Evidence Is a Conditional Promise

Toward a relational and emergent theory of the validation of computational evidence in health, and beyond

I. Three contradictions, and why they are impossibilities

Let us begin by giving up a habit. The habit consists in asking whether a piece of evidence is valid, the way one asks whether a bridge is sound or a theorem correct. The question seems healthy. It is ill-posed, and three ordinary situations suffice to show this, provided one follows through to the end of what they show. That end, as we will see, leads to conceiving evidence as a promise, not as an object.

1. **First situation.** Two clinical studies, conducted with impeccable rigor, bearing on the same intervention, arrive at two incompatible recommendations. Neither is tainted by error: protocols published, analyses pre-registered, populations described. Yet they cannot both serve as the basis of the same decision. If validity were an intrinsic property of each study, two equally valid objects ought not to contradict one another.
2. **Second situation.** An excellent randomized controlled trial, the gold standard, becomes strictly unusable outside its target population. The same trial, identical to the word, is evidence in one context and noise in another. Its validity does not travel with it.
3. **Third situation.** A perfectly calibrated predictive model (its predicted probabilities coincide with observed frequencies) leads to a poor clinical decision, because the decision depends on a trade-off between errors that calibration ignores. A well-calibrated test that fixes its threshold without accounting for the asymmetric cost of a false negative in oncology is calibrated and dangerous.

So far, these situations are paradoxes. A paradox can be repaired, and the community knows the repairs; one must therefore show that none of them holds, failing which the reader will conclude, legitimately, that these examples reveal the inadequacy of certain notions of validity, not the necessity of changing them.

Let us take up the repairs one by one. Restrict ourselves to internal validity? The first contradiction persists: two internally valid studies still diverge, because the incompatibility arises not from an internal defect but from the use made of them. Add external validity? External with respect to what: a population, a decision, a context. One defines external validity only by naming that to which the evidence is meant to apply; one has therefore already conceded that validity is relative to a use. Stratify, meta-analyze?

One moves the decision up a notch (which population, which subgroup, which model of effects) without removing relativity, only displacing it upward. Fall back on a pure statistical property, calibration? The third contradiction forbids it: calibrated is not valid.

The finding is then of another nature. It is not that this or that usual notion of validity is insufficient; it is that no non-relational notion of validity can dissolve the three contradictions, because every repair, in order to work, smuggles back in a decision, a loss, or a domain. Intrinsic validity is not a good idea poorly executed. It is an impossibility. And from this impossibility follows a more modest and more demanding idea: validity is not in the evidence, it is in the relation between a piece of evidence and the use made of it. The rest of this text draws the consequences of this shift, up to a point that may come as a surprise: a piece of computational evidence is neither an object to be certified nor even a static relation; it is a conditional promise, of which the use contract is merely the social formalization.

I do not assert this thesis. I let it emerge, because a doctrine that opens on its axiom asks to be believed, whereas a doctrine that opens on an impossibility compels one to follow it.

II. The minimal structure, and the reversal of the burden of proof

If validity is relational, then "valid" is an incomplete predicate, like "to the left of" or "greater than." It calls for arguments. The question is no longer *is this evidence valid?* but valid for what, under which criterion, within which limits, until when? It remains to know how many arguments are necessary, and the answer is not a matter of taste, for each argument is justified by the fact that its absence resurrects one of the three impossibilities.

There must be a decision. Without a decision to serve, "valid" has no content: this is the first impossibility. Let us denote it **D**. In the strict sense of statistical decision theory, founded by Abraham Wald (*Statistical Decision Functions*, 1950), a decision is not a belief but a rule: a function δ that, to an observation, assigns an action. This point cannot be stressed enough. A decision is not a probability distribution; it is a mapping that turns information into an act. One does not validate a distribution. One validates the quality of a rule of action, and, as we will see, the value of what it teaches us.

I will henceforth call **evidence source** any device producing the data that feed δ (a real cohort, a synthetic population, a digital twin, a simulation), without prejudging its status: it is the source, and not the raw data, that one will seek to render substitutable.

There must be a criterion to order the consequences of the rule: a loss function, denoted **L**. This is the third impossibility: the calibrated model fails because it ignores that erring

in one direction does not cost what erring in the other costs. Wald calls *risk* the expectation of this loss and makes it the sole judge of a rule. I add at once two corrections that medicine imposes and that classical theory evades, which I develop later: this loss is not a scalar, and it is not only the cost of an action; it includes the value of what a piece of evidence lets us know before acting. The loss therefore carries, from its first appearance, two functions: to decide, and to reduce uncertainty.

There must be a domain in which the evidence is deemed reliable, denoted **Δ**. This is the second impossibility: the randomized trial does not transfer because its validity was local and was believed universal. The notion is codified: the third of the OECD's five principles for the validation of structure-activity models (*Guidance Document on the Validation of (Q)SAR Models*, 2007) requires that a model declare its applicability domain. Outside Δ, a piece of evidence is not false; it is off-topic, which is more dangerous, because off-topic evidence looks like evidence.

There must, finally, be a time, denoted **T**: evidence validated today is not validated forever. I devote Section VI to it.

We obtain a structure (D, L, Δ, T) that I claim to be minimal. Minimal does not mean elegant; it means that nothing can be removed without breakage: remove D and the predicate empties, remove L and the first impossibility returns, remove Δ and it is the second, remove T and it is the expiration of evidence. The structure is not chosen; it is what remains once one has removed everything whose absence is paid for with a contradiction.

As for sufficiency, I do not proceed by confession: "I know of no fifth component" would be a weakness, not an argument. I lay down a rule, and I reverse the burden of proof: any candidate component must demonstrate that it is reducible neither to D, nor to L, nor to Δ, nor to T. Absent such a demonstration, it is already contained in one of the four. It is not for the theory to prove that no fifth component exists; it is for whoever proposes one to prove its irreducibility. Minimality thus ceases to be a timid conjecture and becomes an opposable constraint.

III. The theory judges itself

The most serious objection arises here. If validity is never anything but a relation to a use, have we not dissolved every notion of quality? Is not everything of equal worth, once each party invokes its own triplet? Relativism lurks, and with it the ruin of the argument: a doctrine that makes everything relative renders itself indefensible.

The parry is not to slip an absolute validity back in. It is to require of a theory of validation that it satisfy criteria that are themselves explicit. A theory of validation is admissible if and only if it is coherent, composable, transferable, and refutable.

- **Coherent:** it does not contradict itself,
- **Composable:** partial validations assemble along a chain of decisions without the guarantee being lost along the way,
- **Transferable:** a validation carries from one context to another under declared conditions, and not by hope,
- **Refutable:** each of its statements can be contradicted by an observation.

Relativity ceases to be arbitrary: it is governed by a conjunction of four conditions. One does not say "everything is of equal worth"; one says "the value of a piece of evidence is measured relative to a use, and the measure itself obeys rules."

A theory that imposes these criteria on others must submit to them, on pain of hypocrisy. Let us therefore pass it through its own sieve, criterion by criterion.

- **Coherence:** the thesis that all evidence is relative to a decision is itself relative to a decision (that of validating evidence), which places it in agreement with itself rather than in infinite regress; it does not self-refute,
- **Composability:** Section X demonstrates it on a case, where a cohort validated for a sub-decision feeds a larger decision without a break in the guarantee, because substitutability, being defined relative to a decision, composes as decisions compose,
- **Transferability:** Section XI carries the theory beyond health by declaring its transfer conditions, which is exactly what the criterion demands,
- **Refutability:** Section II gave the conditions for it, and the rule reversing the burden provides one more: let an irreducible component be exhibited, and the structure will have to yield.

A theory that survives its own criteria is not thereby true; it is admissible. That is all a theory can claim, and it is already more than the idea of "validity in itself" has ever offered.

One last misunderstanding remains to defuse. This theory does not claim to replace the established frameworks of evidence: neither GRADE for grading recommendations, nor CONSORT for reporting trials, nor TRIPOD for prognostic models. It claims to describe the level at which these frameworks become comparable: each, in its own way, declares a decision, a criterion, a domain, and a window of validity. The relational theory is not one more competitor on the list; it is the common grammar that explains why these frameworks do what they do, and what they share. A theory that knows itself to be a metatheory does not threaten the theories; it situates them.

IV. Loss is contextual before it is vectorial

I said that loss is not a scalar. This must be demonstrated, for the whole of classical decision theory rests on the idea that one can summarize the consequences of an action by a single number and minimize its expectation.

Now a clinical decision weighs, simultaneously, magnitudes irreducible to one another: efficacy, safety, equity across subpopulations, cost, acceptability to the patient, regulatory compliance. These are not shades of a single quantity; they are distinct, sometimes antagonistic dimensions. Loss is a vector **L**, and the optimal decision is not a minimum but a front, what multi-objective optimization calls a Pareto front, the set of trade-offs none of which dominates the others on all dimensions at once.

But one must go further, for stopping at the vector would suggest that the dimensions and their weights are fixed. They are not. The preferences that weigh efficacy against toxicity depend on the patient, the stage, the age, the care plan; what is a good trade-off for one is unacceptable for another. Loss is contextual before it is vectorial: it is not a vector given once and for all, but a vector whose components and weightings are themselves a function of the decision context. This is one more requirement, not a gratuitous complication: it forbids transporting a trade-off from one context to another as though it were neutral.

The consequence is more political than technical. Reducing this contextual vector to a scalar (fixing an exchange rate between a year of life and a euro, between the safety of a group and average efficacy) is not an operation of calculation; it is an encoded value judgment. It may be legitimate; it cannot be clandestine. A theory of evidence that lets "scalarization" happen in silence, in the weightings of a composite score or in the choice of a threshold, lets governance happen without its knowledge. To validate a piece of evidence without making explicit the loss vector it serves, and the context that fixes its weights, is to validate one knows not what for one knows not whom.

V. The domain: internal times external

The applicability domain I have borrowed from the validation of structure-activity models, where it is well equipped: there one knows how to construct a model's region of reliability through the covariate envelope, the convex hull of the support, Mahalanobis distances or leverage values, zones of high density (see the comparison by Sahigara and colleagues, *Molecules*, 2012). These methods are sound. They have one defect: they measure only one face of the domain. A clinical applicability domain has two, and each splits in two.

There is an **internal face**, proper to the modeled relation. It comprises the statistical domain (the region of covariate space where the observed distribution supports inference) and the clinical domain (the set of situations where the relation has a

pathophysiological meaning, which may be narrower than the statistical support, since a learned correlation does not always carry beyond a mechanism).

There is an **external face**, proper to the context that surrounds the decision. It comprises the temporal domain (the window during which the conditions of generation remain comparable to those of use) and the organizational domain (the framework of practices, recommendations, and constraints within which the decision is taken). One can therefore write the domain as a product: Δ = internal × external, the internal uniting the statistical and the clinical, the external uniting the temporal and the organizational. Reading gains by it: the two internal faces describe what the evidence is, the two external faces describe the world in which one wishes to use it.

An example settles the question. Take a cohort perfectly within its internal domain (same covariates, same distribution, same mechanism). Same population, same treatment. Meanwhile, the Haute Autorité de Santé modifies its recommendations on the indication concerned. The optimal decision changes, although no covariate has moved. The evidence has remained within its internal domain and has left its external domain, and this departure suffices to invalidate it for the targeted decision. To reduce the domain to its statistical face is to believe one has shut the door because one has drawn the top bolt.

VI. Evidence is a process

The two external faces of the domain have one thing in common: they move. Windows close, recommendations are revised. This is why the parameter T is not a refinement but a necessity, and it is T that completes the transformation of the nature of the object we are handling.

A piece of evidence is not a stable state. It is valid at a date, under conditions that have a lifespan. A deployed model sees the population it serves drift slowly away from the one on which it was fitted: the phenomenon is so well identified that regulators have ceased to ignore it, as I will return to. A causal identification judged impossible yesterday may become possible tomorrow with a new source, reopening a closed question. A recommendation reverses. In each of these cases, the evidence does not become false: it expires. The distinction is crucial. False evidence was poorly established; expired evidence was well established and has ceased to apply. To confound them is either to discard what was sound or to keep what is no longer.

Hence an abrupt reformulation: outside its domain, a piece of evidence is not false; it is expired. And a piece of evidence that can expire is not an object; it is a process, with a start of validity, a window, and a condition of end. One does not certify a process once and for all. One commits to it, one monitors it, one renews it. The vocabulary of the certificate (issued, acquired, definitive) is ill-suited. The vocabulary that fits belongs to the time-bound commitment, and it is toward this that everything preceding leads.

VII. What a generator cannot render inferable

Before getting there, one must treat a question that the rise of generative models makes burning, and treat it with a precision on which the credibility of the whole edifice depends. Can a generator (a synthetic population, a digital twin, a simulated cohort) produce content that was not in its data? The naive answer is no, and it is right for the wrong reasons, which makes it dangerous.

I will avoid the word *information* here, because it is so loaded that it invites the quarrel rather than settling it. Let us speak of **identifiable content**: what a set of constraints (data plus structural assumptions) makes it possible to distinguish.

- **First lemma:** any content absent from the constraints of the problem is non-identifiable. This is the heart of identification theory, as modern causal inference has formalized it, from Hernán and Robins's target trial emulation (*American Journal of Epidemiology*, 2016) to Pearl's identifiability calculus. One can estimate only what the data, under declared assumptions, make it possible to distinguish.
- **Second lemma:** what is not identifiable is not recoverable by generation, for a generator has access to nothing other than these same constraints. *Conservation theorem:* no generator renders identifiable a content absent from the constraints.

This formulation holds where the previous one was exposed. A theorist could object that a latent representation brings out a structure never explicitly observed, that a model makes exploitable a regularity no one had managed to name, and he was right, as long as one spoke of "information." But what the latent representation exhibits is not a new content; it is a content already identifiable in principle from the constraints, which the generator renders inferable, that is, operational. The right verb is therefore not *to create* but *to render inferable*.

A generator does not create content; it transforms an implicit content, already identifiable within the constraints and the model's inductive bias, into content mobilizable for a decision.

The consequence for practice is at once demanding and liberating. Demanding, because it forbids the sales promise that one generates patients where there are no data: one never generates anything but in the cast shadow of what one already has. Liberating, because it assigns the generator a real and defensible value: rendering a latent structure computable, exploring the region of the plausible that the constraints authorize, at a cost incommensurable with that of a recruitment. To confound the two, to take rendering-inferable for a creation of content, is the exact symmetry of the skeptic's error who takes the observed sample for reality. The naif of data and the naif of the generator hold hands, and neither knows it.

VIII. To decide and to reduce uncertainty

I announced, in introducing loss, that a piece of evidence has two functions, not one. If a decision is a rule that turns information into an act, then a piece of evidence serves two distinct things:

- to choose the act,
- and to judge whether it is worth knowing more before choosing.

The health technology assessment literature has long identified that a piece of evidence fulfills at least two distinct functions. The first consists in supporting a present decision. The second consists in determining whether the remaining uncertainty still justifies producing new knowledge. This second function is precisely what value-of-information analysis (Value of Information, VOI) quantifies. The expected value of perfect information (EVPI) measures the maximal benefit that the complete disappearance of uncertainty before deciding would procure; the expected value of sample information (EVSI) estimates the expected gain of a realistic study of given size. From Claxton's founding work to ISPOR's methodological recommendations, the question is therefore no longer only "which decision is optimal today?" but also "how much is it worth investing to reduce uncertainty before confirming this decision?"

This distinction reveals a property often ignored of synthetic cohorts. A cohort may perfectly preserve the optimal decision under the chosen loss function while deforming the uncertainty structure that grounds the value of future research. In other words, it may lead to the right choice today while wrongly suggesting that an additional study is useless, or conversely that it is indispensable. It is then substitutable for deciding, but not for learning.

The difference is fundamental. The optimal decision depends solely on the position of the minimum expected loss. The value of information, for its part, depends on the possibility that this minimum shifts when new observations become available. Two distributions may thus lead to exactly the same decision while inducing very different EVPI or EVSI. Preserving the decision therefore does not imply preserving the value of information. The two properties fall under distinct mathematical constraints.

We thus propose to distinguish two levels of substitutability.

1. The first, which may be called **decisional substitutability**, requires that the synthetic data lead to the same optimal decision rule as the real data under a given decision context.
2. The second, more demanding, corresponds to an **informational substitutability**: the synthetic data must also preserve the value of the information liable to modify this decision. A piece of evidence that satisfies only the first property remains

intrinsically myopic. It allows one to choose correctly today, but no longer allows one to evaluate correctly whether one should continue searching tomorrow.

This distinction opens a natural extension of current validation frameworks. It is no longer enough to demonstrate that the decisions produced from synthetic data reproduce those obtained on real data. It also becomes necessary to verify that value-of-information analyses lead to the same research priorities, the same trade-offs between immediate decision and acquisition of new data, and the same conclusions on the marginal utility of further studies. A fully substitutable synthetic cohort should preserve not only the optimal action, but also the economy of learning that surrounds this action.

IX. Substitutable is a degree, and a date

Everything preceding has handled substitutability as a state: one source replaces another, or does not. It is a convenience that must be abandoned, for it is false.

Substitutability is gradual. A cohort preserves certain properties better than others; it supports certain decisions and betrays others. The right object is not a Boolean but a degree (let us call it σ, between zero and one), or, more usefully, a profile: substitutable for this family of decisions, not for that one. To say of a synthetic population that it is "substitutable at 0.82" has meaning only when accompanied by its decision, its loss vector, its domain, and the uncertainty interval surrounding this figure, an uncertainty that is twofold, for it combines sampling randomness, which affects any finite cohort, and generation uncertainty, which stems from the fact that the generator is itself estimated on finite data. Conformal prediction (Vovk, Gammerman, and Shafer, 2005; for an introduction, Angelopoulos and Bates, 2021) offers a distribution-free framework for producing such intervals with non-asymptotic guarantees.

The move from Boolean to degree is not a concession of modesty. It is what makes substitutability governable. A binary state is not negotiated; a degree accompanied by a profile and an uncertainty can be discussed, thresholded, audited. One can decide that a substitutability of such a level suffices for an exploratory decision and not for a confirmatory one. Substitutable is not a state: it is a degree, and a date. And a dated degree, attached to a declared use, already bears all the features of a time-bound commitment.

X. An oncological decision: unrolling the structure

The moment has come to cease illustrating and to unroll the structure on a case. The whole theory rests on contextualization; it must therefore let itself be unrolled on a real decision, from the top down to the degree of substitutability. Let us take a precise oncological decision, and follow the chain without skipping a link. The empirical magnitudes that follow are marked as *to be documented*: the theory fixes the structure of

the demonstration, not the figures, which are a matter of measurement and not of argument.

The decision. In a patient with hormone-dependent breast cancer, without HER2 overexpression, should adjuvant hormone therapy be intensified? The decision is a rule δ that, to a patient profile, assigns an action within a finite set of therapeutic options. This profile is not a point in a text: it is a vector over several hundred tabular variables, which is why one does not reason about it as about language, any more than one reasons about a patient carrying a mutation such as BRAF V600E by applying to it the intuition of a word model (LLM).

The loss vector, contextual. L weighs the expected benefit on recurrence-free survival, toxicity, equity of access, cost, acceptability to the patient, and regulatory compliance. The weightings are not universal: for an elderly patient with comorbidities, the weight of toxicity rises; for a young patient, that of recurrence-free survival dominates. The loss is therefore fixed within the context of the decision, not before it.

The domain, internal times external. Internally: the statistical region of the profiles actually represented in the data mobilized, and the clinical relevance of the hormone-dependent, HER2-negative subtype. Externally: the temporal window of the protocols considered, and the organizational framework of the recommendations in force at the moment of the decision. A patient whose profile falls outside the statistical region leaves the internal domain; a change of recommendation leaves the external domain. Both invalidate, for opposite reasons.

Time. The validation is dated. It holds as long as the distributions do not drift and the recommendations are not revised; otherwise it expires. The evidence mobilized therefore carries a condition of end, not only a condition of use.

The validation. One mobilizes a cohort (real, synthetic, or a twin targeting, for example, the toxicity profile, terrain that a device such as ToxTwin explores) to estimate what δ needs. The question is not "is this cohort real?" but "does it preserve, on the declared domain, the properties necessary to δ under L?" One tests it by decision concordance: does it produce, where it is used, the same actions as the reference cohort, weighted by loss, and not by mere distributional proximity, which can be excellent without the decision being preserved. One also submits it to review by expert clinicians, for maximum likelihood smooths anomalies, and the anomaly (the atypical responder, the rare signature) is often the most precious object, the one that no Kullback-Leibler divergence signals.

The degree of substitutability. The result is not a verdict but a graduated measure. We propose to represent substitutability by a continuous index σ, defined relative to a decision rule δ, a loss function L, and an applicability domain explicitly declared. A natural formalization consists in defining σ as a normalized decision regret:

$$\sigma = 1 - \frac{\mathbb{E}_\theta[\, L(\delta_S, \theta) - L(\delta_R, \theta)\,]}{L_{\max}}$$

With:

- δ_R: decision rule obtained from the real data,
- δ_S: decision rule obtained from the synthetic cohort,
- $L(\delta, \theta)$: loss associated with decision $\delta$when the true state is θ,
- $\mathbb{E}_\theta$: expectation over the distribution of states of the world (or of patients),
- $L_{\max}$: maximal regret retained as the normalization reference.

If one uses the definition of Bayesian risk as $R(\delta) = \mathbb{E}_\theta[L(\delta, \theta)]$, then:

$$\sigma = 1 - \frac{R(\delta_S) - R(\delta_R)}{L_{\max}}$$

Thus, σ = 1 means that the substitution induces no average decisional loss; as the clinical, economic, or organizational cost of the divergences increases, σ decreases toward zero. The index therefore measures not the statistical proximity between two distributions, but the expected cost of replacing one by the other for the decision considered.

This definition makes substitutability intrinsically contextual. The same cohort may present a high value of σ for a therapeutic-intensification decision and a much lower value for a de-escalation decision, because the consequences of errors, and therefore the loss function, differ. There exists, consequently, no universal certification of a cohort, only degrees of substitutability relative to an explicitly defined use. Any estimate of σ should be reported with its uncertainty interval, its domain of validity, and its decision profile. A natural extension consists in verifying also that the cohort preserves the value of information (EVPI, EVSI) associated with this decision, in order to assess not only its capacity to reproduce present choices, but also its capacity to preserve the value of future research.

That the same chain applies word for word to a real-world-data cohort (that one declare its decision, its contextual loss, its two-faced domain, its date, and that one measure its degree of substitutability) makes the essential explicit, and verifies in passing the composability criterion of Section III: the validation of a sub-decision (estimating toxicity)

assembles with that of the encompassing decision (to intensify or not) because both are indexed by the same apparatus. The doctrine is not an apologia for the synthetic. The synthetic is merely its first demonstration, because it renders visible, by contrast, what real data concealed behind its self-evidence.

XI. Health is only the first terrain

A theory is judged also by what it illuminates beyond its point of departure. Nothing in the structure D, L, Δ, T is proper to medicine. Wherever computational evidence informs a decision under constraint, the same impossibilities lie in wait and the same resolution imposes itself.

In finance, a risk model calibrated on a market regime expires when the regime changes: its temporal domain is short, and confounding it with a stable property is a classic source of mishap. In aeronautics, a certification simulation is valid only for the declared flight envelope: the domain there is an engineering notion before being a statistical one. In cybersecurity, a detector trained on known attacks has no guarantee outside their distribution, and extrapolation there is precisely the adversary. In each of these fields, evidence is relative to a decision, under a contextual loss, within a two-faced domain, at a date.

This transfer is not a gratuitous extrapolation: it satisfies the transferability criterion that the metatheory demanded, provided one declares its conditions, which I have just done for three domains. These transfers remain, at this stage, conjectures of scope: the theory demands their test domain by domain; it does not provide it here. To present health as the first terrain rather than as the perimeter is therefore not a façade ambition; it is the direct consequence of a theory that, if it is correct, could not stop at the border of a discipline. Health offers only its densest version, because the stakes there are vital, the subpopulations numerous, and the loss trade-offs the most morally charged.

XII. Why this theory was hard to see

A reviewer will no longer, at this stage, pose a methodological question. He will pose a question of history: who has already said something similar? It is a good question, and the worst answer would be to feign absolute originality. The truth is that nothing here is new taken in isolation. Everything is new in the assembly.

Each of the fields summoned already held a fragment of the problem, and saw only its own.

- Wald had the decision and the risk, but under a scalar loss, and without a word on the domain or on time.

- The GRADE family had the graded quality of evidence, but treated as a property of a body of evidence, whereas we read it as a degree relative to a decision.
- The validation of structure-activity models, codified by the OECD, had the applicability domain, but geometric and static.
- Pearl, Hernán, and Robins had identifiability (the boundary of what the data can support), but for causal estimands, not as a general conservation law of evidence.
- The value-of-information tradition had the value of reducing uncertainty, but as a research-prioritization tool, not as the second function of all evidence.
- Recent regulation, finally, with the FDA's predetermined change control plans (statutory authority written into Section 515C of the Federal Food, Drug, and Cosmetic Act by the FDORA of 2022, specified by the final guidance of December 2024 and the lifecycle management draft of January 2025).
- The intended purpose of the European regulation on artificial intelligence had lifecycle validation, but as a practice without the theory that explains it.

The contribution is therefore not a creation; it is an integration. The theory inherits the decision from Wald, the domain from the QSAR tradition, the identifiability boundary from causal inference, the value of uncertainty from health technology assessment, the lifecycle unit of use from regulation, the gradation of quality from GRADE. It retains the core of each fragment, reread relationally. It abandons the assumption these fragments shared without knowing it, the one that makes validity, quality, or domain properties of an object. And it finally renders explicable a troubling coincidence: why these frameworks, developed independently, kept repeating the same gestures, namely declaring a use, a criterion, a limit, a window. They did so because each touched a different face of one and the same relation.

If the theory was hard to see, it is because it dwelt in the in-between of the disciplines, and none had reason to lift its eyes from its own fragment.

XIII. From emergent property to promise, and from promise to contract

It remains to name what emerges from the journey, and to name it in the right order, for the order here is the thought itself.

The primary fact, the one from which all the rest follows, is that validation belongs to none of the terms present. It is not a property of the data, nor of the model, nor even of the decision. It is an emergent property of a relation (an evidence source, a decision, a context), indexed by time and revisable. It is born of their encounter and dies with it; none of the three possesses it, as neither of the two sides of a valley possesses the river. This is

the truly new statement, and it is from this that one must start, for the two notions that follow are merely its consequences. That validation is called emergent is not an unverifiable metaphysical assertion: it is an interpretive framework whose refutable burden rests entirely on the theses that precede. Let a single validity be established independent of the decision, the loss, and the domain, and emergence falls with them.

1. **First consequence:** if validity emerges from a relation indexed by time, then a piece of evidence never affirms an unconditional fact. It states a conditional promise: "if the decision is this one, if the loss is that one, if one remains within this domain, and as long as this window lasts, then the source supports the action to this degree." Evidence, correctly understood, has the logical form of a guarded implication, not that of a finding. And the conditional promise is not proper to health: it is the form of every guarantee in mathematics, where a theorem holds under its hypotheses; in physics, where a law holds within its regime; in machine learning, where a bound holds under its distribution; in regulation, where an authorization holds for a use. To see evidence as a conditional promise is to render it comparable from one domain to another.
2. **Second consequence:** the contract. When a conditional promise must be kept between parties (a manufacturer and a regulator, an evidence provider and a decision-maker), it formalizes socially. It becomes a **use contract**: a commitment of substitutability, graduated, dated, with testable clauses and a declared domain, honored as long as its conditions hold and renegotiated as soon as they give way. The contract is therefore not the fundamental concept; it is the particular case where the conditional promise becomes juridically or organizationally explicit. And it is, as we have seen among the fragments, the direction regulation has already taken: lifecycle change control at the FDA, the intended purpose of the European regulation, already make the declared use, and not the object alone, the unit of certification. The regulatory world already formalizes conditional promises without having the theory that says so. This note proposes that theory.

One must conclude on the symmetry that will have run through the whole text, because it is its lesson.

- There is a naivety that takes the observed sample for reality,
- there is a naivety that takes the generator for a source of content,
- and there is, deeper and more widely shared, a naivety that takes evidence for an object, when it has never been anything but a relation, and the relation for a state, when it has never been anything but a promise.

To renounce the first without renouncing the others is to change idols. A piece of evidence is not an object one certifies; it is a conditional promise one keeps, and a contract, when

it comes about, is merely the social name of that promise, for parties, for an object, and for a time.

References

1. Wald A. *Statistical decision functions*. New York (NY): John Wiley & Sons; 1950.
2. Organisation for Economic Co-operation and Development. *Guidance document on the validation of (Q)SAR models*. OECD Series on Testing and Assessment No. 69. Paris: OECD Publishing; 2007.
3. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926. doi:10.1136/bmj.39489.470347.AD.
4. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c332. doi:10.1136/bmj.c332.
5. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. *BMJ*. 2015;350:g7594. doi:10.1136/bmj.g7594.
6. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*. 2012;17(5):4791-4810. doi:10.3390/molecules17054791.
7. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183(8):758-764. doi:10.1093/aje/kwv254.
8. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669-688. doi:10.1093/biomet/82.4.669.
9. Vovk V, Gammerman A, Shafer G. *Algorithmic learning in a random world*. New York (NY): Springer; 2005.
10. Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv* [Preprint]. 2021. arXiv:2107.07511.
11. Rothery C, Strong M, Koffijberg HE, Basu A, Ghabri S, Knies S, et al. Value of information analytical methods: Report 2 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force. *Value Health*. 2020;23(3):277-286. doi:10.1016/j.jval.2020.01.004.
12. U.S. Food and Drug Administration. *Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device*

software functions: guidance for industry and Food and Drug Administration staff. Silver Spring (MD): U.S. Food and Drug Administration; 2024.