\mathrm{lo}}/L^{\mathrm{b}}_{\mathrm{lo}}$ and $L^{\mathrm{h}}_{\mathrm{up}}/L^{\mathrm{b}}_{\mathrm{up}}$ for different hypotheses and outcomes Y. Here, $L^{\mathrm{h}}_{\mathrm{lo}}$ and $L^{\mathrm{b}}_{\mathrm{lo}}$ denote the lengths of intervals $[y_{\mathrm{lo}}, R^{\alpha}_{\mathrm{up}}]$ obtained using Hoeffding's inequality and bootstrapping respectively. Likewise, $L^{\mathrm{h}}_{\mathrm{up}}$ and $L^{\mathrm{b}}_{\mathrm{lo}}$ correspond to the lengths of $[R^{\alpha}_{\mathrm{lo}}, y_{\mathrm{up}}]$.

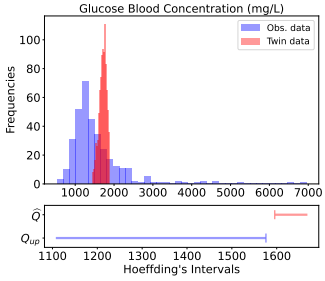


(a) Not rejected

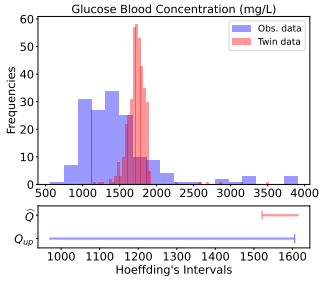
(b) Rejected



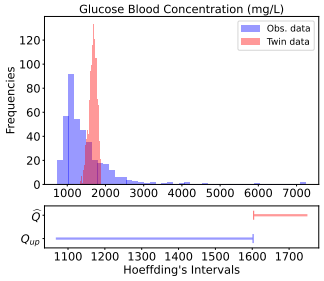
(c) Not rejected



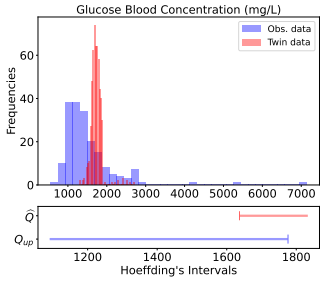
(d) Rejected



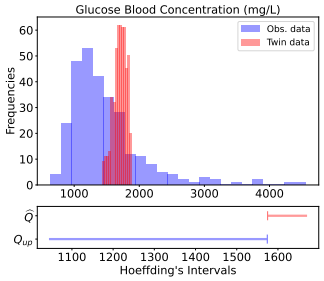
(e) Not rejected



(f) Rejected



(g) Not rejected



(h) Rejected

Figure 9: 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 $\widehat{X}_{0:t}(a_{1:t}) \in B_{0:t}$. Each row shows two distinct choices of $(B_{0:t}, a_{1:t})$. Below each figure are shown 95% Hoeffding confidence intervals for $\widehat{Q}$ 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.

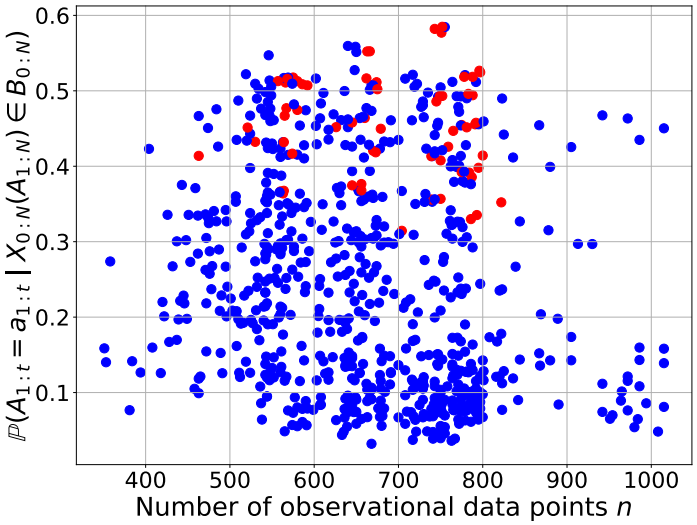


Figure 10: Sample mean estimate of $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$ for each pair of hypotheses $(\mathcal{H}_{\text{lo}}, \mathcal{H}_{\text{up}})$ corresponding to the same set of parameters $(t, f, a_{1:t}, B_{0:t})$ that we tested, along with the corresponding number of observational data points used to test each hypothesis. Red points indicate that either $\mathcal{H}_{\text{lo}}$ or $\mathcal{H}_{\text{up}}$ were rejected, while blue points indicate that both $\mathcal{H}_{\text{lo}}$ and $\mathcal{H}_{\text{up}}$ were not rejected.

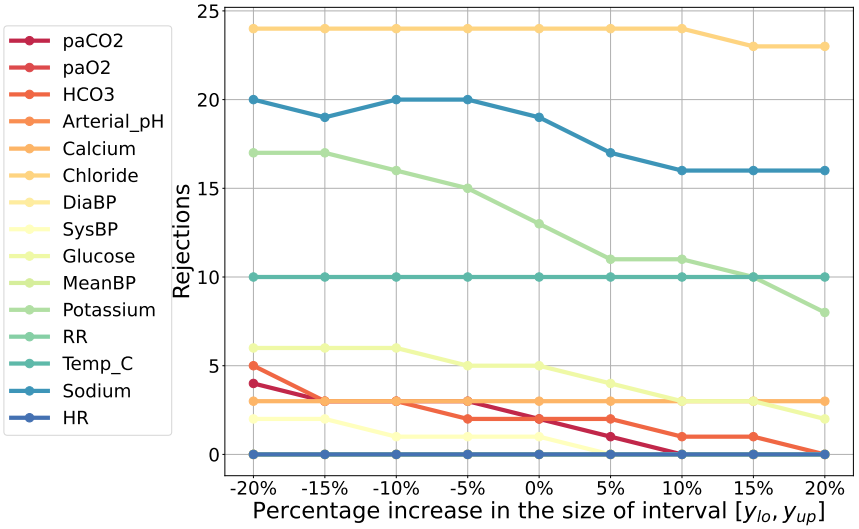


Figure 11: Rejections obtained as the width of the $[y_{\text{lo}}, y_{\text{up}}]$ interval changes. Here, the interval is increased (or decreased) symmetrically on each side.