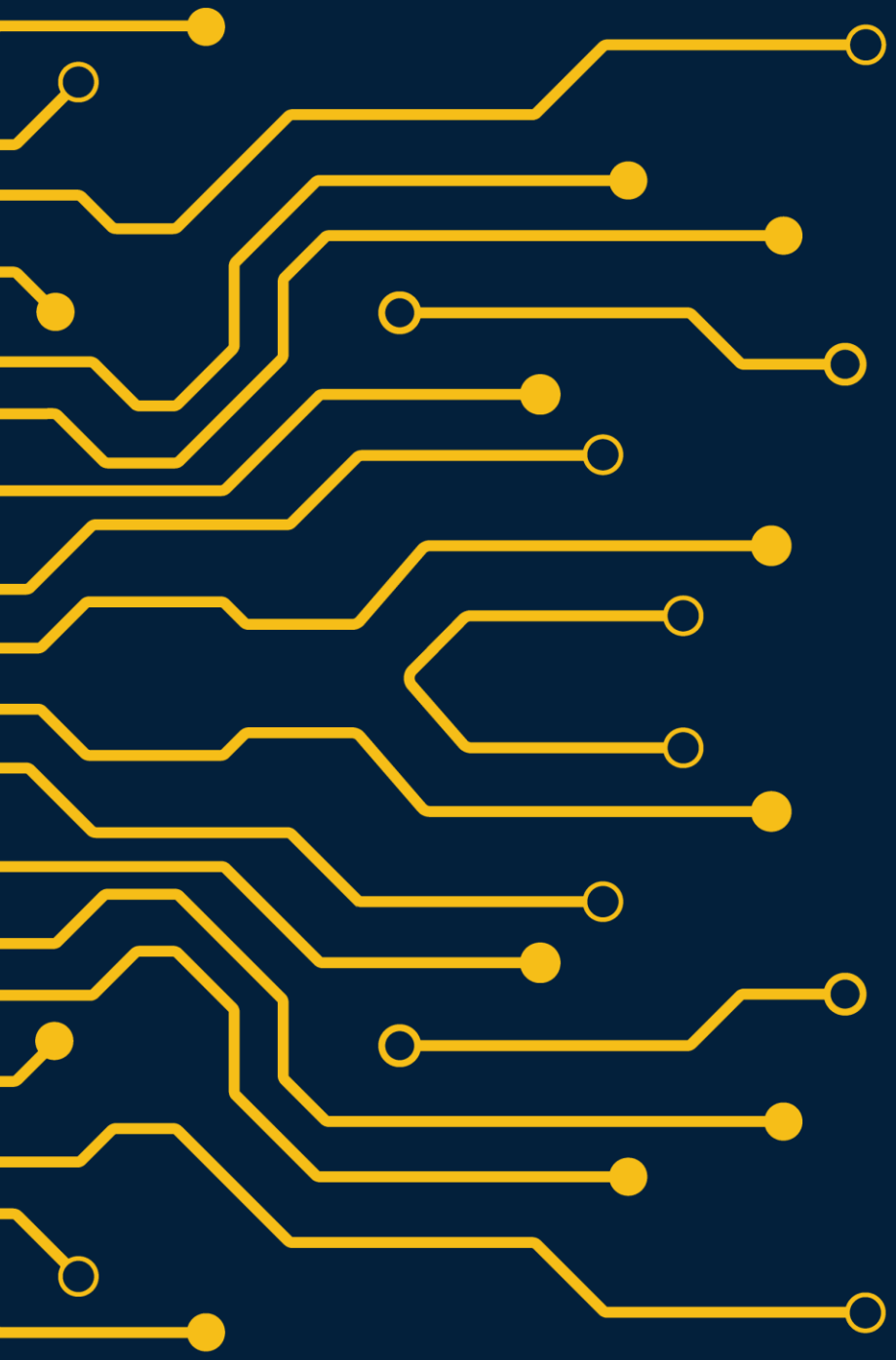


MRCT CENTER

THE MULTI-REGIONAL CLINICAL TRIALS CENTER OF BRIGHAM AND WOMEN'S HOSPITAL AND HARVARD

Data Quality, Model Validation, and Governance for AI Digital Twins and Synthetic Data

Now On-Demand

Disclaimer

The views expressed today do not imply endorsement or reflect the views or policies of Mass General Brigham, Harvard Medical School, or any affiliated organization or entity.

The MRCT Center is supported by voluntary contributions from foundations, corporations, international organizations, academic institutions, and government entities (see www.MRCTCenter.org), as well as by grants.

We are committed to autonomy in our research and to transparency in our relationships. The MRCT Center and its directors retain responsibility and final control of the content of any products, results, and deliverables.

Welcome~

Thank you for joining this webinar today!

Tips for today's session:

- Please use the Q&A for questions. We will do our best to answer live.
- Feel free to use the Closed Captioning available on the Zoom toolbar.
- Most of the links in our presentations will be shared in the Chat.

This meeting is being recorded.

The recording, slides, and any additional materials will be available next week. If you registered, you will receive an email about their availability and a notification about future webinars in this series.

The Multi-Regional Clinical Trials Center (MRCT Center)

The MRCT Center is a research and policy center focused on addressing the conduct, oversight, ethics and regulatory environment for clinical trials.

Our Vision

Improve the integrity, safety, and rigor of global clinical trials.

Our Mission

Engage diverse stakeholders to define emerging issues in global clinical trials and to create and implement ethical, actionable, and practical solutions.

AI and Clinical Research

https://mrctcenter.org/project/ethics-ai/

The **Framework for Review of Clinical Research Involving AI** offers IRB and other oversight bodies a structured, practical approach to evaluating protocols that involve AI in research with human participants.

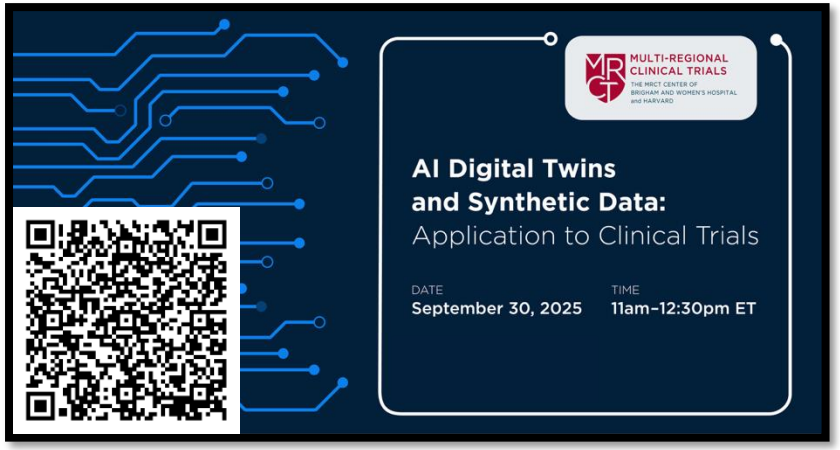
Available to download

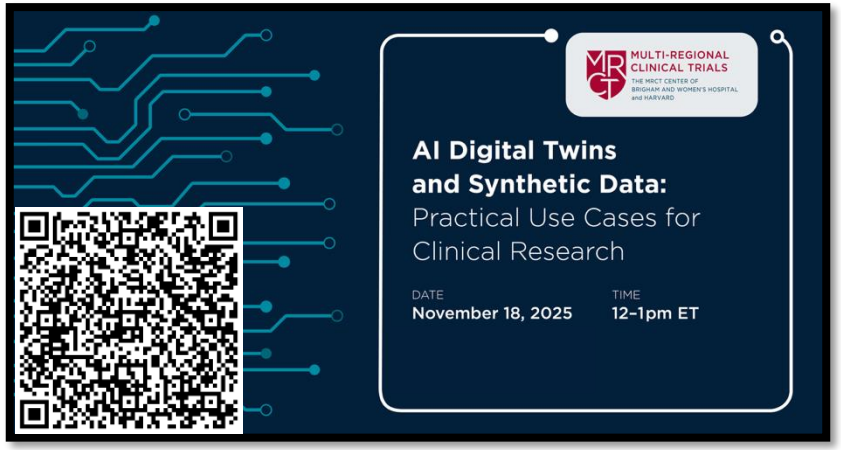
Ethical considerations during IRB review:

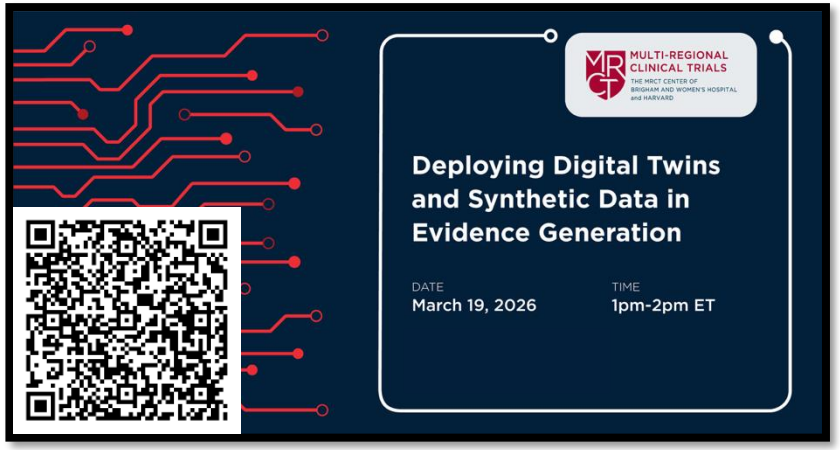
- Human Agency and Oversight
- Technical Robustness and Safety
- Privacy, Confidentiality, and Data Governance
- Transparency
- Representativeness and Fairness
- Informed Consent

Now: AI in the Administration of Research

AI Digital Twins and Synthetic Data Webinar Series







- Today's session is the fourth in the MRCT Center's AI Digital Twins and Synthetic Data webinar series
- The first three sessions covered:
  - An introduction to digital twins and how they are built for clinical research
  - Specific use cases for clinical research of digital twins and synthetic data
  - The current landscape of evidence generation expectations and benchmarks
- On-demand recordings and slides are available on the MRCT Center's website

Data Quality, Model Validation, and Governance for AI Digital Twins & Synthetic Data

MULTI-REGIONAL CLINICAL TRIALS
THE MRCT CENTER OF BRIGHAM AND WOMEN'S HOSPITAL and HARVARD

Barbara E. Bierer
Professor of Medicine, HMS
Faculty Director,
MRCT Center of BWH & Harvard
bbierer@bwh.harvard.edu

Daniele Bertolini
Machine Learning Scientist

UNLEARN

Tina Morrison
Vice President, Scientific Strategy
EQTY Labs

EQTYLAB

Chao-Yi Wu
Assistant Professor of Neurology
Massachusetts General Hospital

MGH 1811

https://mrctcenter.org/resource/ai-digital-twins-and-synthetic-data-practical-use-cases-for-clinical-research/

Digital Twins

A digital twin of a patient is a computational model of disease progression that **represents a specific subject** and predicts their expected clinical trajectory under defined conditions (e.g., standard of care).

- The twin is typically linked to the real patient through bidirectional information flow: patient characteristics are used to generate the twin's predictions, which can subsequently be compared with the patient's observed outcomes.
- AI-generated digital twins are typically produced using machine learning models trained on historical clinical data, rather than purely mechanistic models.

Key Properties

- Individual-level counterfactual:* corresponds to a specific trial participant
- Predictive: Estimates outcomes for that patient under alternative scenarios

* relating to or expressing what has not happened or is not the case.

Example Use Cases

- **In Clinical Trials:**
  - Improving trial efficiency (e.g., increasing statistical power or reducing sample size
  - Creating synthetic control arms in trials without a randomized control group
- **Outside clinical trials:**
  - Individualized medicine applications, like guiding treatment choices. But we won't cover these use cases

Synthetic Data

Synthetic clinical data are artificially generated patient records designed to reproduce the statistical properties of real datasets, but that **do not correspond to specific real individuals.**

Key Properties

- No one-to-one linkage to real patients: synthetic records represent population-level patterns rather than predictions for specific individuals
- Static generation: there is no direct dynamic interaction between synthetic records and real-world patients.

Example Use Cases

- **In Clinical Trials:**
  - Simulate and optimize different trial designs
- **Outside clinical trials:**
  - Generating privacy-preserving datasets for research and data sharing
  - Augmenting datasets used to train machine learning models

Risk-Informed Credibility Assessment: From Computational Models, Synthetic Data to Digital Twins

Tina Morrison, PhD
tina.morrison@eqtylab.io

EQTYLAB

A presentation for MRCT's Series on
Digital Twins and Synthetic Evidence

MULTI-REGIONAL CLINICAL TRIALS
THE MRCT CENTER of
BRIGHAM AND WOMEN'S HOSPITAL
and HARVARD

Regulatory Science as the foundation for TRUST

Regulatory science is the science of developing tools, standards, and approaches to assess the safety, efficacy, quality, and performance of all FDA-regulated products. [FDA]

Credibility is the trust,

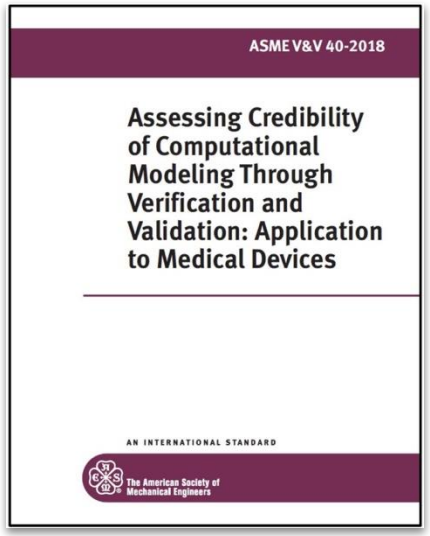
obtained through the collection of evidence,

in the *predictive capability* of a model

for a context of use. (ASME V&V40)

Understands Approach & Evidence Assessment

Conferred by

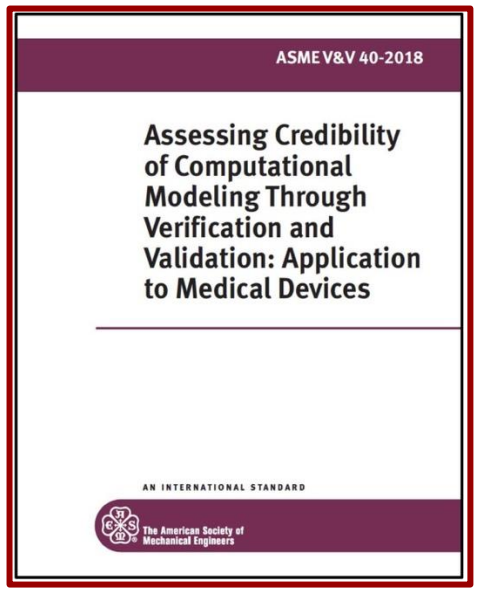


Stakeholders:

- R&D Teams
- Business Leaders
- Regulatory Affairs
- Regulators
- ….

Original Concept:
Jeff Bodner,
Medtronic

Relevant Guidance that harness the Credibility Assessment Framework



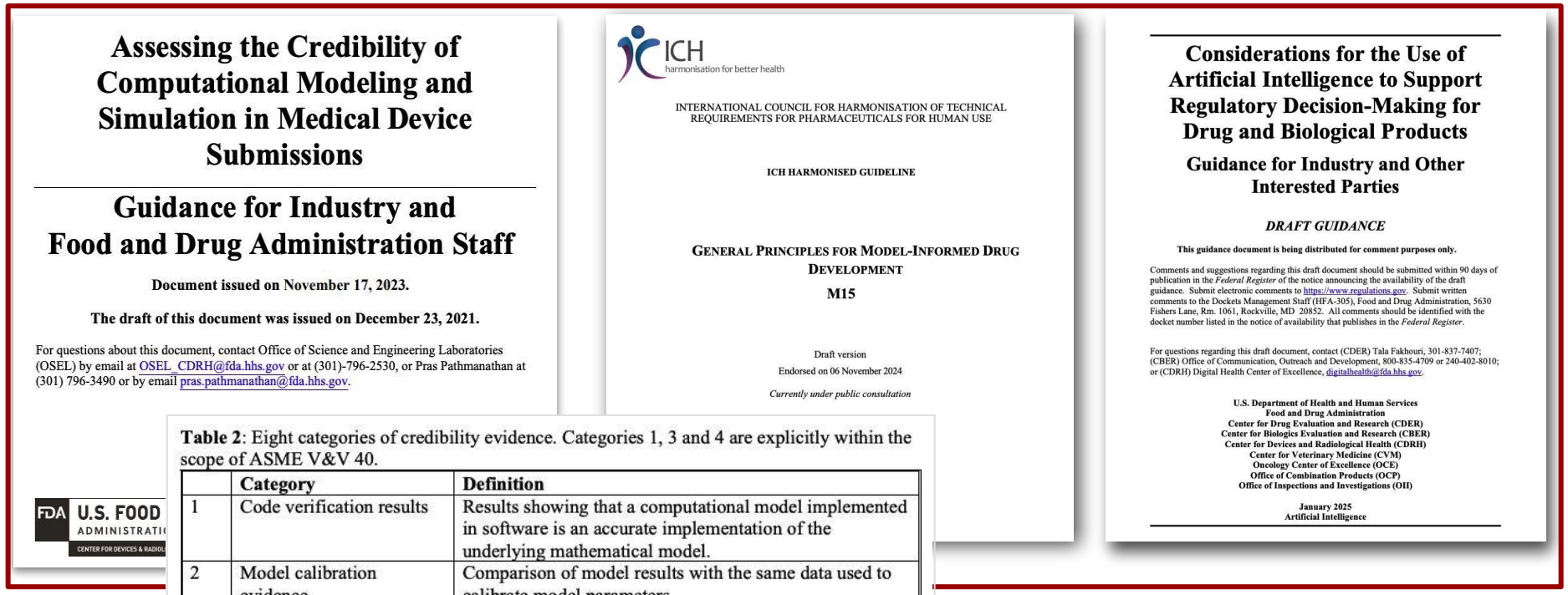


Table 2: Eight categories of credibility evidence. Categories 1, 3 and 4 are explicitly within the scope of ASME V&V 40.

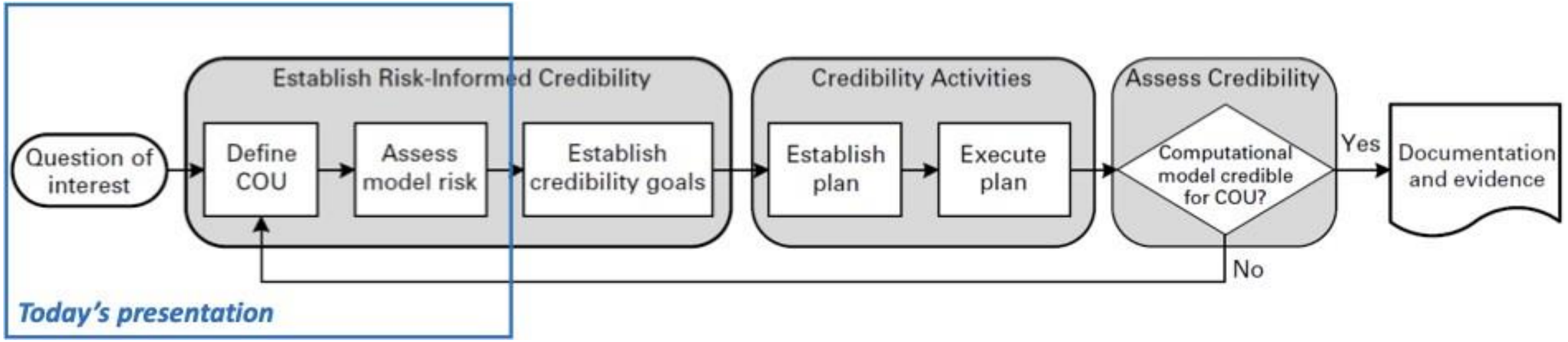
	Category	Definition
1	Code verification results	Results showing that a computational model implemented in software is an accurate implementation of the underlying mathematical model.
2	Model calibration evidence	Comparison of model results with the same data used to calibrate model parameters.
3	Bench test validation results	Validation results using a bench test comparator. May be supported by calculation verification and/or UQ results using the validation conditions.
4	*In vivo* validation results	Same as previous category except using *in vivo* data as the comparator.
5	Population-based validation results	Comparison of population-level data between model predictions and a clinical data set. No individual-level comparisons are made.
6	Emergent model behavior	Evidence showing that the model reproduces phenomena that are known to occur in the system at the specified conditions but were not pre-specified or explicitly modeled by the governing equations.
7	Model plausibility evidence	Rationale supporting the choice of governing equations, model assumptions, and/or input parameters only.
8	Calculation verification /UQ results using COU simulations	Calculation verification and/or UQ results obtained using the COU simulations, that is, the simulations performed to answer the question of interest

Why is the context of use important?

Typically, there are many evidence sources on the table informing regulatory decisions.

Each source of evidence is not necessarily weighed equally; depends on its role; depends on its scope; depends on risk.

Foundation of the Risk-informed Credibility Assessment Framework



1. State the decision or question of interest that is being informed by the model.
2. Define the context of use (COU) for the model, which defines the specific role and scope of the model to address the question of interest.
   a. This is *important* because there are likely going to be other sources of evidence to inform the decision being made.
3. Assess the risk associated with using the model, which represents the *possibility* that the use of the model leads to a decision that results in patient harm and/or other undesirable impacts.

Remarks on Context of Use

- COU describes how the model will be used to address the question of interest, i.e., the specific role and scope.
  - Alongside the COU under development should be a description of additional evidence sources that will also be used to inform the question of interest (e.g., in vitro tests, in vivo data).
- From experience, ambiguity in the question of interest and COU can result in
  - reluctance to accept model in a given drug development or regulatory review scenario or
  - an undesirably protracted dialogue between drug developers and regulators on the data requirements needed to establish model credibility.
- It is, therefore, critical to unambiguously and explicitly state the question of interest and how the proposed modeling approach will address the COU.

Example from drug development

A small molecule drug is in clinical development for the treatment of a chronic, non-life-threatening symptomatic condition that affects people of all ages.

Planned clinical studies include assessment of PK and long-term safety and efficacy in adults, adolescents, and children.

The drug is primarily eliminated by cytochrome P450 (CYP) 3A4 and has a broad therapeutic window.

Clinical drug–drug interaction (DDI) studies in adults demonstrate that drug PK are affected by strong CYP3A4 modulators such that patients require altered dosing.

Citation: CPT Pharmacometrics Syst. Pharmacol. (2020) 9, 21–28; doi:10.1002/psp4.12479

1. Question of Interest and 2. Context of Use

Because the drug model will serve multiple purposes, there are two questions of interest, each with a different COU.

Question of interest 1: How should the investigational drug be dosed when coadministered with CYP3A4 modulators?

COU 1: The PBPK model will predict the effects of weak and moderate CYP3A4 inhibitors and inducers on the PK of the investigational drug in adult patients. Simulated peak plasma concentration (Cmax) and area under the plasma concentration-time curve (AUC) ratios of the investigational drug after a single dose and at steady state will be used to *provide dosing recommendations for adults in labeling* without the need for additional clinical DDI studies.

Question of interest 2: What is the optimal labeled dose for pediatric patients?

COU 2: Relevant physiological parameters will be changed in the adult PBPK model to predict plasma concentration-time course and exposure metrics in adolescents and children. Predictions at steady state will be used *to inform the starting dose for pediatric patients* in a clinical trial assessing the PK, efficacy, and safety of the investigational drug. The results of the trial will determine the final labeled dose.

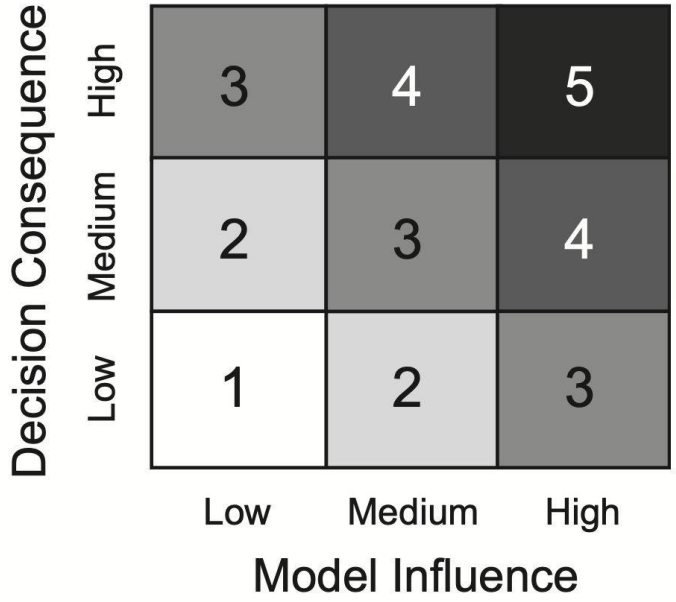
3. Model (use) Risk

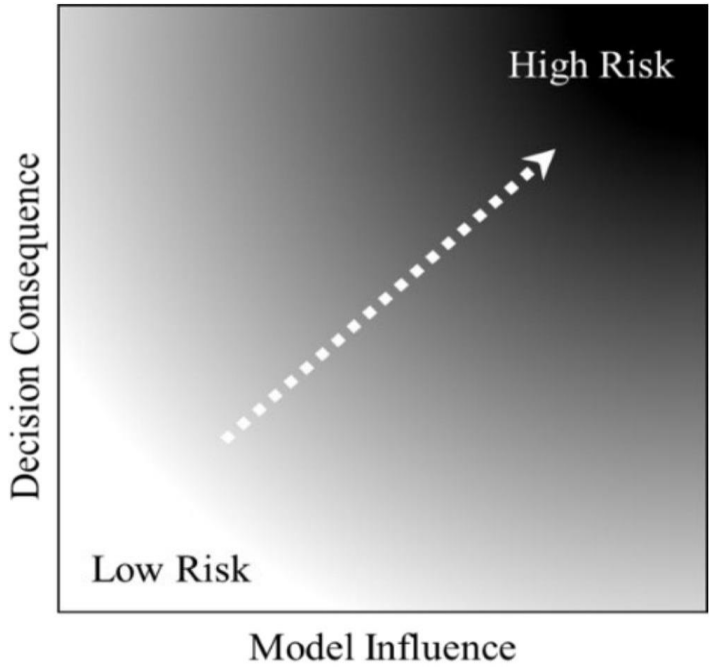
*Start with: **decision consequence**, which describes the significance of an incorrect decision based on all available evidence.

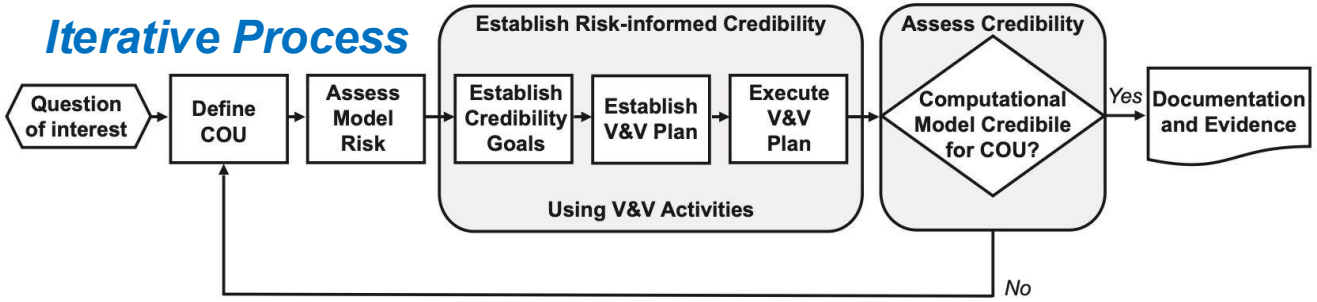
→ NOT consequence of an incorrect model

Then examine the **model influence**, the contribution of the model output as compared to all available evidence in addressing the question of interest.

→ model influence is not static; the rigor of the evidence should drive its influence in the decision.

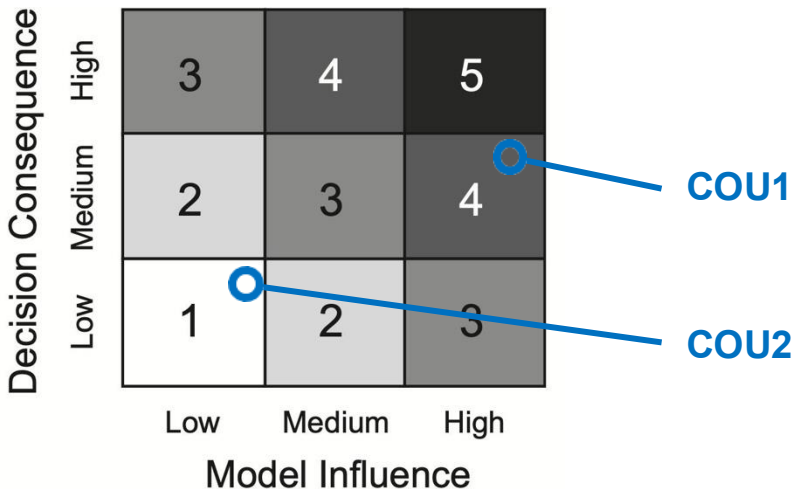






Example from drug development

	Context of use 1	Context of use 2
Question of interest	How should the investigational drug be dosed when coadministered with CYP3A4 modulators?	What is the optimal labeled dose for pediatric patients?
Context of use	–Simulation to predict effects of weak and moderate CYP3A4 modulators on investigational drug PK –Predictions will be used for dosing recommendations in label –No DDI studies proposed with weak and moderate CYP3A4 modulators; have clinical data with strong CYP3A4 modulators	–Simulation to predict investigational drug PK in children and adolescents –Prediction will be used to inform starting dose for clinical trial –Final labeled dose will be based on clinical trial data in pediatric patients
Model risk	High	Low
Model influence	High: –Model provides substantial evidence –Limited clinical data from similar scenarios to support the decision	Low: –Model provides minor evidence –Primary evidence for labeled dose is pediatric clinical trial
Decision consequence	Medium: –Incorrect decision could result in minor to moderate adverse patient outcomes	Low: –Incorrect decision would not result in adverse outcomes in patient safety or efficacy

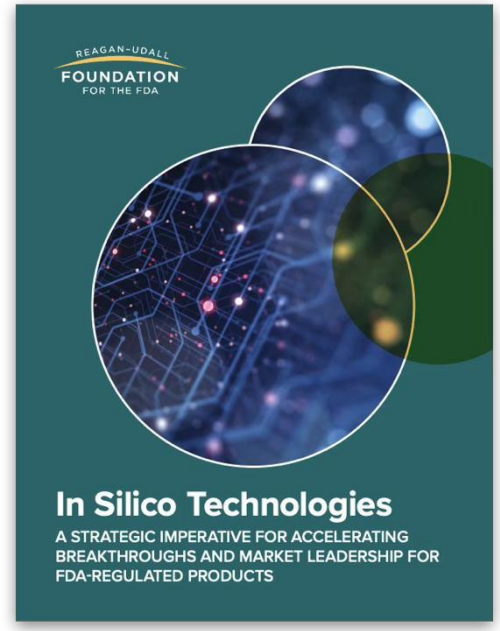


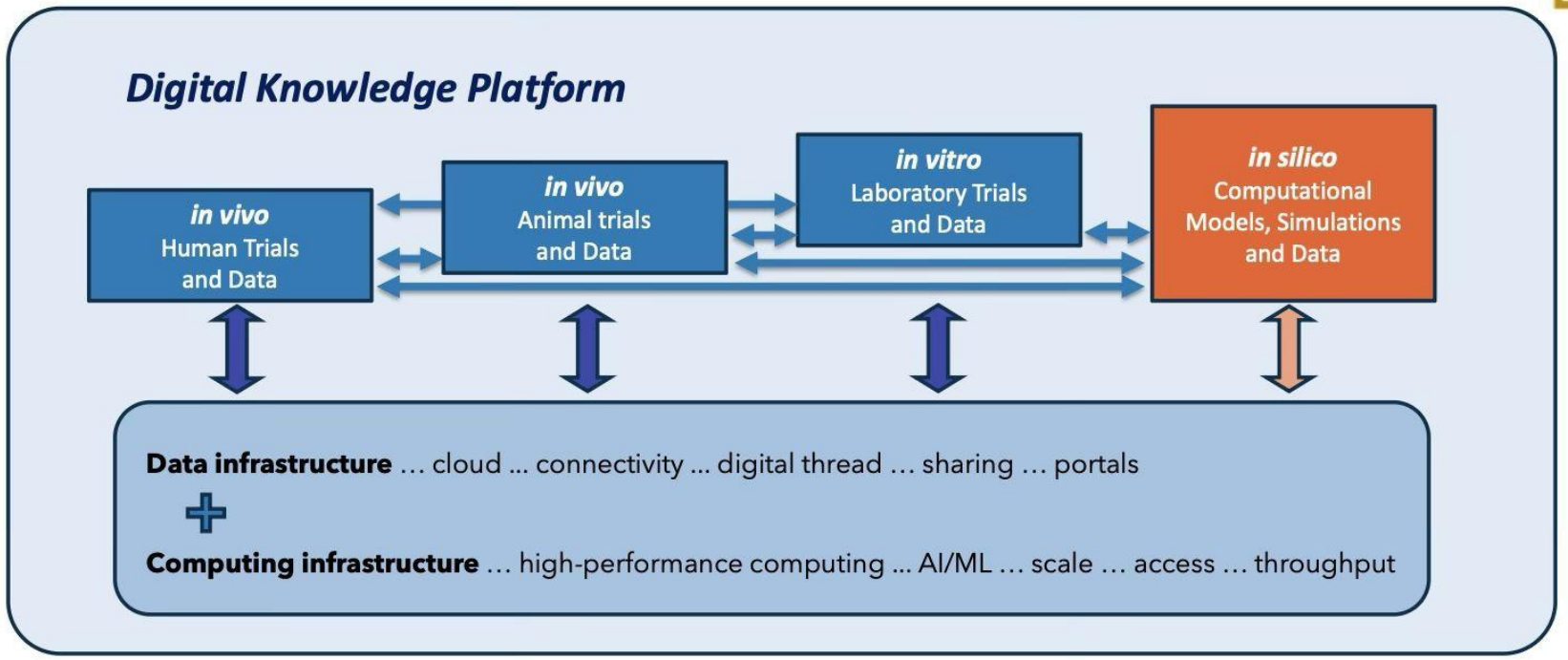
Report for Business Leaders and Regulatory Affairs

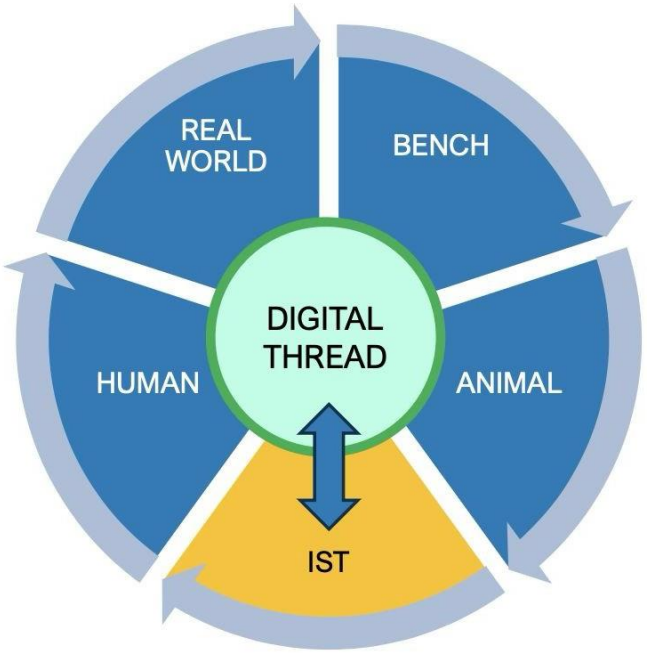
Written by industry experts, 2024

Role of *In Silico* Technologies

- *Screen before human testing*
- Complementary evidence
- Accelerate therapeutic development
- Achieve first-pass success on standards tests
- Reduce cost, time, human & animal testing
- Drive affordable and safe therapies







Adapted from FDA Reagan Udall Foundation, "In Silico Technologies: A strategic imperative for accelerating breakthroughs and market leadership for FDA-Regulated products"

Additional COU Discussion for *In Silico Trials Methodologies Applied to 3Rs*

The rigor of evidence should be commensurate with the the role of the model to inform the decision.

2021 Oct;25(10):3977-3982.
doi: 10.1109/JBHI.2021.3090469.
Epub 2021 Oct 5.

	Reduce	Refine	Replace
Preclinical In Vitro/Ex Vivo Experiments	Reduce the number or duration of in vitro/ex vivo experiments	Improve the predictive accuracy of safeness and/or effectiveness provided by the in vitro or ex vivo experiment	Replace entirely a portion or all the required in vitro or ex vivo experiments
Preclinical Animal Experiments	Reduce the number of animals involved in the experiment, or its duration	Alleviate the suffering of the animals involved, or improve the predictive accuracy of the safeness and/or effectiveness provided by the animal experiment	Replace animal experiments in the prediction of the expected safety and/or efficacy for a new treatment during the clinical experimentation
Clinical Human Experiments	Reduce the number of humans involved in the experiment, or its duration	Reduce the risks for the humans involved, or improve the predictive accuracy of the safeness and/or effectiveness provided by the human trials	Replace human experiments in the prediction of the expected safety and/or efficacy for a new treatment during real-world, post-marketing use

UNLEARN + MULTI-REGIONAL CLINICAL TRIALS
THE MRCT CENTER of BRIGHAM AND WOMEN'S HOSPITAL and HARVARD

May 18, 2026

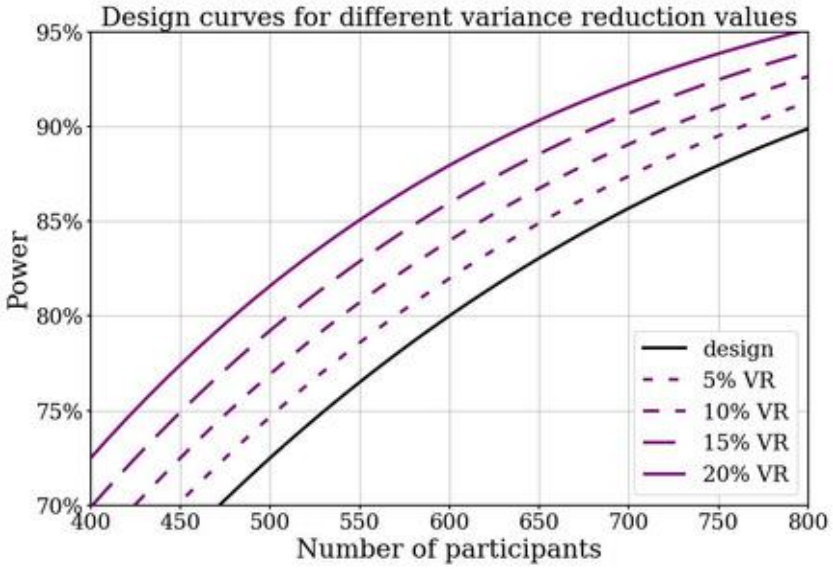
Digital Twin Models Validation

Daniele Bertolini
Unlearn.AI

1. Define the Question of Interest	→	What is the specific decision or clinical/scientific question the AI model will inform?
2. Define the Context of Use	→	How and when the model will be used in the development process
3. (COU) Assess the AI Model Risk	→	Evaluate how much the model output will influence decisions and resulting potential harms
4. Develop a Credibility Plan	→	Plan how you will gather evidence to support the model’s trustworthiness
5. Execute the Plan	→	Run the planned validation, testing, analyses
6. Document Results & Deviations	→	Create a “Credibility Assessment Report” detailing what was done, results, and any deviations from the original plan
7. Determine Adequacy for COU	→	Decide whether model is credible enough for its intended use

Example: digital twins as “super covariates” to increase power in a randomized clinical trial.

Digital twins increase power in an RCT



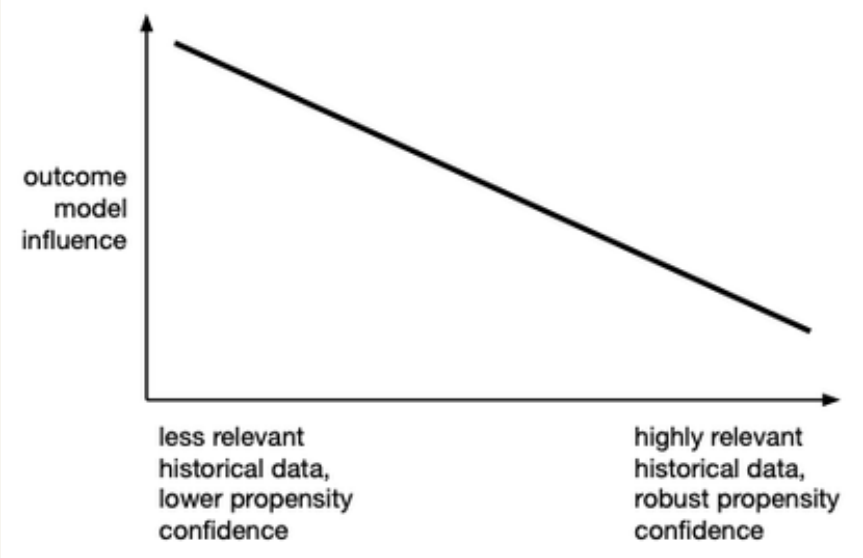
PROCOVA. AI-generated digital twins of trial participants used as covariates to **increase power** –i.e., the probability of detecting a treatment effect if there is one. Digital twins "explain" a portion of the outcome variance and reduce the residual uncertainty. The larger the variance reduction, the larger the power boost (or sample size reduction).

Schuler, Alejandro, et al. "Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score." The International Journal of Biostatistics 18.2 (2021): 329-356.

- **Question of interest**
  - Unchanged, is the drug safe / effective?
- **Context of use**
  - Narrow, use digital twins as covariates. They do not alter trial conduct, endpoints, eligibility.
- **Risks**
  - **Statistical**: minimal, if it's not prognostic, you lose one degree of freedom.
  - **Operational**: data/model lineage control. Only use pre-randomization data, use a locked, version-controlled pipeline. Traceable, auditable, reproducible.
- **Credibility Plan**
  - Determine whether it is beneficial by testing model predictions on past completed trials. The key metric is **correlation** between model predictions (digital twins) and observed outcomes.

Example: digital twins as synthetic controls in a single-arm trial.

Digital twins as synthetic controls



Synthetic Controls. Digital twins of treated participants can be used as an synthetic comparator in a single-arm trial. They predict how each participant would have progressed under standard of care. It is often helpful to construct a synthetic control by **combining these predictions with natural history/historical control data (doubly robust estimators)**. The influence of the digital twin model depends on the relevance/quality of these historical data.

Bertolini et al. Digital Twins as Synthetic Controls in Single-Arm Trials. arXiv: 2605.12832.

- **Question of interest**
  - Is the drug effective?
- **Context of use**
  - Digital twin models predict outcomes under a defined standard of care for the endpoints of interest. We consider methods that combine both model predictions and historical data (doubly robust).
- **Risks**
  - **Decision consequence**, regardless of the model/method, is invariably high: bias in the treatment effect estimate translates directly into an incorrect efficacy conclusion.
  - **Model influence**, by contrast, depends sensitively on the relevance of the historical data that's used in combination with digital twins (and on the quality of the propensity model used to reweight it).
- **Credibility plan**
  - Evaluate model bias and relevance of historical data on relevant past trials/data.

The Role of Training Data

- The data used to *train* digital twin models does not need to be restricted to the target population of interest
- Models trained on large and diverse datasets learn generalizable representations that can then be fine-tuned and adapted to the target population, typically with improved predictive performance (**transfer learning**)
- For example, for an Alzheimer's disease trial than enrolls participants with low severity, training a model on a broader population that includes also more severe patients will typically improve predictive performance on the the mild population
- It is crucial to properly document training data and protocols. However, the core question for AI model credibility assessments is *evaluation* of the trained model for its intended context of use (using appropriate test data, metrics, etc), as mentioned in the two examples we discussed

UNLEARN + MULTI-REGIONAL CLINICAL TRIALS
THE MRCT CENTER of BRIGHAM AND WOMEN'S HOSPITAL and HARVARD

Thank you.

Building a Goal-Oriented, Synthetic Comparator for Early-Stage Alzheimer's Trials

Chao-Yi Wu
Assistant Professor of Neurology
Massachusetts General Hospital
Harvard Medical School
chwu3@mgh.harvard.edu

My view on digital twins: clinical trials

- Involved in 10+ neurodegenerative diseases clinical trials.
- Across case studies, crossover, single arm, parallel group RCT designs as PI, Co-I, and statistician.

Go/no-go decisions

If go, who should be the target population (responders)?

- The persistent challenge across all of them: the comparator problem.

THE BOTTLENECK

Why Current AD Trials Struggle

Patient heterogeneity

Mixed pathologies and comorbidity

Placebo Ethics Crisis

With Lecanemab & Donanemab approved, pure placebo controls are increasingly unjustifiable for vulnerable patients

CONTROL GROUPS

Three approaches to constructing a comparator in trials.

Matched Control

What it is
Real patients from historical data matched to treated patients on observed characteristics (e.g., propensity score).

Key assumption
No unmeasured confounding: treatment assignment is as good as random given matched covariates.

Counterfactual level
Real untreated people who resemble the treated patient.

In Silico Control

What it is
A population-level computational model used to simulate expected outcomes without treatment.

Key assumption
The simulation model correctly captures disease dynamics; model misspecification directly introduces bias.

Counterfactual level
A forward prediction from population parameters.

Digital Twin

What it is
A population backbone model calibrated to an individual's own data (biomarkers, imaging, history) to simulate their personal counterfactual.

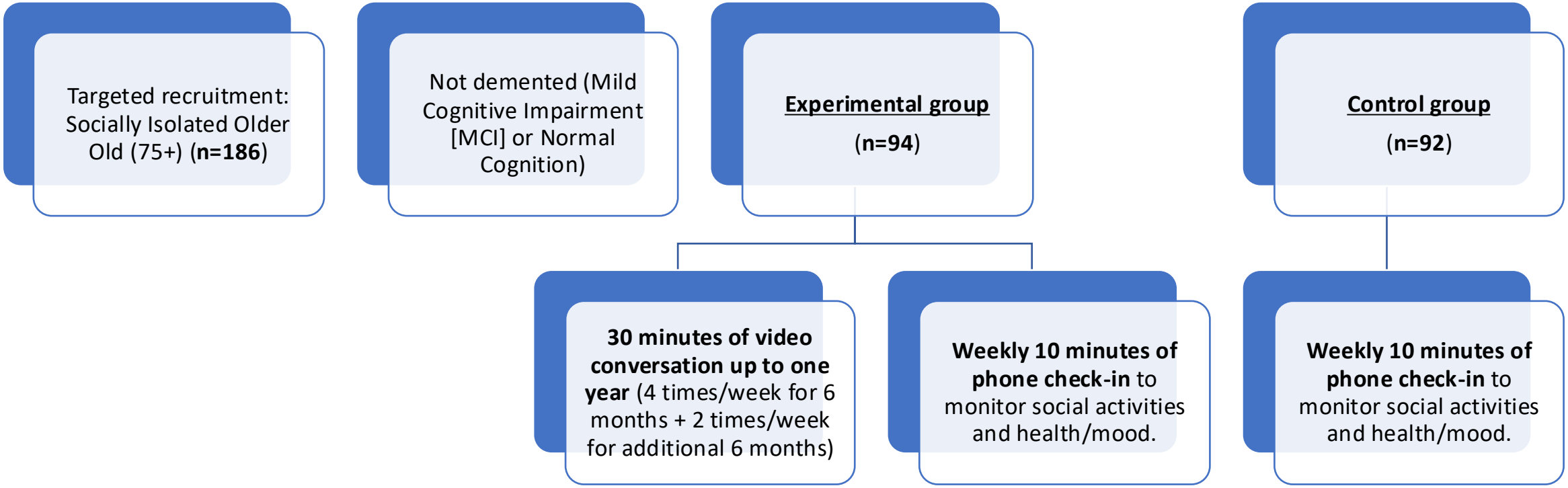
Key assumption
The backbone model is well-validated; individual calibration does not overfit to sparse personal data.

Counterfactual level
The simulated trajectory is specific to patient, updated as new data arrive.

Goal oriented definition of digital twins in the trial context: a simulated version of a participant under a different condition, built to answer a specific trial question.

PROOF-OF-CONCEPT

A completed phase II RCT, with outcomes being cognition (MoCA, animal naming)



PI: Hiroko Dodge

I. NIA R01 AG0033581 (2010-2014) Completed (ClinicalTrials.gov: NCT01571427)
II. NIA R01 AG051628 (2016- 2021) Completed (Normal)
III. NIA R01 AG056102 (2017- 2022) Completed (MCI)
www.i-conect.org (ClinicalTrials.gov: NCT02871921)

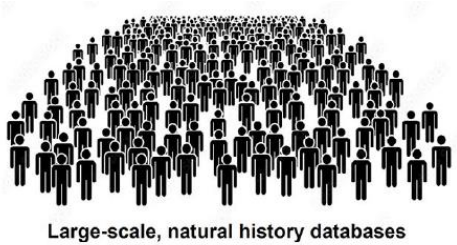
SOLUTIONS

Results -- Matched Control

1 **Mine Historical Data**

2 **Matched control;**
Propensity; distance matching

3 **Clinical validation**
Check group allocation; E-values



Large-scale, natural history databases

Satisfy eligibility for a given RCT

NACC Dataset as of March 28, 2024
(n = 50,259)

NACC UDS version 3
(n = 27,418)

MoCA total score > 0 & <= 30
(n = 23,310)

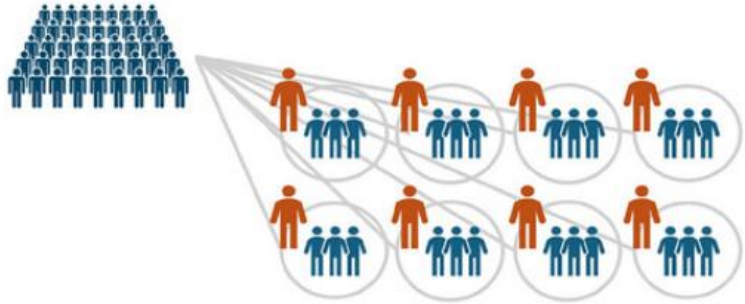
Category Fluency Animal > 0 & < 80
(n = 22,956)

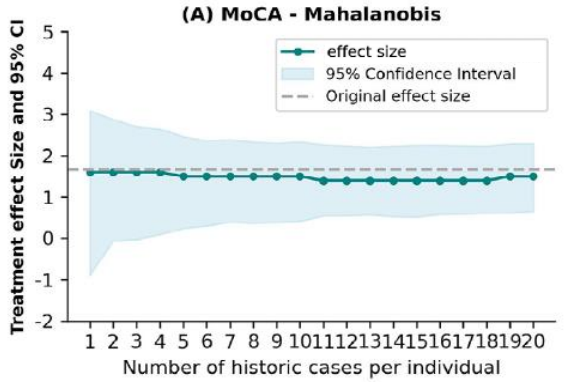
Age >=75
(n = 9,856)

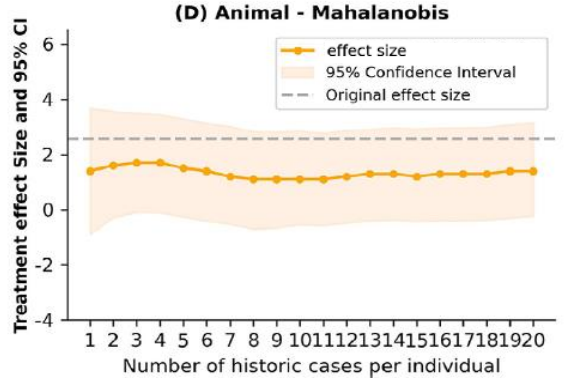
GDS-15 <= 7
(n = 9,316)

CDR = 0 or 0.5
(n = 7,935)

Data with more than 1 follow-ups
(n = 5,132)







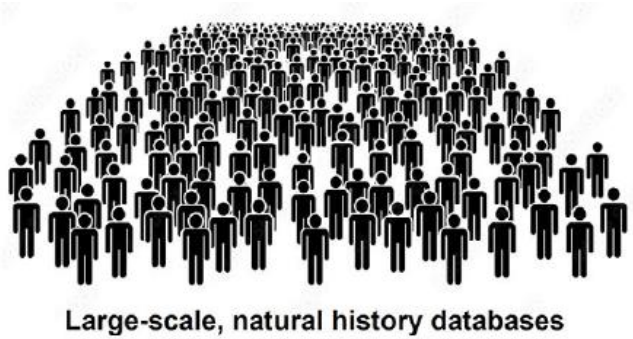
SOLUTIONS

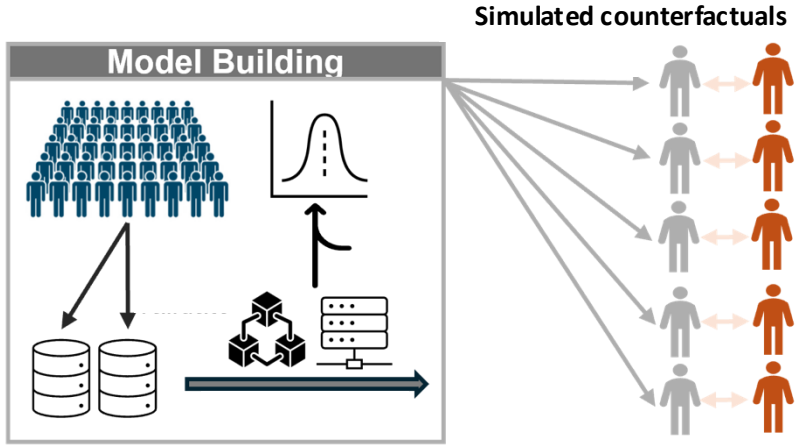
Results -- In Silico Controls

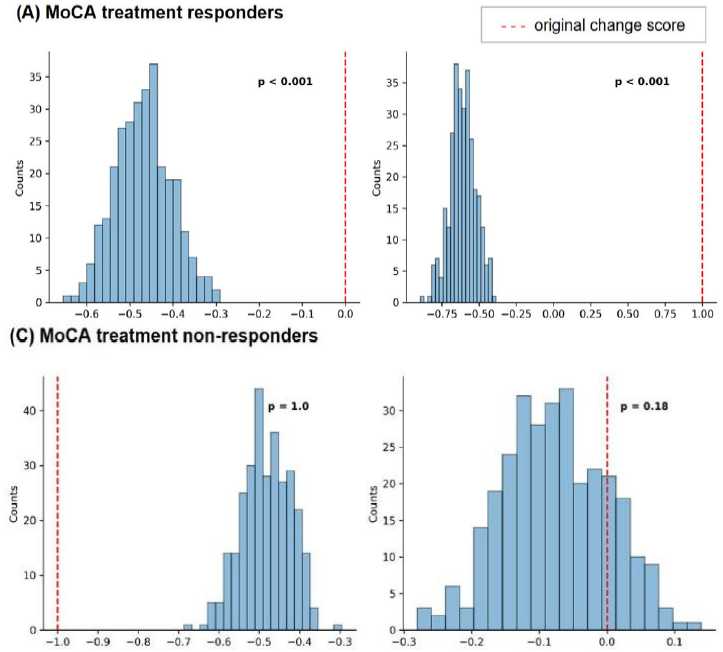
1 **Mine Historical Data**

2 **Simulate Controls; Model validation**

3 **Person-specific control**
Check group allocation; E-values







~58-67% are responders

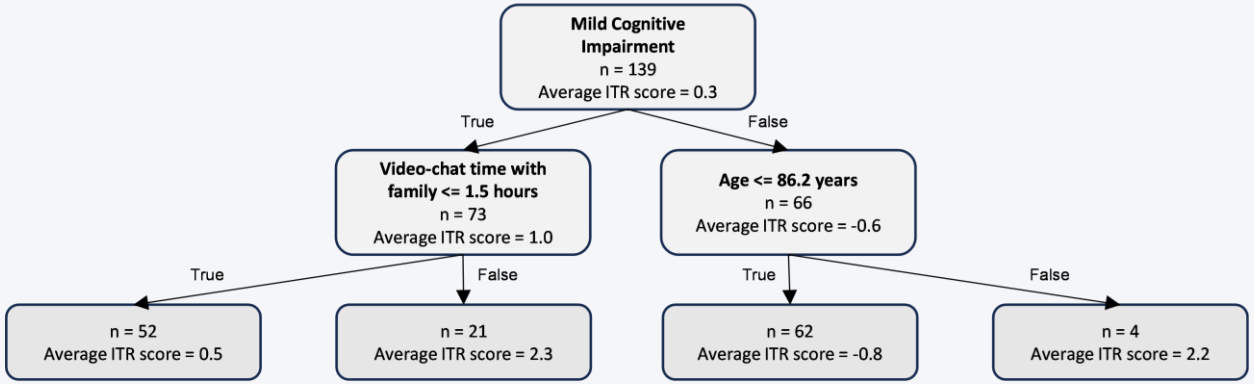
Comparison of trial findings using two control groups (matched; in silico)

Each patient is compared to their own constructed comparator: a matched real-patient group (matched control) or a simulated group (in silico control).

Identify treatment responders and their characteristics, cluster-based analytics needed (e.g., Bayesian hierarchical)

Compare the two methods:

The matched and in silico controls identified largely but not uniformly the same treatment responders: agreement of 0.92 (MoCA) and 0.87 (animal naming)



Reflections

01 **A "perfect" model can still produce an unusable synthetic control group**

- Model performance and trial assumptions are two separate hurdles, and you need to clear both
- This is the difference between "can we build twins?" and "are these twins actually usable?"

02 **Dropouts: Can we twin the hardest-to-follow patients?**

- ITT requires we account for everyone enrolled, including those who left early
- But dropouts are systematically different - sicker, less engaged, more burdened

03 **Trial endpoint modeling (e.g., 2 month) vs historical data collection interval (e.g., 12 month)**

04 **Would you design your trial differently knowing you'd use synthetic/ digital twins controls later?**

THANK YOU!