The Fourth Generation of Clinical Data

Why the observed individual was never the scientific object.

Medicine treats individuals; the science that grounds it has only ever studied distributions. At the bedside, the clinician treats a person. The researcher, by contrast, has only ever had access to samples from which populations are inferred. This asymmetry is old, and no one disputes it. What is new is that a technology now makes it literal.

The debate over "synthetic data" is pitched at the wrong level. It sets fake patients against real ones, when the question is not about patients at all. A generative model does not, fundamentally, manufacture fictional individuals; it makes explicit the representation of the distribution that clinical research was already pursuing without saying so. The thesis of this article is therefore historical before it is technical: we are entering the fourth generation of clinical data, the one in which data ceases to be a thing one stores and becomes a model one queries. And the corollary is more uncomfortable than the thesis: the scientific object of a study was never the observed individual, but the distribution that individual allowed to be estimated. The individual was the index, not the target.

A boundary, set down at once, because it governs everything that follows. This thesis holds for the science of inference: epidemiology, treatment effects, risk structures, trajectories. It does not hold for the singular decision, where the individual becomes the target again rather than the index. The distinction between treating and studying is not a rhetorical detail; it is the boundary that keeps this article from saying something foolish.

The three ages of data, and the fourth

Clinical data has known three ages, and each treated data as an accumulation to be stored and then mined.

- The paper record was its archival age: the individual, a trace filed away in a cabinet.
- The relational database was its transactional age: the individual, a row that could be queried, joined, aggregated.
- The real-world data warehouse was its analytical age: the cohort, an aggregate on which a biostatistician writes an analysis plan for one question, then another for the next.

Three ages, one same logic: data is a stock, and knowledge is extracted from it by query.

The fourth age no longer extracts; it learns. From the cohort it builds a representation of the population, and it is that representation, now, that one queries. The center of gravity leaves the archive for the model. Data ceases to be an archive; it becomes a model. That sentence is not a stylistic flourish. It is the exact statement of the break, and the rest of this article unfolds its conditions and its limits.

The image that holds this displacement together from end to end is cartographic, and I extend it deliberately, because it works everywhere the reasoning must go. The first three ages collected the territory. The fourth draws its map. A map is not a miniaturized territory; it is a selective representation, one that keeps what serves movement and discards the rest. That is precisely what a population model does, and it is why the objection of the missing individual falls flat: one does not fault a map for failing to contain every tree.

The emergence of the model

The word "model" arrives too quickly if it is set down without being built. Let us therefore retrace the chain, slowly, because it is the chain that makes the rest intelligible.

One starts from a cohort. That cohort is a sample, produced by a particular recruitment.

From that sample, one infers. What one infers are the properties of an underlying distribution, which the sample shows only in part.

Of that distribution, one may be content to estimate a few parameters, as classical statistics does; or one may learn a complete representation of it, and that is the gesture of the generative model.

The synthetic population appears only at the end of this chain, as a realization of the model, never as its starting point.

This detour shows that nothing, in the fourth age, breaks with biostatistics. The objective remains to infer an invisible structure from a limited sample. What changes is the object learned: no longer a few numbers summarizing the distribution, but a rule capable of regenerating it. And this is why the real-world data warehouse, however exhaustive, was already no longer reality. It was a topographic survey: a measurement of the territory, taken with an instrument, from a vantage point, on a given date. Moving from the survey to the map does not lose the territory, since the territory was never in the survey.

This patient does not exist, this cohort is representative: the formula ceases to be a slogan once one has understood that the object was always the distribution, and never the surveyed point.

Compression: from an accumulation to an equation

Here is the gesture that the vocabulary of "fake patients" prevents one from seeing, and which is probably the deepest contribution of this change of generation. For two centuries we treated data as an accumulation: the more one had, the better, and knowledge lay at the end of the pile.

The generative model does the opposite. It seeks the most compact description capable of reproducing the observed regularities. It does not preserve the cohort; it summarizes it into a rule from which the cohort can be regenerated. The cohort ceases to be a set of rows; it becomes an equation.

This operation carries an implicit name that can be described without being theorized. A cohort contains three things intermingled:

- Information,
- Redundancy,
- And noise.

Redundancy is the regularities repeated from one patient to the next. Noise is the idiosyncrasy proper to each trace, what belongs to one individual and to no one else. Useful information is the structure of the population: the dependencies, the correlations, the shapes of trajectories. The model retains only this information; it discards the noise, and it compacts the redundancy. One understands then, and only then, why the individual matters little in this operation: the individual was, in large part, the noise that compression discards, not the signal it keeps.

A point of rigor, without which the image would mislead. What is at stake is a compression of information, not necessarily an economy of parameters: some deep models count more parameters than the cohort contains values. The compression is that of the relevant structure, not that of the raw count. To say that the cohort "becomes an equation" is a rhetorical compression of this idea, not a literal claim about the number of terms. It is the survey becoming a map: not lighter in ink, but selective in what it keeps.

A representation, not patients

Once it is granted that the object is the distribution, the generator joins a well-established family.

A large language model does not store sentences; it learns a representation of language from which sentences can be produced.

A diffusion model does not store images; it learns a representation from which images can emerge.

There is no reason for a clinical population to escape this logic: a population generator is a model that represents populations, and the movement is exactly the same.

This representation has a geometry. Rather than name it by its implementation, I will name it by its function: it is a probabilistic geometry, the space of patient configurations that the distribution renders plausible, with their relative densities. Deep models often implement this geometry in the form of a latent space; but the concept does not depend on that implementation, and to conflate the two would be to confuse the map with the printing technique.

In this geometry, a synthetic patient is not an invented individual: it is a point one projects, a position one reads off the map. One does not ask that point to exist; one asks it to be in the right place.

The blank zones of the map

Every map has blank zones, and it is there, exactly, that rigor is won or lost. As long as one reads the map within a densely surveyed region, it is reliable, because it interpolates between real measurements. At the margins, where the surveyor never passed, the map is no longer a measurement of the territory; it is a conjecture of the cartographer. A map is reliable only where it has been surveyed. The distinction is sharp and must be held: to interpolate within the surveyed support is not to extrapolate into the blank zone.

Three dangers live in these blank zones, and the metaphor names them all.

- The first is survey bias: what was over-represented in the surveying will be over-represented on the map, with all the more assurance as the drawing is clean; the literature documents that deep generative models amplify the biases of their data and then produce unbalanced, unrepresentative populations [1].
- The second is the unsurveyed zone: a plausible but scarcely observed region, where the model extrapolates without any data to constrain it.
- The third is the rare: a generator does not invent the rare, it copies the little it has seen of it, and propagates the peculiarities of that handful as though they described the population. A map does not reveal a buried city from a single shard; a model does not reveal a subpopulation from a dozen cases.

The interface: data becomes an itinerary

If the model is the map, then a synthetic cohort is merely its use: the itinerary a human asks the map to trace. And this use changes in nature. Data becomes conversational.

Yesterday, one queried a warehouse in SQL, and each question demanded its own query. Tomorrow, one converses with a population model. "Show me patients comparable to this

one, but without renal failure" does not describe a query on a table; it describes an itinerary through a space. The shift is of the same nature as that of search engines moving from the index to the language model: one no longer searches for a row, one navigates a representation.

One must resist the euphoria this fluidity inspires, because it hides the inference beneath the conversation. Each itinerary remains a computation. "Without renal failure" is legitimate only if the map has correctly surveyed the dependency between that condition and the rest; otherwise the itinerary crosses a blank zone while giving the illusion of a road. The conversational interface does not remove the inference; it slips it beneath the surface, which makes it easier to forget, and therefore more dangerous to leave unvalidated.

Validation demonstrates substitutability, not resemblance

To validate a population model is to measure how far its map remains reliable for a given use. And the decisive criterion is not resemblance. A representation is not validated on its fidelity of appearance; it is validated on its operational substitutability.

Two levels must be distinguished, and their confusion explains most of the misunderstandings.

- *Statistical fidelity* measures how closely the learned distribution matches the real one: low Wasserstein distance, non-significant log-rank between survival curves, pMSE near indistinguishability [2][3].
- *Operational substitutability* measures something else: does a model trained on the generator's outputs, then tested on real data, preserve the conclusions? This is the Train-on-Synthetic, Test-on-Real protocol, which asks not whether the map resembles the ground, but whether one reaches the destination by trusting it.

The distinction is not a refinement. A map can be faithful in its broad outlines and misleading on the precise itinerary, because it missed a dependency that the margins do not reveal. This is why validation is always done against external real measurements, never against the model's judgment of itself alone, as established by the work that confronts reference generators with clinical indicators computed on real data [4].

Conditional generation toward under-represented subpopulations confirms this: augmenting a dataset with conditionally generated cohorts can improve generalization to those subpopulations [5], but in exact proportion to the fidelity with which they had been surveyed.

The condition is not a detail; it is the subject. A representation without a protocol of validation against real data is not a population model: it is a mapped assertion. TweenMe®, as a terrain of implementation, does not exempt itself from this requirement; it makes it operational, and it is at that level that the matter is judged, not at the level of the promise.

The principle of clinical representation

Everything above condenses into a single principle, and it is this principle, more than the technology, that deserves to be retained.

Principle of clinical representation. A cohort is never studied for itself. Its sole function is to provide a representation of the distribution that generated it, faithful enough to support a family of decisions. Generative models do not modify this principle; they make that representation explicit and manipulable.

The value of this statement is that it depends on nothing we have discussed. It depends neither on neural networks, nor on synthetic data, nor even on medicine. It describes a general property of statistical inference: the scientific object is the representation of the distribution, and not the observations that allowed it to be estimated. The decisive word is *enough*: fidelity is never absolute, it is relative to the family of decisions targeted. It is exactly the substitutability criterion, raised to the rank of a principle. The first three ages of data applied this principle without formulating it, leaving the representation implicit in the biostatistician's head. The fourth makes it explicit, and with it makes explicit its conditions of validity, which is at once a gain and an exposure.

Medicine leaves objects behind

It remains to close where this article joins a trajectory larger than itself. I have argued elsewhere that, in a regulated system, evidence is not an object one holds but a relation one establishes, valid under explicit conditions and void outside them. The present article is its counterpart: the representation, too, is not an object. Three displacements say so, and they belong to the same movement.

A cohort is not an object; it is a sampled observation.

A synthetic population is not an object; it is a representation.

Evidence is not an object; it is a relation.

Each time, medicine leaves the thing for what links it to the distribution from which it proceeds. The representation is an emergent property, just as evidence is: it resides in no patient, real or synthetic, but in the relation a map maintains with a territory it will never contain.

The fourth generation of clinical data does not, then, consist in manufacturing patients who do not exist. It consists in recognizing that the object was never the patient, and in finally handling the distribution as a map rather than as an archive. This is a considerable gain, and it is for that very reason that one must mark its limit rather than pass over it in silence. The map becomes richer; the territory does not change. The model extends what one can ask of a cohort; it does not extend what that cohort knew. To confuse the two is precisely the error that validation exists to prevent.

References

1. Fasseeh AN, ElShafie S, ElMahalawy II, et al. Generating realistic synthetic patient cohorts: enforcing statistical distributions, correlations, and logical constraints. *Algorithms*. 2025;18(8):475. doi:10.3390/a18080475.

2. Shi J, Xu Y, McKenzie FE, et al. Generating high-fidelity privacy-conscious synthetic patient data for causal effect estimation. *J Am Med Inform Assoc*. 2022;29(12):2078-2088. doi:10.1093/jamia/ocac174.

3. Cipriani M, Di Rocco L, Puopolo M, Alfò M. A flexible parametric approach to synthetic patients generation using health data. *Stat Methods Appl*. 2025;34(4). doi:10.1007/s10260-025-00800-5.

4. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. *BMC Med Res Methodol*. 2020;20:108. doi:10.1186/s12874-020-00977-1.

5. Jarrett D, Cebere B, Liu T, et al. HealthGen: generative model to enhance medical time series for extrapolation to underrepresented populations. *Nat Commun*. 2023;14:3290. doi:10.1038/s41467-023-36938-1.

6. Vétillard J. Evidence is a conditional promise: toward a relational theory of the validation of computational proofs and synthetic populations. Twingital Institute Working Paper No. 1. Paris: Twingital Institute; 2026.

7. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27(3):379-423;27(4):623-656. doi:10.1002/j.1538-7305.1948.tb01338.x.

8. Cover TM, Thomas JA. *Elements of Information Theory*. 2nd ed. Hoboken (NJ): John Wiley & Sons; 2006.

9. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge (MA): MIT Press; 2016.