A prediction error degrades a metric, an intervention error degrades a patient

Health digital twins, and the proportionality between claim, proof, and responsibility

1. The right question

A health digital twin that displays a good performance curve has demonstrated one thing, and one thing only: that it predicts. It has not demonstrated that it explains, nor that it simulates, nor that it can serve to intervene on a patient. Predicting, explaining, simulating, intervening are four distinct claims, and each commits a proof and a responsibility of its own. The thesis of this article holds in one sentence: a model's claims must be proportionate to the proof it provides and to the responsibility it commits, and a proof of prediction covers neither when the claim is to intervene.

This is not a critique of AI, and that must be said before the article gets read crooked. A model can be extraordinarily useful without explaining, without simulating, and without intervening. A classifier that sorts retinal images or prioritizes a queue renders a real service while claiming nothing about the mechanism of the disease. The problem is never that a model is *limited*. The problem is that it claims more than it proves, and that in health this inflation is paid on the patient, not on a metric.

The scope must be set from the outset, for *digital twin* covers three families. The purely statistical twin learns regularities with no model of the mechanism. The mechanistic twin rests on explicit physiological equations, as in physiologically based pharmacokinetics or in cardiovascular modeling. The hybrid twin combines the two. The target of this article is the first family, and the statistical part of the third. A twin that carries its causal assumptions in its equations has already done part of the work that will be required; a twin that holds only by statistical fitting has not.

This article builds a deliverable: an identity card for the twin, a direct consequence of the audit of its dependencies, without which the system is not governable whatever its performance. The hurried reader may skip to section 9; there they will lose the reasons, not the conclusion.

2. Four claims: a hierarchy of responsibility, not of knowledge

To *predict* is to associate an output with an input. To *explain* is to know why that association holds. To *simulate* is to produce reliable counterfactuals, to answer "what would happen if the treatment were changed." To *intervene* is to act on a real patient.

The temptation is to see here a staircase where each step would presuppose the previous one. That would be wrong, and a clinician would refute it in one sentence.

Medicine has always intervened without explaining, by empirical approach. Aspirin was prescribed for decades before its mechanism was understood, lithium given without knowing why it stabilizes, general anesthesia practiced though its mechanism remains debated, historical anticancer agents used whose efficacy long preceded the explanation. One can intervene without explaining, explain without being able to intervene, predict without understanding. The relation between these verbs is not a hierarchy of knowledge.

It is a hierarchy of responsibility. In moving from predicting to intervening, what changes is not the degree of understanding required, it is what bears the error. A prediction error degrades a metric. An explanation error compromises a model. A simulation error invalidates hypotheses. An intervention error degrades a patient. The fundamental jump is not cognitive; it is moral, clinical, and legal. It is what orders the four verbs, and what justifies that the required proof grows heavier at each notch.

Claim	**Question**	**Minimal proof**	**An error is borne by**
Predict	What will happen?	Performance: discrimination and calibration	a metric
Explain	Why?	Declared dependencies, explicit assumptions	a model
Simulate	What would happen if?	Causal structure, domain of validity	hypotheses
Intervene	What should I do?	Assurance proportionate to risk	a patient

This reading by responsibility illuminates the apparent exception. One can reach "intervene" without passing through "explain": this is exactly what the randomized controlled trial does, establishing the effect of a treatment without its mechanism, replacing understanding with experimental control. But it then provides proof at the

intervention level, experimental, commensurate with the responsibility committed. The fault is never to skip the explanation. The fault is to intervene with a proof of prediction, which carries neither the mechanism nor the experiment, and therefore not the responsibility.

The industrial conflation consists in presenting a system that has filled the first row as if it had filled the following ones. The slippage lodges itself in a verb: the twin "models" the patient, "anticipates" the response, "tests" scenarios. Each one promises a simulation that the provided proof does not establish. The governable question is not "is the model performant?" but "what does this performance authorize, and who answers for it?"

3. The bridge to decision: calibration

One thing is surprising in most twin evaluations: the outsized place given to discrimination, and the meager portion left to calibration. It should be the reverse, for to move from prediction to decision, calibration matters more than discrimination.

Discrimination measures a sorting capacity: does the model place the patients most at risk above those least at risk? This is what the AUC or the c-index summarize. Calibration measures something else: do the announced probabilities correspond to the observed frequencies? When a model predicts a twenty percent risk, does the event indeed occur in one patient out of five? A model can sort perfectly and be systematically wrong about the levels. A high AUC does not mean a reliable decision.

Yet a clinical decision does not rest on a rank, it rests on a probability threshold: treat above such a risk, monitor below. A threshold makes sense only on calibrated probabilities. A model that discriminates well but calibrates poorly will send the wrong patients to the other side of the threshold, with a deceptive assurance, and this is often what distribution shift degrades first. The bridge between prediction and intervention runs through calibration; discrimination alone does not cross it.

4. Signal and scarcity: aggravating factors, not the root

One might believe the root of the problem is the poverty of the clinical signal. It is an aggravating factor, to be situated without dwelling on it. Three facts from information theory bound what a model can do: no transformation of a datum increases the information it holds about a target; a model does not create information, it extracts it [Cover & Thomas]; and at a given noise level, there exists a bound beyond which no processing recovers further signal [Neyman-Pearson]. Creating information is impossible, extracting a latent structure is possible, estimating that structure reliably is conditional on the signal and the number of patients.

When the signal is weak, estimation becomes fragile, and the high-dimension, low-sample regime, common in health, makes things worse: with far more variables than patients, illusory structures appear stable [Hall, Marron & Neeman 2005; clinical synthesis to be added]. But this regime does not ground the hierarchy of verbs; it makes it more urgent. The proof: one can remove it without the thesis giving way. "My twin rests on fifteen million patients, and my imaging model has a very strong signal." Granted. In electrocardiography, radiology, anatomic pathology, deep learning reaches high performance on rich signals and massive cohorts. And the problem remains identical: excellent discrimination on images demonstrates neither the mechanism, nor the effect of an intervention, nor the validity of a counterfactual. Fifteen million patients establish a very reliable association; they establish neither the explanation, nor the simulation, nor the right to intervene. A strong signal moves the system higher on the first row of the table; it does not make it change rows.

Reality recalls this wall when a clinical design takes it seriously: an adaptive platform trial such as OCTOPUS/PLATYPUS, in progressive multiple sclerosis, exists only because the effect to be measured is small and the cohort costly [ms-octopus.mrcctu.ucl.ac.uk]. A cautious design is a metrological confession. One will add only that, on the data side, many measurements per patient do not make many independent patients: the quantity that counts is the effective sample size, closer to the number of patients than to the number of rows.

5. Declaring the dependencies

If performance is the entry of a dossier, the first item answers a question: on what does this performance depend? One answers by ablation: one removes one by one the sources likely to carry it, and measures what it loses.

Four sources stand out, for "prior" covers different things: the injected equations (mechanistic prior, as in physics-informed networks, already used in the small-data regime [proceedings.neurips.cc, 2024]); the regularities of a massive pretraining (statistical prior, as in MedGemma [developers.google.com, 2025]); the shape of the learned latent space (geometric prior); the nomenclatures (ontological prior).

The gap between raw performance and performance after removal of a source measures a real quantity, which must be named correctly: a dependency, not a causal provenance. The sources interact. Three priors can together produce a performance that collapses as soon as a single one is removed; then the sum of contributions exceeds the total, attribution depends on the order of ablations, and several decompositions coexist without any being the right one. This is an allocation problem akin to Shapley values, with their own assumptions and their computational cost [Shapley 1953; Lundberg & Lee 2017]. The ablation matrix measures what the model depends on today, not what causally produces its performance. A reader trained in statistics will object, and will be right.

This honesty costs a test one would have wished decisive. One might believe it suffices, to refute the idea of imported performance, to show a gain obtained without addition of information or dimension reduction. The test is weak: a better architecture produces a gain without adding information, by narrowing the range of possible hypotheses, which is already an imposed structure, and passes through the test. It corroborates a dependency; it does not prove the absence of an imported source.

The instrument keeps its value. A strong dependency on a pretraining is not a flaw, it is a declaration to be produced. What remains a fault is the absence of the declaration, or an out-of-distribution collapse that no one looked at. This declaration, put into form, is the card of section 9: the card is not an idea separate from the audit, it is its output.

6. Provenance is not transferability

There remains a stubborn association to undo: the idea that a heavily imported performance would, for that reason, be fragile. This is false, and the most visible example in the field shows it. AlphaFold depends on a gigantic exogenous prior, and its robustness exceeds that of many models trained on the data of a narrow task alone. Importing a regularity is not a weakness; in health, it is often what saves, when the patient's cohort is too small to provide it itself.

Hence a distinction that cuts: *provenance explains, transferability decides*. Knowing where performance comes from helps understand why a model holds or gives way. Knowing whether it holds outside its training distribution and by subpopulation is what authorizes, or not, deployment. The two dimensions are decoupled: a performance can be largely imported and highly transferable, or largely "homegrown" and collapse at the first change of site. The decisive criterion is not provenance, it is stability under distribution shift. The dependency audit does not serve to purge the exogenous in the name of some endogenous purity; it serves to declare it and to test it where it counts.

7. Simulate what, within what domain of validity

A twin rarely claims to merely predict; it claims to simulate. And simulation demands something other than good performance: a sufficiently correct causal structure. The statement must be held without softening: even at a strong signal, a twin can be wrong if it simulates on an incorrect causal structure.

It remains to say what "sufficiently correct" means. A structural causal model describes the variables and their relations, and allows one to distinguish what it can support: an association, the effect of an intervention, or a counterfactual, the three rungs of Pearl's ladder [Pearl, Causality]. Two properties condition the scope of a simulation: transportability says whether a result established in one population holds in another, and under what conditions [Bareinboim & Pearl]; invariance says whether a relation holds

across environments, the most reliable indication that it is causal and not accidental [Peters, Bühlmann]. A sufficiently correct causal structure is one that has been shown to transport and to remain invariant over the domain where one wishes to simulate. Outside this domain, the simulation is not wrong by misfortune; it is out of warranty.

Hence three levels, not two states: predictive correlation, with no causal claim; local interventional causality, which predicts the effect of certain interventions in a narrow domain; generalizable structural causality, which captures a transportable mechanism. Most serious therapeutic systems live at the second level, over a bounded domain: they do not imitate a patient, they illuminate certain precise interventions. The fault consists in selling this second level as the third. The question is therefore not "does it simulate well?" but "what does it simulate, at what scale, and within what domain of validity?" These questions should be asked from the model's design onward. Every simulation indeed implies a choice of scope, of resolution, and of level of abstraction (often dictated by the structure, in quantity and quality, of the available data). Yet a frequent analytical bias leads to pursuing the ideal of a universal digital twin able to reproduce the patient in full. But a model's value does not depend on its exhaustiveness, but on the fit between what it claims to represent and the decision it must illuminate.

An example at the frontier of the state of the art fixes the idea. The protein language models of the ESM family reach remarkable structural accuracy, and the literature acknowledges that it is not yet settled whether they internalize physical principles of folding or whether they mainly capture evolutionary correlations. If this question remains open for one of the most advanced bio-digital systems currently available, the industrialist who claims that their twin "simulates the patient" should at minimum specify what it simulates, at what scale, and within what limits this simulation can be considered valid.

8. Responsibility commands the proof

Everything converges toward a simple principle. It is not the lack of causal information that forbids, on its own, intervening on the faith of a prediction. It is that a prediction error degrades a metric, where an intervention error degrades a patient. The required proof must therefore be set according to the responsibility committed, and not according to the displayed performance alone. It is a principle of governance before being an epistemological one.

Critical engineering has translated it into a scale. Aeronautical software grades its requirements according to the severity of the failure: the DO-178C standard defines five assurance levels, and the number of verification objectives decreases with criticality, from seventy-one at the highest level to twenty-six at a minor level [RTCA DO-178C, 2011]. The proof is proportionate to the impact, therefore to the responsibility. Health has a framework for risk classification, but no scale for the legitimacy of simulation. The

European AI regulation classifies as high-risk any AI integrated into a CE-marked medical device, that is, classes IIa, IIb, and III of medical devices and A to D of in vitro diagnostics, with a deadline set at August 2027 for the highest classes [Regulation (EU) 2024/1689; (EU) 2017/745; (EU) 2017/746]. It requires robustness, absence of bias, human oversight. It says nothing of the causal level required before authorizing an intervention.

This is the missing link, and one can propose it without excess: an assurance level that grows with the verb, because responsibility grows with it. Predicting in order to triage demands discrimination, calibration, and transferability. Explaining adds the dependency audit. Simulating further demands an established causal level and a declared domain of validity. Intervening demands a causal structure shown to transpose, and explicit conditions of non-deployment. The cost of proof rises with the verb because the cost of the error, and the name of who bears it, rise with it.

9. The twin's identity card

Here is the deliverable, and it is nothing other than the preceding audit put into form. Rather than a checklist half of which is forgotten, four blocks that one remembers: the *object*, the predictive target and the decisional target, which are not always the same; the *data*, the signal available for that target, the effective sample size, and the dependencies declared by the audit; the *validity*, out-of-distribution transferability, the causal level attained, the domain beyond which the twin must not simulate; the *assurance*, the targeted level, the responsibility assumed, and the conditions of non-deployment.

Object, data, validity, assurance: what is claimed, with what, how far, and under what warranty.

A card has value only if it does not lie about its own precision. This is the classic flaw of governance metrologies: displaying values more solid than they are. A mature card therefore fills each entry with a value, an uncertainty interval, and a confidence level. Governing a fragile estimate as a fact is the symmetric fault of the conflation of verbs: in one case predicting and intervening are confused, in the other estimating and knowing.

This card is not a standard in force; it is a proposal that makes operational the spirit of existing regulations for the case of twins. Qualifying a twin as a "candidate for such an assurance level" is not to declare it compliant, and remains a doctrinal framework, not a legal opinion.

10. Limits

The dependency audit does not deliver the causal provenance of performance: in the presence of interactions, fine attribution is undecidable without strong assumptions. The proposed refutation test is partial: a better architecture produces a gain without added

information, and it corroborates without refuting. The instruments have a cost that this article does not quantify: the scale says that proof must be proportionate to risk, it does not say where exactly to place the thresholds. The framework assumes that one can name the decisional target, which often remains implicit. The fine allocation of responsibility overflows epistemology and belongs to a distinct legal work, out of scope here: this article shows that each verb calls for a different regime, it does not adjudicate it. Finally, the critique targets the predominantly statistical and probabilistic twin; a mechanistic twin that carries its causal assumptions in its equations deserves a distinct evaluation, on the correctness of those equations rather than on the declaration of its dependencies.

11. Conclusion

The dominant debate asks: is the model performant? It is the first of a series of questions that we have learned not to distinguish.

Performing authorizes prediction. It does not authorize explaining, simulating, or intervening on the sole faith of the same proof. What this article proposes is not a critique of AI but a rule of proportionality: to each claim, the causal level, the required proof, and the responsibility that correspond to it. The rule will seem obvious once read. It is so little so in practice that the market makes a living from ignoring it.

This rule is not peculiar to twins, nor even to health. The same proportionality governs any algorithmic claim that crosses the frontier from prediction toward action: bank scoring, the autonomous vehicle, the software agent that acts on its own, weapons systems. The health twin is only one case of it, but it is the acute case, the one where the error, at the end of the chain, is borne not by a metric but by a body. Health does not invent the principle; it is the place where ignoring it hurts most.

Hence the sentence that sums up all the rest, and the only one to retain if one retains but one: a prediction error degrades a metric, an intervention error degrades a patient. A digital twin that claims to intervene with a proof of prediction is not a virtual patient. It is a decision disguised as a measurement, and a responsibility that no one has signed.

References

Information theory and detection

- Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley. [data processing inequality]
- Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A*, 231, 289-337.
- Van Trees, H. L. (2001). *Detection, Estimation, and Modulation Theory, Part I*. Wiley. [matched filter]

High dimension, low sample size

- Hall, P., Marron, J. S., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B*, 67(3), 427-444. [methodological reference; a clinical synthesis remains to be added]

Contribution attribution

- Shapley, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games II* (pp. 307-317). Princeton University Press.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions (SHAP). *Advances in Neural Information Processing Systems (NeurIPS)* 30.

Twins, priors, and models

- Med-Real2Sim: Non-Invasive Medical Digital Twins using Physics-Informed Self-Supervised Learning. NeurIPS 2024. https://proceedings.neurips.cc/paper_files/paper/2024/file/0b081a44ed0b8c0c4aa6bd886a60bea4-Paper-Conference.pdf
- Physics-informed neural networks for modeling physiological time series for cuffless blood pressure estimation. *npj Digital Medicine* (2023). https://www.nature.com/articles/s41746-023-00853-4
- MedGemma model card. Google Health AI Developer Foundations (2025). https://developers.google.com/health-ai-developer-foundations/medgemma/model-card ; MedGemma Technical Report, arXiv:2507.05201 (2025).

- Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583-589. https://www.nature.com/articles/s41586-021-03819-2
- Lin, Z., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model (ESM-2 / ESMFold). *Science*, 379(6637), 1123-1130. https://www.science.org/doi/10.1126/science.ade2574
- Protein Language Models Encode Evolutionary Grammar but Conflate Topological and Thermodynamic Phases. *bioRxiv* (2026). https://www.biorxiv.org/content/10.64898/2026.04.07.717117v1.full [open question: physical principles vs evolutionary correlations]

Causality

- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.
- Pearl, J., & Mackenzie, D. (2018). *The Book of Why*. Basic Books.
- Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *PNAS*, 113(27), 7345-7352. https://www.pnas.org/doi/10.1073/pnas.1510507113 [transportability]
- Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5), 947-1012. https://rss.onlinelibrary.wiley.com/doi/10.1111/rssb.12167 [invariance]

Field and regulatory framework

- OCTOPUS, Optimal Clinical Trials Platform for Progressive Multiple Sclerosis. MRC Clinical Trials Unit at UCL. Protocol. https://ms-octopus.mrcctu.ucl.ac.uk
- RTCA DO-178C (2011). *Software Considerations in Airborne Systems and Equipment Certification*. [assurance levels / DAL]
- Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act).
- Regulation (EU) 2017/745 on medical devices (MDR).
- Regulation (EU) 2017/746 on in vitro diagnostic medical devices (IVDR).