A digital twin in healthcare is not validated in the mirror

A clinical digital twin's deployability follows neither from its realism nor from its generator's performance alone. It rests on an architecture demonstrating that, for a declared task and domain, the decisions produced remain substitutable for the real, and on the system's capacity to refuse when it leaves that domain.

Digital twins enter regulated healthcare through the door of scarcity. Clinical data are costly to produce, hard to share, and sometimes nonexistent for certain populations or rare events. Synthetic cohorts promise to widen that space: training models, exploring therapeutic scenarios, testing protocols, building virtual control arms, or simulating decisions that cannot be experimented directly on humans.

The promise is credible. The criterion by which an executive committee will defend it before a regulator is far less so, because it still too often bears on the wrong property.

The methodological community has, for the most part, already settled the matter: the statistical fidelity of a synthetic cohort does not measure its utility for a task. The displacement persists elsewhere, at the layer that truly matters for an industrial dossier, that of deployment, governance, and regulatory accountability, where a twin is still judged by its apparent realism. That is where the trap sits, and that is where this note is situated.

The realism trap

A twin's validation still frequently rests, in deployment trade-offs, on measures of resemblance: distribution comparison, expert inspection, a discriminator's ability to tell real from synthetic data. These measures are locally useful and become misleading the moment they are promoted into a general validation criterion.

For the relevant question is never: "Does the twin resemble the real?"

It is always: "For what use does one wish to replace the real, and under what guarantees?"

A twin is never deployed "in general". It is always mobilized for a precise purpose: training a model, exploring a therapeutic scenario, estimating a recruitment strategy, building a control arm, calibrating a threshold, or testing a clinical policy. Each use carries its own notion of fidelity.

Jérôme Vétillard – VP R&D + Engineering / Chief Product Officer TweenMe® by Qualees®

In some cases, faithfully preserving a statistical distribution is precisely the objective. In others, what matters is the preservation of a decision boundary, a calibration, or a predictive capacity. The debate therefore does not oppose resemblance to its absence. It opposes two levels of resemblance:

1. Global distributional resemblance, often measured,
2. And the resemblance of the structure relevant to the task, rarely isolated.

Realism is not a bad indicator. It is a specialized one, relevant only when the level it captures matches the task being claimed.

From statistical portrait to decision instrument

All twins begin as statistical portraits. Some then become decision instruments. The difference is structuring.

A portrait seeks to represent a population. An instrument seeks to produce a decision faithful enough to replace, in a given context, a decision that would have been made from real data.

The moment a twin enters a clinical, regulatory, or industrial decision, the question changes in nature. Validation no longer bears principally on the quality of the imitation. It bears on the robustness of the substitution. What is evaluated is no longer the portrait, but the decision loop in which it intervenes.

Three proofs then become necessary

1. The first is to demonstrate the absence of leakage. The real cohort used for evaluation must never have contributed to generating the synthetic data. Without strict separation between generation and validation, the measured performance may be no more than the reflection of data contamination.
2. The second is to demonstrate operational substitutability. The Train-on-Synthetic / Test-on-Real methodology, formalized by Esteban, Hyland and Rätsch (2017) for medical time series, provides a relevant frame here: the model is trained exclusively on the synthetic data, then evaluated on an independent real cohort. But discrimination alone does not suffice. A credible substitution must preserve the properties that condition the use: discrimination when that is what is sought, but also calibration, decision benefit, and stability when those dimensions govern the clinical decision.
3. The third is to declare the applicability domain explicitly. This domain does not merely describe a population. It defines the twin's space of validity: patient characteristics, collection modalities, clinical context, time period, therapeutic

practices, and technical conditions under which the preceding guarantees remain demonstrated. This declaration is not a guarantee: it is a refutable hypothesis, one that surveillance must be able to disprove and that must be revised accordingly.

Outside that space, a twin must not produce a confident answer. It must produce a refusal. This refusal is not a spontaneous property of the generator; it is an architectural requirement, presupposing an explicit mechanism for detecting out-of-domain situations. Governance begins precisely where the system recognizes that it is leaving its space of validity (the promotion port and the refusal taxonomy, developed elsewhere in this series, provide its framework).

The scarcity objection

An objection arises here, and it is a serious one. The proof of substitutability requires an independent real cohort. Yet the twin is summoned precisely where the real is missing: underrepresented populations, rare events, orphan diseases, situations never observed.

For these cases, Train-on-Synthetic / Test-on-Real validation is unavailable by construction. One does not test on a real that does not exist.

This objection must be held rather than circumvented. It compels a distinction between two regimes.

- In the *anchored regime*, a real validation cohort exists. Substitutability is demonstrated there, in the strong sense, and it is to this regime that the thesis of this note fully applies.
- In the *extrapolated regime*, no ground truth is available in the zone of use. Substitutability is not demonstrated there; it is bounded and monitored. Its value then rests on a generalization hypothesis about the generator, one that only the declared applicability domain frames, and that out-of-domain detection must be able to invalidate in real time.

The consequence is uncomfortable and must be assumed: a governable twin will sometimes refuse to answer precisely where it was hoped for most. This is not an architectural failure. It is the honest form of governability, set against the silent confidence of a system that answers everywhere without ever knowing where it ceases to be valid.

The generator does not suffice

This distinction leads to another frequent confusion. A twin's deployability cannot be deduced from its generator's performance alone.

An excellent generator can produce data exhibiting confidentiality leaks, a collapse of rare modes, poor preservation of the relevant dependencies, or an inability to extrapolate beyond the observed domain. Conversely, a generator whose data remain easily distinguishable from the real may nonetheless allow excellent substitutability for a given operational task.

The generator's properties remain important. They simply do not suffice to establish that the decisions built from the twin will remain valid.

What ToxTwin actually shows

ToxTwin illustrates this distinction, and it is worth framing what it establishes exactly.

The result: the synthetic data generated to model exposure-response relationships remain readily separable from the real trials by a discriminator.

The protocol: a toxicological classifier is trained on these synthetic data alone, then evaluated under Train-on-Synthetic / Test-on-Real on compounds never observed during training, within the declared domain of validity.

What this proves: within that domain, discrimination performance remains compatible with that obtained from the real data.

What this does not prove: general statistical indistinguishability, nor, as this demonstration stands, the preservation of calibration and benefit.

Global distributional resemblance fails. Substitution, for the task considered, succeeds.

The example obviously does not demonstrate that all twins become substitutable. It shows, more modestly, that indistinguishability is neither a necessary condition nor a sufficient proof of deployability.

PREDICARE sheds light on the symmetric problem. A twin meant for clinical triage is not only obligated to produce a decision. It must also identify the situations for which that decision is no longer guaranteed. Here again, governance begins when the system recognizes its own exit from the domain.

What the executive committee actually carries before the regulator

The strategic error would be to expect a generation of ever more realistic models to resolve a question that belongs to architecture. The sought property is not maximal realism. It is the governability of substitution.

This governability does not remain a mere formula if it is anchored to the regulatory instrument that already encodes it. A dossier is not filed as a resemblance score. It is filed as a chain of guarantees: an explicitly defined task, a strict separation between generation and validation, a demonstration of substitutability on independent real data, a declared applicability domain, mechanisms for detecting out-of-domain situations, and continuous surveillance verifying that these guarantees remain valid as clinical practices, populations, or treatments evolve.

This is precisely the logic of the predetermined change plans that regulators are beginning to recognize for learning-based devices (the FDA's Predetermined Change Control Plan is its most accomplished expression): authorizing in advance a bounded envelope of changes, under surveillance, rather than freezing a model. The doctrine of the promotion port offers its architectural primitive.

Before a regulator, accountability therefore bears neither on the aesthetic quality of the synthetic data nor on the generator's isolated performance. It bears on the capacity of the entire decision chain to produce, within an explicitly declared domain, decisions whose substitution for the real remains demonstrated and continuously monitored.

A digital twin is therefore not an autonomous object. It is a component of a decision architecture. And like any critical architecture, it is not validated in the mirror. It is validated by the guarantees it brings to the decisions it replaces.

A twin that resembles is shown. A twin that replaces is governed.

[Series: Digital Twin in Healthcare - 9/12 - Sunday closing article of the weekly series]