Causal Falsification of Digital Twins

Rob Cornish*, **Muhammad Faaiz Taufiq***, Arnaud Doucet, Chris Holmes

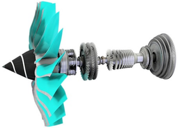
Department of Statistics, University of Oxford

June 25, 2023

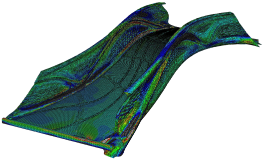
*Equal contribution

Motivation

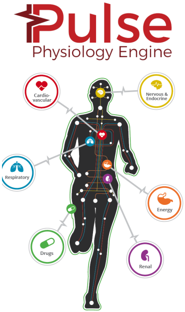
Simulators called Digital Twins are increasingly used to guide safety-critical decision-making



rolls-royce.com



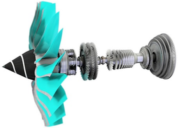
turing.ac.uk



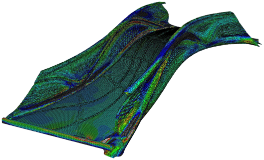
pulse.kitware.com

Motivation

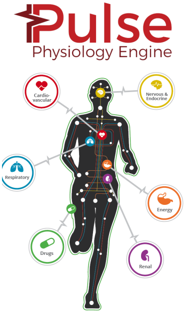
Simulators called Digital Twins are increasingly used to guide safety-critical decision-making



rolls-royce.com



turing.ac.uk



pulse.kitware.com

In these environments, the accuracy of a twin is paramount

High-level goal

Our question: Often large datasets taken from the underlying phenomena are available

How can we use this data to assess the accuracy of a given twin?

High-level goal

Our question: Often large datasets taken from the underlying phenomena are available

How can we use this data to assess the accuracy of a given twin?

Constraints: Assessment procedure itself must be reliable:
$\Rightarrow$ Prefer soundness over completeness

Want a procedure that can realistically scale to real twins
$\Rightarrow$ Want to make minimal assumptions

Key insight and challenge

An natural approach is to compare directly the output of the twin with observational data

Key insight and challenge

An natural approach is to compare directly the output of the twin with observational data

However, if causal conclusions are sought (e.g. for planning), then this is unsound for most datasets in practice

Motivating example

Toy scenario

Consider modelling effect of drug on weight for some population

Drug interacts with an enzyme $U \in \{0, 1\}$ present in a subpopulation:

- If $U = 1$, drug increases weight
- If $U = 0$, drug has no effect

Suppose drug is only administered when $U = 1$

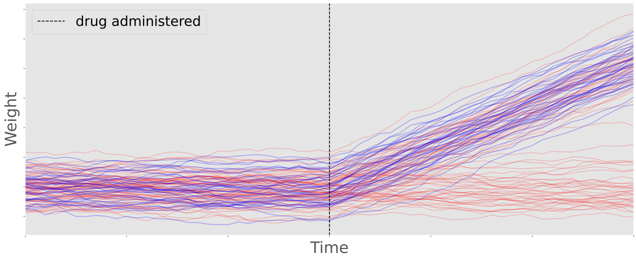
Toy scenario

Consider modelling effect of drug on weight for some population

Drug interacts with an enzyme $U \in \{0, 1\}$ present in a subpopulation:

- If $U = 1$, drug increases weight
- If $U = 0$, drug has no effect

Suppose drug is only administered when $U = 1$



Blue: outcomes that *were* observed for patients administered drug;
Red: outcomes that *would* be observed across whole population

Key point

This phenomenon occurs because the data are confounded

Confounding is well-studied in the causal inference literature

However, implications for simulators are less appreciated

Key point: in general wrong to compare the data with the output of twin under the corresponding actions

Overview

Motivated by this observation, our paper:

- Formulates twin assessment as a causal inference problem
- Argues for an approach based on falsification rather than verification
- Presents a statistical methodology valid under minimal assumptions
- Illustrates via a large-scale case study

Aside: Causal Inference

Overview

Causal inference provides a mathematical framework for reasoning about the causal effects of interventions based on observational data

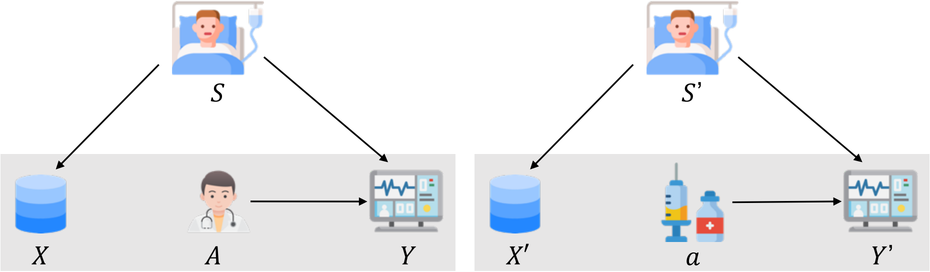
Many questions we care about in practice are of a causal nature

- "What should I do to make things a certain way?" vs. "How do things evolve on their own?"

For this reason, highly suitable for Twins, for which decision-making and acting in the world are primary concerns

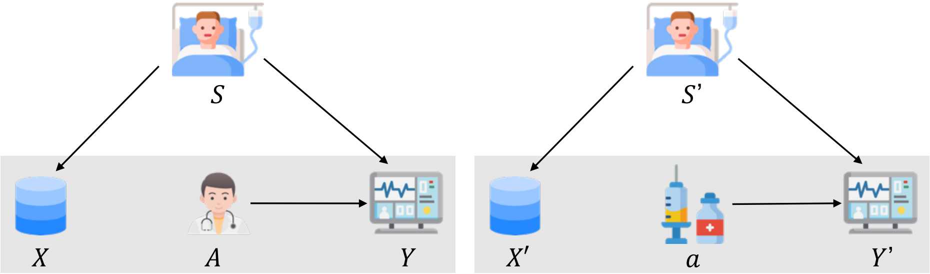
A Typical Problem

Straightforward problem: Given distribution of (X, A, Y) from the left-hand system, what is distribution of (X', Y') in the right-hand system?



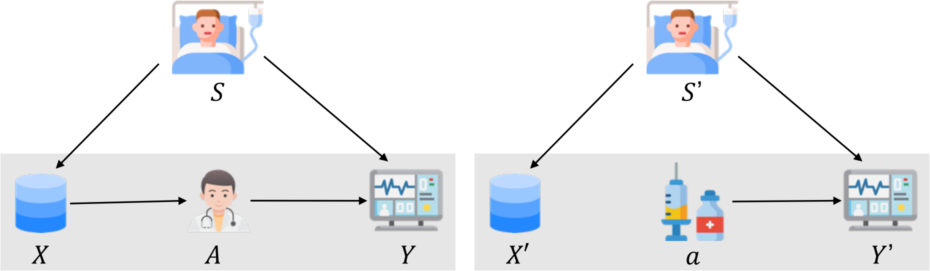
A Typical Problem

Straightforward problem: Given distribution of (X, A, Y) from the left-hand system, what is distribution of (X', Y') in the right-hand system?



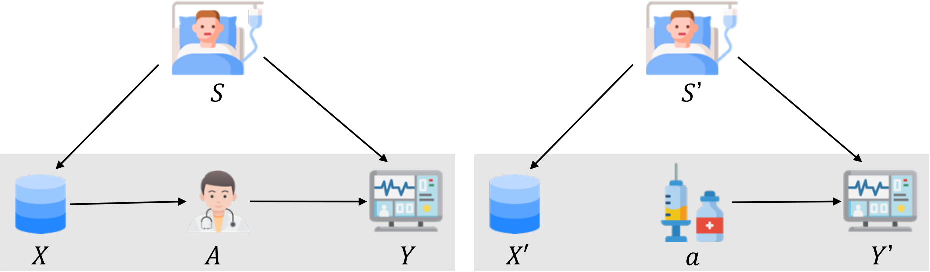
Answer: $P(X' = x, Y' = y)$ on right is $P(X = x, Y = y \mid A = a)$ on left

Given distribution of (X, A, Y) from the left-hand system, what is distribution of (X', Y') in the right-hand system?



More general example

Given distribution of (X, A, Y) from the left-hand system, what is distribution of (X', Y') in the right-hand system?

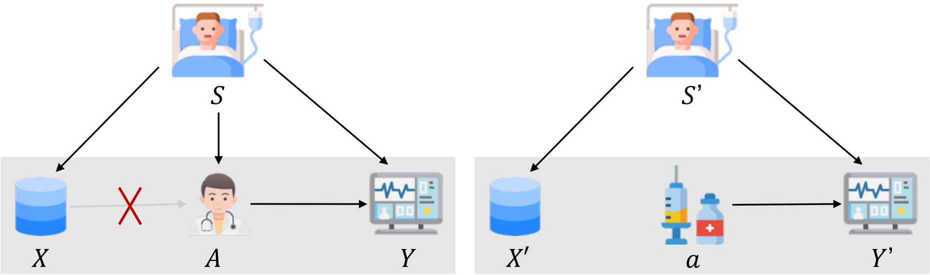


Answer:

$P(X' = x, Y' = y)$ on right is $P(X = x)\, P(Y = y \mid X = x, A = a)$ on left $(\neq P(X = x, Y = y \mid A = a))$

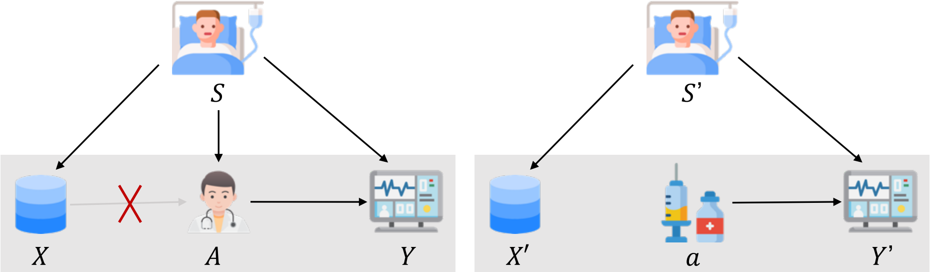
Unidentifiable example

Given distribution of (X, A, Y) from the left-hand system, what is distribution of (X', Y') in the right-hand system?



Unidentifiable example

Given distribution of (X, A, Y) from the left-hand system, what is distribution of (X', Y') in the right-hand system?



Answer: Don't know! (without further assumptions)

Unmeasured confounding

In last case, the data contains unmeasured confounding (cf. second case)

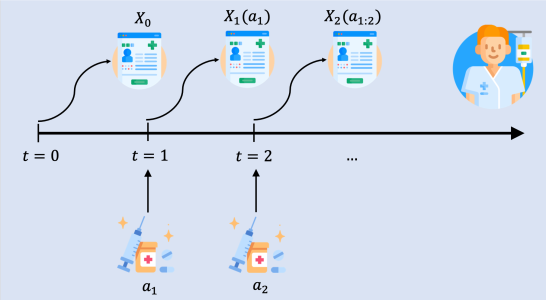
Unmeasured confounding is usually assumed away, but it is in fact extremely common (e.g. U as enzyme from earlier)

For no unmeasured confounding, every factor that affects both A and Y must be included explicitly in the data

- Often tenuous, especially for safety-critical applications

Our Problem Setup

Real World Process

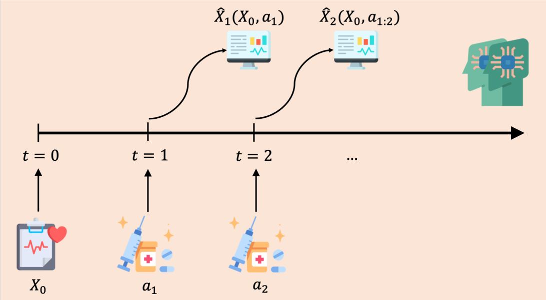


Model real-world process via potential outcomes:

$$X_0, X_1(a_1), X_2(a_{1:2}), \ldots, X_T(a_{1:T}) \qquad \text{for each sequence } a_{1:T} \text{ of actions.}$$

Idea: $X_t(a_{1:t})$ represents what would be observed after actions $a_{1:t}$

Digital Twin Process

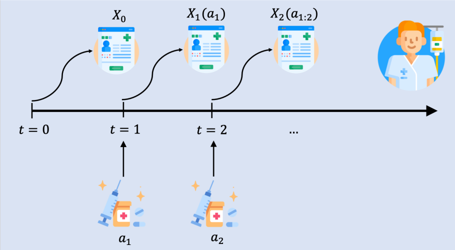


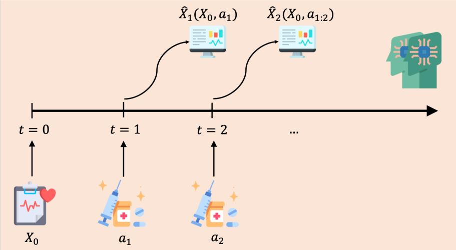
Model twin similarly as

$$\widehat{X}_1(x_0, a_1), \ldots, \widehat{X}_T(x_0, a_{1:T}) \qquad \text{where additionally } x_0 \text{ is an initialisation}$$

Idea: $\widehat{X}_t(x_0, a_{1:t})$ represents output of twin after inputs x_0 and $a_{1:t}$

Interventional Correctness

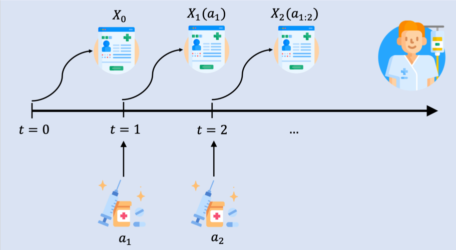


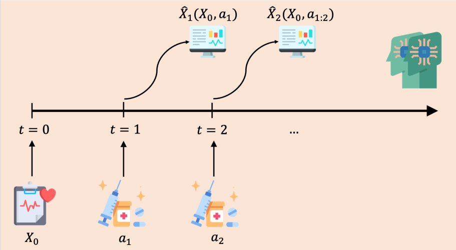


Interventional correctness

Would like the distribution of each $\widehat{X}_{1:t}(x_0, a_{1:t})$ to be equal to the conditional distribution of $X_{1:t}(a_{1:t})$ given $X_0 = x_0$

Interventional Correctness





Interventional correctness

Would like the distribution of each $\widehat{X}_{1:t}(x_0, a_{1:t})$ to be equal to the conditional distribution of $X_{1:t}(a_{1:t})$ given $X_0 = x_0$

$\Rightarrow$ Can recover real-world distribution via Monte Carlo (e.g. for planning)

Data-driven assessment problem

Behavioural agent takes an action A_t at each timestep

Data-driven assessment problem

Behavioural agent takes an action A_t at each timestep

Obtain dataset of i.i.d. copies of

$$X_0, A_1, X_1(A_1), A_2, X_2(A_{1:2}), \ldots, A_T, X_T(A_{1:T})$$

Data-driven assessment problem

Behavioural agent takes an action A_t at each timestep

Obtain dataset of i.i.d. copies of

$$X_0, A_1, X_1(A_1), A_2, X_2(A_{1:2}), \ldots, A_T, X_T(A_{1:T})$$

Goal is to use this dataset to assess whether the twin is interventionally correct

Data-driven assessment problem

Behavioural agent takes an action A_t at each timestep

Obtain dataset of i.i.d. copies of

$$X_0, A_1, X_1(A_1), A_2, X_2(A_{1:2}), \ldots, A_T, X_T(A_{1:T})$$

Goal is to use this dataset to assess whether the twin is interventionally correct

Overall model is intentionally very weak, which seems appropriate for the assessment problem

- Do not assume $X_t(a_{1:t}) \perp\!\!\!\perp A_t \mid X_{0:t-1}(A_{1:t-1}), A_{1:t-1}$ (sequential randomisation assumption, i.e. no unmeasured confounding)

Verification and falsification

Verification approaches

Standard assessment approaches have the following logical structure:

Verification assessment

1. Choose a hypothesis $\mathcal{H}$ such that, if $\mathcal{H}$ is true, then the twin is correct
2. Try to show that $\mathcal{H}$ is true
3. If successful, consider the twin verified

Verification approaches

Standard assessment approaches have the following logical structure:

Verification assessment

1. Choose a hypothesis $\mathcal{H}$ such that, if $\mathcal{H}$ is true, then the twin is correct
2. Try to show that $\mathcal{H}$ is true
3. If successful, consider the twin verified

Problem with this approach:

Theorem

The distribution of $X_{0:t}(a_{1:t})$ is not identifiable from the distribution of $(X_{0:t}(A_{1:t}), A_{1:t})$.

Verification approaches

Standard assessment approaches have the following logical structure:

Verification assessment

1. Choose a hypothesis $\mathcal{H}$ such that, if $\mathcal{H}$ is true, then the twin is correct
2. Try to show that $\mathcal{H}$ is true
3. If successful, consider the twin verified

Problem with this approach:

Theorem

The distribution of $X_{0:t}(a_{1:t})$ is not identifiable from the distribution of $(X_{0:t}(A_{1:t}), A_{1:t})$.

$\Rightarrow$ Does not exist $\mathcal{H}$ with this property whose truth can be determined from the data alone

Our alternative: falsification

We consider the following alternative structure:

Falsification assessment

1. Choose hypotheses $\mathcal{H}$ such that, if the twin is correct, then $\mathcal{H}$ is true
2. Try to show that $\mathcal{H}$ is false
3. If successful, we have determined a failure mode of the twin

Our alternative: falsification

We consider the following alternative structure:

Falsification assessment

1. Choose hypotheses $\mathcal{H}$ such that, if the twin is correct, then $\mathcal{H}$ is true
2. Try to show that $\mathcal{H}$ is false
3. If successful, we have determined a failure mode of the twin

Advantage: can choose $\mathcal{H}$ with this property whose falsity can be determined from data

Our alternative: falsification

We consider the following alternative structure:

Falsification assessment

1. Choose hypotheses $\mathcal{H}$ such that, if the twin is correct, then $\mathcal{H}$ is true
2. Try to show that $\mathcal{H}$ is false
3. If successful, we have determined a failure mode of the twin

Advantage: can choose $\mathcal{H}$ with this property whose falsity can be determined from data

However: lack of falsification does not imply the twin is correct

Hypotheses from causal bounds

Key result

Define real-valued outcomes $Y(a_{1:t}) \coloneqq f(X_{0:t}(a_{1:t}))$ for some f

Key result

Define real-valued outcomes $Y(a_{1:t}) := f(X_{0:t}(a_{1:t}))$ for some f

Fix $a_{1:t}$ and let

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

Key result

Define real-valued outcomes $Y(a_{1:t}) \coloneqq f(X_{0:t}(a_{1:t}))$ for some f

Fix $a_{1:t}$ and let

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

Theorem (Causal bounds)

If $\mathbb{P}(y_{\mathrm{lo}} \le Y(a_{1:t}) \le y_{\mathrm{up}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) = 1$, *then*

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

Key result

Define real-valued outcomes $Y(a_{1:t}) := f(X_{0:t}(a_{1:t}))$ for some f

Fix $a_{1:t}$ and let

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

Theorem (Causal bounds)

If $\mathbb{P}(y_{\mathrm{lo}} \leq Y(a_{1:t}) \leq y_{\mathrm{up}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) = 1$, *then*

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

Key point: left and right-hand sides are identifiable (in fact, unbiasedly) from observational data

Intuition

Theorem (Causal bounds)

If $\mathbb{P}(y_{\mathrm{lo}} \leq Y(a_{1:t}) \leq y_{\mathrm{up}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) = 1$, *then*

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

Intuition

Theorem (Causal bounds)

If $\mathbb{P}(y_{\mathrm{lo}} \leq Y(a_{1:t}) \leq y_{\mathrm{up}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) = 1$, *then*

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

Take $B_{0:t}$ to be the whole space and recall

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

Lower bound becomes:

$$\mathbb{E}[Y(a_{1:t})] \geq \mathbb{E}[\mathbb{I}(A_{1:t} = a_{1:t})\, Y(A_{1:t}) + \mathbb{I}(A_{1:t} \neq a_{1:t})\, y_{\mathrm{lo}}]$$

Intuition

Theorem (Causal bounds)

If $\mathbb{P}(y_{\mathrm{lo}} \leq Y(a_{1:t}) \leq y_{\mathrm{up}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) = 1$, *then*

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

Take $B_{0:t}$ to be the whole space and recall

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

Lower bound becomes:

$$\mathbb{E}[Y(a_{1:t})] \geq \mathbb{E}[\mathbb{I}(A_{1:t} = a_{1:t})\, Y(A_{1:t}) + \mathbb{I}(A_{1:t} \neq a_{1:t})\, y_{\mathrm{lo}}]$$

Essentially, choose worst-case for unseen subpopulation.

Intuition

Theorem (Causal bounds)

If $\mathbb{P}(y_{\mathrm{lo}} \leq Y(a_{1:t}) \leq y_{\mathrm{up}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) = 1$, *then*

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

Take $B_{0:t}$ to be the whole space and recall

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

Lower bound becomes:

$$\mathbb{E}[Y(a_{1:t})] \geq \mathbb{E}[\mathbb{I}(A_{1:t} = a_{1:t})\, Y(A_{1:t}) + \mathbb{I}(A_{1:t} \neq a_{1:t})\, y_{\mathrm{lo}}]$$

Essentially, choose worst-case for unseen subpopulation.
Corresponds to Manski [1990] (cf. Zhang and Bareinboim [2019])

Optimality of bounds

- Without further assumptions, these bounds cannot be improved upon for general $Y(a_{1:t})$ (or if $Y(a_{1:t}) = f(X_t(a_{1:t}))$)

Optimality of bounds

- Without further assumptions, these bounds cannot be improved upon for general $Y(a_{1:t})$ (or if $Y(a_{1:t}) = f(X_t(a_{1:t}))$)
- Also, cannot bound $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})]$ nontrivially if $X_{1:t}(a_{1:t})$ is continuous

Derived hypotheses

The twin is interventionally correct iff $(X_0, \widehat{X}_{1:T}(X_0, a_{1:T})) \overset{\mathrm{d}}{=} X_{0:T}(a_{1:T})$

Derived hypotheses

The twin is interventionally correct iff $(X_0, \widehat{X}_{1:T}(X_0, a_{1:T})) \stackrel{\mathrm{d}}{=} X_{0:T}(a_{1:T})$

Therefore, if the twin is interventionally correct,

$$\mathbb{E}[Y(a_{1:t}) \mid X_{1:t}(a_{1:t}) \in B_{1:t}] = \underbrace{\mathbb{E}[\widehat{Y}(a_{1:t}) \mid X_0 \in B_0, \widehat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}]}_{=:\widehat{Q}}$$

Derived hypotheses

The twin is interventionally correct iff $(X_0, \widehat{X}_{1:T}(X_0, a_{1:T})) \stackrel{\mathrm{d}}{=} X_{0:T}(a_{1:T})$

Therefore, if the twin is interventionally correct,

$$\mathbb{E}[Y(a_{1:t}) \mid X_{1:t}(a_{1:t}) \in B_{1:t}] = \underbrace{\mathbb{E}[\widehat{Y}(a_{1:t}) \mid X_0 \in B_0, \widehat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}]}_{=:\widehat{Q}}$$

Let Q_{lo} and Q_{up} be causal bounds from earlier
$\Rightarrow$ If the twin is interventionally correct, then $\mathcal{H}_{\mathrm{lo}}$ and $\mathcal{H}_{\mathrm{up}}$ hold, where

$$\mathcal{H}_{\mathrm{lo}} : Q_{\mathrm{lo}} \leq \widehat{Q} \qquad\qquad \mathcal{H}_{\mathrm{up}} : \widehat{Q} \leq Q_{\mathrm{up}}$$

(Note dependence on $(t, f, a_{1:t}, B_{0:t})$)

Derived hypotheses

The twin is interventionally correct iff $(X_0, \widehat{X}_{1:T}(X_0, a_{1:T})) \stackrel{\mathrm{d}}{=} X_{0:T}(a_{1:T})$

Therefore, if the twin is interventionally correct,

$$\mathbb{E}[Y(a_{1:t}) \mid X_{1:t}(a_{1:t}) \in B_{1:t}] = \underbrace{\mathbb{E}[\widehat{Y}(a_{1:t}) \mid X_0 \in B_0, \widehat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}]}_{=:\widehat{Q}}$$

Let Q_{lo} and Q_{up} be causal bounds from earlier
$\Rightarrow$ If the twin is interventionally correct, then $\mathcal{H}_{\mathrm{lo}}$ and $\mathcal{H}_{\mathrm{up}}$ hold, where

$$\mathcal{H}_{\mathrm{lo}} : Q_{\mathrm{lo}} \leq \widehat{Q} \qquad\qquad \mathcal{H}_{\mathrm{up}} : \widehat{Q} \leq Q_{\mathrm{up}}$$

(Note dependence on $(t, f, a_{1:t}, B_{0:t})$)

Interpretation: (e.g.) if $\mathcal{H}_{\mathrm{lo}}$ is false, then when $(X_0, \widehat{X}_{1:t}(X_0, a_{1:t})) \in B_{0:t}$, the outputs $f(X_0, \widehat{X}_{1:t}(X_0, a_{1:t}))$ are on average too small

Statistical methodology

High-level overview

Consider testing a given $\mathcal{H}_{\mathrm{lo}} : Q_{\mathrm{lo}} \leq \widehat{Q}$

Recall: we have an observational dataset of i.i.d. copies of

$$X_0, A_1, X_1(A_1), A_2, X_2(A_{1:2}), \ldots, A_T, X_T(A_{1:T}).$$

For given $a_{1:t}$, generate dataset of i.i.d. copies of

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

High-level overview

Consider testing a given $\mathcal{H}_{\text{lo}} : Q_{\text{lo}} \leq \widehat{Q}$

Recall: we have an observational dataset of i.i.d. copies of

$$X_0, A_1, X_1(A_1), A_2, X_2(A_{1:2}), \ldots, A_T, X_T(A_{1:T}).$$

For given $a_{1:t}$, generate dataset of i.i.d. copies of

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

Use e.g. Hoeffding's inequality to obtain one-sided conf. intervals R^{α}_{lo}, $\widehat{R}^{\alpha}$,

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

and reject $\mathcal{H}_{\text{lo}}$ if $\widehat{R}^{\alpha} < R^{\alpha}_{\text{lo}}$, or return a p-value

Other aspects

Control for multiple testing via e.g. Holm-Bonferroni or Benjamini-Yekutieli

Other aspects

Control for multiple testing via e.g. Holm-Bonferroni or Benjamini-Yekutieli

Can choose parameters $(t, f, a_{1:t}, B_{0:t})$ for each $\mathcal{H}_{\mathrm{lo}}$ and $\mathcal{H}_{\mathrm{up}}$ in a data-dependent way, provided we use sample splitting

- Useful e.g. for y_{lo} and y_{up}

Other aspects

Control for multiple testing via e.g. Holm-Bonferroni or Benjamini-Yekutieli

Can choose parameters $(t, f, a_{1:t}, B_{0:t})$ for each $\mathcal{H}_{\mathrm{lo}}$ and $\mathcal{H}_{\mathrm{up}}$ in a data-dependent way, provided we use sample splitting

- Useful e.g. for y_{lo} and y_{up}

No additional assumptions required by construction

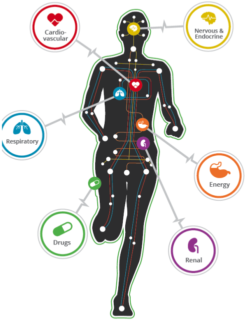
Case study: Pulse Physiology Engine

Pulse Physiology Engine

We apply our methodology to Pulse Physiology Engine, an open source computational model designed for human physiology simulation

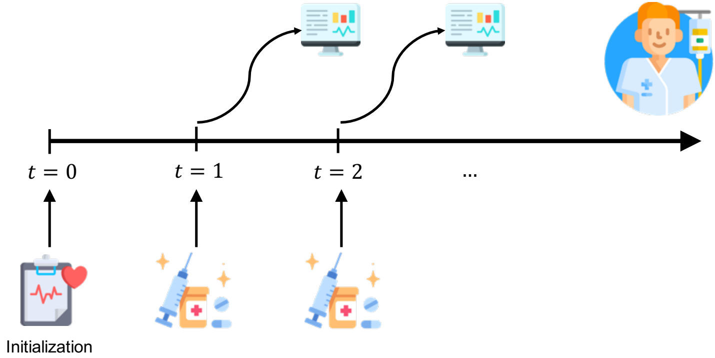
Validate using the MIMIC-III dataset, generated from 40,000+ ICU patients at Beth Israel Hospital

Pulse Physiology Engine



pulse.kitware.com

Pulse Physiology Engine

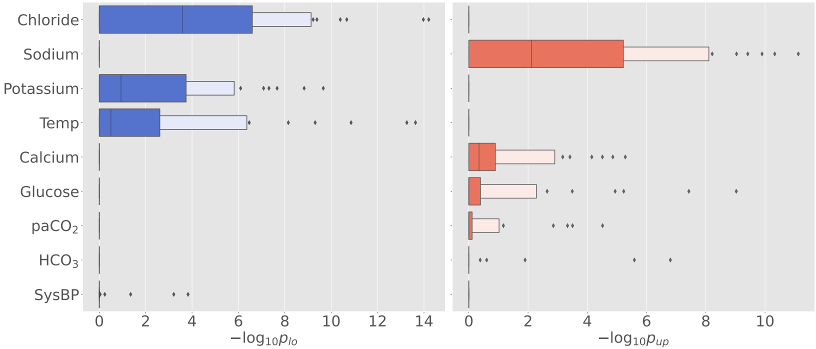


Results

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: Overall rejections (FWER $= 0.05$)

Additional granularity

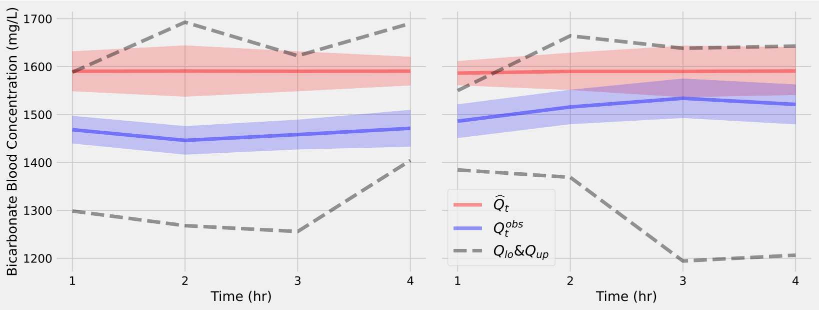


p-values for physiological quantities some rejections (notice consistent over/underestimation)

Pitfalls of naive twin assessment

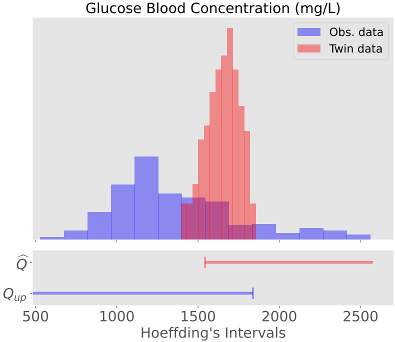
For two separate choices of $(a_{1:t}, B_{1:t})$, compare

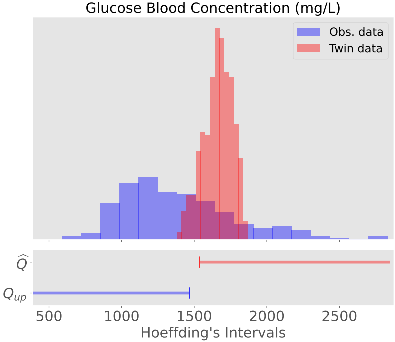
$$\widehat{Q}_t := \mathbb{E}[\widehat{Y}(a_{1:t}) \mid \widehat{X}_{0:t}(a_{1:t}) \in B_{0:t}],$$
$$Q_t^{\text{obs}} := \mathbb{E}[Y(A_{1:t}) \mid X_{0:t}(A_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t}].$$



Left case looks worse, but in fact only right case leads to some rejection

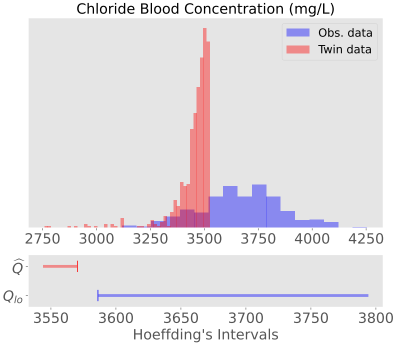
Pitfalls of naive twin assessment (2)

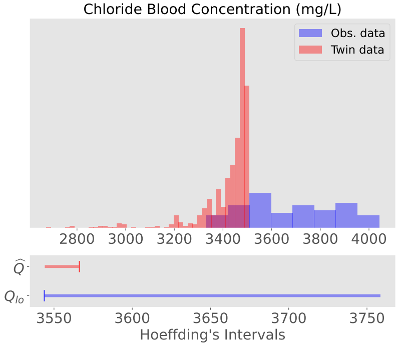




Despite apparent similarity, right hypothesis is rejected but left one is not

Pitfalls of naive twin assessment (3)





Despite apparent similarity, right hypothesis is rejected but left one is not

Thank you!



Joint work with Rob Cornish, Arnaud Doucet, and Chris Holmes

References I

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

Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. *Advances in Neural Information Processing Systems*, 32, 2019.