To represent is not to reproduce

Projection, validation, and the domain of substitutability of synthetic cohorts.

From principle to instrument

The previous article defended a counter-intuitive thesis: in the fourth generation of clinical data, the scientific object is no longer the observed patient but the population distribution that observations allow us to estimate. A synthetic cohort therefore never holds value because it resembles real patients. It holds value if it preserves the properties required by the analyses for which it was built. This thesis remained abstract. The present article leaves doctrinal ground for that of instrumentation, and it stays there.

One proposition sums it up. A tabular generator is an instrument that projects an imperfect clinical survey. Its validation consists in delimiting the zones where this projection remains substitutable for the observed distribution, and those where it becomes speculative.

The word *projection* is the pivot, and it must be fixed before it drifts. A projection is a transformation that preserves certain properties of an object, discards others, and reorganizes the rest. The three operations are not equivalent: to preserve is not to reorganize, and a latent representation can bring out regularities that the observations did not display. To choose a generative architecture is therefore to choose what one preserves, what one loses, and what one reorganizes. Nothing in this article uses *projection* in any other sense.

A bound, set at the outset. The present article stops at the construction and validation of a synthetic cohort. Predictive modeling, calibration, clinical decision, and medico-economic evaluation belong to a distinct layer; they presuppose a valid representation without constituting its criterion, and are not treated here.

A clarification is needed before going further. The real is never directly observed. Observations allow one to estimate a distribution of it. The generator builds a representation of that distribution, not of the real itself. Validation therefore never bears on the real, but on the adequacy of this representation to a declared domain of use. Everything else in the article follows from this hierarchy.

It remains to justify that this object deserves separate treatment, and this is the reasoning's best foothold. Tabular generation is not a special case of image generation. An image possesses a geometry given by the sensor: a grid of pixels, invariance under translation, a natural neighborhood between nearby points. A clinical table has none of this. Its variables are heterogeneous, its marginals often multimodal, its categories

strongly imbalanced, and no adjacency metric pre-exists the analysis. Synthesizing an image exploits a given geometry; synthesizing a clinical table first requires fabricating one. The consequence is immediate and deserves to be stated plainly: a tabular generator learns an implicit metric as much as a distribution. The evaluation literature has recognized this in observing that metrics cut for images poorly diagnose the failures of generators elsewhere (Alaa et al., ICML 2022). It is this asymmetry that forbids treating the tabular as a variant of the visual.

Three words, and a fallible instrument

Three definitions fix the frame. A tabular generator learns, from a finite sample, an approximation of the joint distribution of a set of heterogeneous variables, then produces new observations compatible with it. It does not memorize a cohort; it learns a rule. A synthetic cohort is the sample drawn from this estimated distribution; it is not a copy of the training cohort, no more than a second throw of the die is a copy of the first. The *space of the plausible* is the set of configurations to which the model assigns non-negligible probability: the support of the estimated distribution.

This last notion calls for a correction. The space of the plausible is not an intrinsic property of the population; it is the estimate, at a given moment, of the support that the data permit one to infer. The support is not discovered, it is estimated. And the estimate inherits the conditions of collection: inclusion criteria, practices, coding, recruitment, period.

Here lodges a blind spot that most manuscripts sweep under the rug. The clinical survey is treated as an imperfect sample. It is also an imperfect instrument. Coding errors, inter-observer variability, nomenclature changes, evolving practices, and the biases of measurement devices modify the distribution before any modeling. The generator does not merely project a finite sample; it projects the traces left by a measurement system that is itself fallible. The distinction is not rhetorical: an imperfect sample is corrected by more data, an imperfect instrument is corrected only by knowledge of its biases. One particular case makes it tangible: the estimated support depends as much on the observations present as on the mechanism of the absences, and depending on whether data are missing completely at random, conditionally at random, or not at random (MCAR, MAR, MNAR), the projected territory is not the same.

The support has an extent; it also has a resolution, and this second axis will govern the rest of the article. The extent says where the map exists; the resolution says with what fineness it informs. And this fineness is not a matter of density but of local information: a thousand nearly identical patients resolve a region poorly, whereas thirty sufficiently varied patients resolve it well. Resolution is not a spatial property but an informational one, a function of the variance, entropy, and separability of the observations, not of their mere number. Hence a formula that holds the rest together: a region can be covered

without being sufficiently resolved. A BRAF V600E profile can appear in a cohort without depositing enough information to support a conditional decision. Coverage reassures; resolution decides. One finds this pair at every stage: in the metrics, in extrapolation, in guided generation, in validation.

Every architecture is an inductive bias

Families of architectures are often presented as a succession of advances. The reading is misleading: all answer the same question from different assumptions. This part is illustrative, not thesis. A single sentence must remain from it: every architecture is a different inductive bias, that is, a different choice of what the projection preserves, discards, and reorganizes. And this bias is not a defect to be corrected. Inductive bias is not a weakness of the algorithm; it is the mathematical formalization of the assumptions that the projection imposes when observations become insufficient. One survey, moreover, admits several compatible representations; the choice of architecture therefore selects an identification hypothesis among those that the finite sample under-determines.

Bayesian networks factorize the joint explicitly according to an acyclic graph, $P(X) = \prod_i P(X_i \mid \text{parents}(X_i))$: they preserve auditable dependencies, they discard strongly nonlinear interactions at scale. Copulas dissociate, by Sklar's theorem, the marginals from the dependence, $F = C(F_1, \ldots, F_d)$: robust on small datasets, legible, at the price of an imposed form of dependence.

CTGAN (Xu et al., NeurIPS 2019) solves two problems specific to the tabular. Against the multimodality of continuous variables, a variational Gaussian mixture normalization represents each value by a mode indicator and a normalized scalar, $\alpha = (c - \eta_k)/(4\,\phi_k)$. Against categorical imbalance, a conditional vector and a training-by-sampling scheme impose regular exposure to rare modalities, under a Wasserstein loss with gradient penalty (Gulrajani et al. 2017). CTGAN does not learn to draw patients; it learns not to let the mean erase the modes that sub-populations contain. TVAE optimizes a lower bound on the likelihood, $\text{ELBO} = E_q[\log p(x|z)] - KL(q(z|x) \,\|\, p(z))$, more stable but more smoothed: where the GAN fools a judge, the VAE compresses then reconstructs. TabDDPM (Kotelnikov et al., ICML 2023) noises then denoises, Gaussian on the numerical, multinomial on the categorical (Hoogeboom et al., 2021): it does not learn the data, it learns the distance that separates it from noise. Causal generators, finally, such as DECAF (van Breugel et al., 2021), preserve mechanisms and not the joint alone, since two generators indistinguishable on the distributions can diverge on the effect of an intervention: preserving $P(X)$ does not preserve $P(Y \mid do(X))$.

What these families learn is, moreover, not the same object. A GAN learns a discriminative boundary, a diffusion a score field, a variational autoencoder a latent manifold, a Bayesian network a conditional factorization, a transformer a set of

contextual dependencies. They converge nonetheless toward one same target, a distribution. These architectures therefore differ less by what they represent than by the operator with which they build the representation. It is this operator, and not the targeted distribution, that the choice of architecture selects.

A stylized example fixes the idea, and it holds as illustration, not as measurement. Take one same cohort and one same task, entrusted to three generators. The GAN faithfully restores the marginals but may relax the high-order joint structure. The diffusion better captures this joint, as recent syntheses confirm. The causal generator preserves what the two others ignore, the effect of an intervention. None is better in the absolute; each projects according to its operator, and the right choice is the one whose losses fall outside the targeted task.

These families are not pure lineages, and it is their recombination that shows it best. One can graft a transformer or attention blocks onto a tabular GAN: the operation does not produce a better generator, it changes the inductive bias to capture non-local conditional dependencies between columns, where modal normalization and the conditional vector remained mute. This is the path of TT-GAN in health (Ryu et al., Sci Rep 2025), who report on this point a measured advantage over CTGAN and copula GAN. One can likewise hybridize diffusion and transformer, replacing the denoising perceptron with an encoder-decoder with conditional attention (MTabGen, Villaizán-Vallelado et al., ACM TKDD 2025), or diffuse in the continuous latent space of an autoencoder before decoding toward the table (TabSyn, Zhang et al., ICLR 2024). Accelerator schemes, where a GAN approximates a denoising in fewer steps to reduce inference cost, or regularizer schemes, where a discriminator penalizes the outputs of a diffusion model, follow the same principle.

The trap would be to conclude that the hybrid is by nature superior. It is not. An architecture can be hybridized, but hybridization has scientific value only if each component answers an identified failure mode of the projection: a poorly learned dependence, an erased rare mode, an extrapolated support, degraded privacy, excessive inference cost. Mode collapse calls for diffusion or PacGAN (Lin et al. 2018); non-local conditional dependencies, attention; column heterogeneity, typed embeddings and modal normalization; conditional rarity, controlled conditioning; inference cost, distillation or an adversarial accelerator; privacy, differential regularization and membership auditing; causality, an explicit structure and not one more layer. Grafting attention with no failure to correct adds no information: on small clinical cohorts, attention readily learns noise, reinforces memorization, and produces a matrix that impresses committees without explaining anything.

Hence a posture, and it extends the thesis instead of departing from it. One does not elect a queen architecture; one orchestrates a family of generators, each referred to a data profile, an applicability domain, a fidelity-utility-privacy trade-off, and a validation

protocol. The choice of a generator is not a decision of raw performance; it is a decision of projection under constraints.

There remains the most serious objection to this posture, and it must be stated at full force. If diffusion models now dominate most rankings, as recent syntheses report across a wide range of tabular benchmarks (Diffusion Models for Tabular Data, arXiv:2502.17119, 2025; Shi et al. 2025), why treat architecture as a second-order decision? The answer does not contest the benchmarks. They measure, better and better, several dimensions, fidelity, diversity, privacy, utility. They nonetheless aggregate these performances over average regions, and remain poorly equipped to qualify a conditional substitutability in a rare clinical margin. The dominance of diffusion is real at the center and undemonstrated at the edges. The leaderboard is true and beside the point. What is decided in the clinic is decided precisely where it falls silent.

Measuring a representation

The dominant question is badly posed: is this generator better than another? It presupposes a single measure of quality, which does not exist. The quality of a synthetic cohort is multidimensional, and its dimensions are not independent: improving one often displaces the others. To validate is not to maximize a score, but to characterize a trade-off.

Fidelity is examined at three scales. The marginals, variable by variable, by tests of Kolmogorov-Smirnov type or divergences suited to categoricals. The dependencies, by correlations, mutual information, and the logical constraints of the domain, a pregnancy in a man or a death prior to birth being a violation of structure and not a bad correlation. The joint, finally, by the Wasserstein distance, the Maximum Mean Discrepancy, or the pMSE, which trains a classifier to separate real and synthetic: the harder the distinction, the closer the distributions. Two datasets can share almost identical marginals and describe different populations if the dependencies have vanished; preserving each column does not guarantee preserving the cohort.

A recent advance renders this level legible at the scale of the sample and gives the notion of blank zones its metric anchoring. The triptych extends the precision-recall lineage for generative models (Kynkäänniemi et al. 2019) and specializes it at the sample level (Alaa et al., ICML 2022). The triptych separates three failures that aggregate distances confound. α-Precision measures the share of synthetic observations falling outside the real support: the un-surveyed zone, where the model extrapolates. β-Recall measures the share of real regions not covered: the survey bias transmitted. Authenticity measures whether the generator generalized or copied: exactly the trap of the rare, the handful of observations propagated instead of an estimated structure. Three axes name, term for term, the three ways of deceiving. And resolution reads directly there: a β-Recall

acceptable at coarse grain can mask a region covered but under-resolved, where the decision demanded a fine grain.

Utility is another dimension, and one of the most important observations of the recent literature. A representation can be remarkably faithful and strongly degrade the models trained on it; a slight loss of fidelity can, on the contrary, remain without consequence for the analysis actually conducted. The reason is structural: fidelity preserves P(X), predictive models exploit P(Y|X), and these two objects do not evolve together. Fidelity is necessary to utility; it is never its proof. The reference protocol is Train-on-Synthetic, Test-on-Real (Esteban et al. 2017), whose exact status will be specified later, for it does not measure an intrinsic quality but a relation.

Privacy is a third dimension, and answers another question: did the generator learn a structure or memorize individuals? Distance to the nearest real record, membership inference attacks, differential privacy formulations estimate the risk that a real observation be recovered. These measures are not rewards; they quantify a risk.

Presented as three independent objectives, these dimensions mislead. They form a Pareto trade-off: strengthening privacy adds noise, which can reduce fidelity, which can degrade utility; aiming for maximal fidelity pushes toward memorization, which increases disclosure risk. There is no optimal generator in the absolute sense; there are only operating points suited to a use.

Adequacy to use is therefore not the last question but the first: a cohort meant to train a diagnostic classifier does not impose the same requirements as a synthetic control arm or an inter-institutional share. Validation does not ask whether a cohort is good; it asks whether this representation is sufficiently faithful, useful, and protective for this family of analyses, in this declared domain.

To interpolate is not to extrapolate

No generator observes the population it claims to represent; it observes a finite sample and builds an approximation of it. Its natural domain is interpolation within the documented region. In a densely covered zone, it combines abundantly constrained dependencies, and the estimates are robust. As density decreases, these constraints fade and the architecture's assumptions take over. The further the model moves from the surveyed territory, the less it estimates the observed distribution and the more it expresses its own assumptions. The transition is continuous; it has no sharp boundary, and all validation consists precisely in determining where it becomes incompatible with the claimed use.

It is here that resolution and extrapolation part ways, and the confusion between the two is costly. Extrapolation is a coverage failure: one exits the support. Under-resolution is a grain failure: one remains within the support but without sufficient density to constrain it.

Both produce observations plausible in appearance, and this is precisely what makes them dangerous. The conversational interface of modern tools accentuates the risk: fluency conceals the transition between estimation and extrapolation, and the more natural the generation appears, the easier it becomes to forget that a region was never truly observed.

Hence a clarification. Generators produce so-called new patients, which is true in the combinatorial sense and false in the statistical sense. The lines produced are new combinations compatible with the learned distribution; they are not the discovery of new regions. Novelty bears on the realizations, never on the structure. And if recruitment is biased or the dependencies dated, the generator will faithfully interpolate an imperfect survey, the product of an imperfect instrument. No architecture creates information: it redistributes the available information. This is a direct consequence of information theory (Shannon 1948; Cover & Thomas 2006), not a limitation of current architectures. The nuance is important, for certain models do render exploitable an information previously inaccessible: they do not invent it, they reorganize it. The quality of the representation therefore depends as much on the quality of the survey as on the sophistication of the algorithm.

Fidelity is not substitutability

Here is the article's result, and everything preceding leads to it. An objection appears natural: if the cohort faithfully reproduces the distributions, why would it not replace the real cohort? The intuition is seductive and false. It confuses the quality of a representation with the validity of a reasoning built upon it. Fidelity is a descriptive property; substitutability, an operational property. The two are linked without being equivalent.

Two road maps illustrate it. The first reproduces the relief perfectly but omits secondary roads. The second simplifies the relief but preserves the whole network. The first resembles the territory more; the second leads more surely to destination. Fidelity asks whether the representation resembles the observations. Substitutability asks whether the analyses conducted on it lead to the same conclusions. One can answer the first excellently without satisfying the second, because fidelity approximates the preservation of $P(X)$ when the conclusions depend on $P(Y|X)$; a slight alteration of certain dependencies can be negligible on the global distributions and heavy on performance. It would be surprising for fidelity and utility to be strongly correlated.

Hence the exact status of Train-on-Synthetic, Test-on-Real. Its interest is not only to compare real and synthetic learning; it is to measure a relational property. TSTR does not assign a grade to the generator. It qualifies a representation relative to a family of tasks, and its result depends simultaneously on four elements: the generator, the task, the downstream model, and the evaluation protocol. Modifying one of them modifies the result without the generator having changed, and two generators can invert their ranking

depending on whether one trains a survival model, a classifier, or a causal estimator. Substitutability never characterizes a generator; it characterizes a representation relative to a family of decisions. A cohort built for a prognostic model can preserve the risk structures necessary to that task and remain insufficient to estimate a treatment effect, without contradiction, because the tasks do not exploit the same properties of the representation. This thesis is falsifiable, and it must be said so that it remains scientific: a strong and stable correlation between fidelity metrics and operational substitutability, measured across varied tasks, would refute it. It is the same logic as the applicability domains of QSAR models in predictive toxicology: a model is not valid because it predicts well in general, but because its domain of validity is known, documented, and respected.

The whole article then condenses into a stable statement.

Principle of clinical projection. Every cohort is a sample. Every projection applies an inductive bias. Every synthetic cohort is an estimate arising from that projection. Every estimate has a domain. Every validation is conditional on that domain. All substitutability vanishes outside that domain.

This principle does not repeat the principle of clinical representation of the parent article; it extends it. The first established that the scientific object is the distribution and not the observed individual. This one adds that this distribution, once projected by a generator, is substitutable only within the bounds of a domain, and null outside it. The one bears on what one studies, the other on what one can make of it.

Validation thus ceases to be a certificate. It becomes a conditional proof, valid for an explicitly declared domain, and whose very restriction makes its scientific value.

Validation is a process

A conditional proof has four components: it holds for a population, a period, a family of tasks, a protocol. Modifying one of them modifies the validated object. A cohort validated on a breast cancer does not become valid for a bronchial cancer; a validation on 2015 data does not necessarily survive the introduction of a new therapeutic strategy; a representation substitutable for a prognostic model is not necessarily so for a causal effect. Validation is never transportable without justification.

The deep reason is twofold, and the second is rarely named. Not only does the population drift, but the measurement instrument drifts with it. Medicine is not stationary: treatments evolve, diagnostic criteria are revised, nomenclatures and codings are renewed. Today's survey is not yesterday's survey, and a representation perfectly valid at its date progressively loses this property without the generator ever having ceased to function. Drift is not a defect of validation; it is the normal consequence of a world that changes, measured by an instrument that changes too.

The consequence is operational, and industry regularly underestimates it. The real cost of a generative platform is not the cost of producing a cohort; it is the cost of maintaining its domain of validity. Monitoring, revalidating, recalibrating, versioning, tracking: these operations, and not the initial training, decide the viability of a system in real conditions. The logic is known elsewhere, from QSAR applicability domains to change-management procedures for software medical devices and drift monitoring of deployed models. TweenMe constitutes one implementation of this doctrine: the product does not seek to optimize a generator, it seeks to industrialize the maintenance of the domain of validity.

No metric crosses this frontier. Neither a Wasserstein distance, nor a pMSE, nor a TSTR score, nor a level of differential privacy answers a question that is not its own. An excellent fidelity does not guarantee the relevance of a decision; an excellent privacy does not guarantee utility; an excellent TSTR does not guarantee a causal analysis. To seek a single score summarizing the quality of a cohort would amount to asking a single biomarker for the complete physiological state of a patient. The useful information is in the combination of indicators, not in their reduction.

A tabular generator therefore does not fabricate patients. It projects a fallible survey into an interrogable representation, and all its demonstrable value rests on the honest delimitation of the zones where this projection remains substitutable. Fidelity describes the proximity of two distributions; substitutability demonstrates that they lead to the same conclusions for a family of tasks. The rest is a matter of domain.

There remains a question this article does not open, and which another will have to carry. If the observed distribution were itself only the statistical projection of a geometry of constraints that the generator learns first, then the survey would be only a shadow, and the true object would stand one notch further. This displacement, from distributions toward geometries, is the thesis of a coming article, not the conclusion of this one.

References

1. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. In: Advances in Neural Information Processing Systems. 2019;32. Available from: https://arxiv.org/abs/1907.00503
2. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems. 2017. Available from: https://arxiv.org/abs/1704.00028
3. Lin Z, Khetan A, Fanti G, Oh S. PacGAN: The power of two samples in generative adversarial networks. In: Advances in Neural Information Processing Systems. 2018. Available from: https://arxiv.org/abs/1712.04086
4. Hoogeboom E, Nielsen D, Jaini P, Forré P, Welling M. Argmax flows and multinomial diffusion. In: Advances in Neural Information Processing Systems. 2021. Available from: https://arxiv.org/abs/2102.05379
5. Kotelnikov A, Baranchuk D, Rubachev I, Babenko A. TabDDPM: Modelling tabular data with diffusion models. In: Proceedings of the 40th International Conference on Machine Learning (ICML). PMLR. 2023;202:17564-17579. Available from: https://arxiv.org/abs/2209.15421
6. Zhang H, Zhang J, Srinivasan B, Shen Z, Qin X, Faloutsos C, Rangwala H, Karypis G. Mixed-type tabular data synthesis with score-based diffusion in latent space. In: International Conference on Learning Representations (ICLR). 2024. Available from: https://arxiv.org/abs/2310.09656
7. Villaizán-Vallelado M, et al. Diffusion models for tabular data imputation and synthetic data generation. ACM Trans Knowl Discov Data. 2025. doi:10.1145/3742435.
8. Ryu KS, et al. Tabular transformer generative adversarial network for heterogeneous distribution in healthcare. Sci Rep. 2025;15:10254. doi:10.1038/s41598-025-93077-3.
9. Borisov V, Seßler K, Leemann T, Haug J, Pawelczyk M, Kasneci G. Deep generative models for tabular data: A survey. IEEE Trans Knowl Data Eng. 2023;35(12):12340-12364.
10. Alaa AM, van Breugel B, Saveliev ES, van der Schaar M. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In: Proceedings of the 39th International Conference on Machine Learning (ICML). 2022. Available from: https://arxiv.org/abs/2102.08921

11. Kynkäänniemi T, Karras T, Laine S, Lehtinen J, Aila T. Improved precision and recall metric for assessing generative models. In: Advances in Neural Information Processing Systems. 2019. Available from: https://arxiv.org/abs/1904.06991

12. van Breugel B, Kyono T, Berrevoets J, van der Schaar M. DECAF: Generating fair synthetic data using causally-aware generative networks. In: Advances in Neural Information Processing Systems. 2021. Available from: https://arxiv.org/abs/2110.12884

13. Esteban C, Hyland SL, Rätsch G. Real-valued (medical) time series generation with recurrent conditional GANs. 2017. Available from: https://arxiv.org/abs/1706.02633

14. Snoke J, Raab GM, Nowok B, Dibben C, Slavković AB. General and specific utility measures for synthetic data. J R Stat Soc Ser A Stat Soc. 2018;181(3):663-688.

15. Li Z, Huang Q, Yang L, Shi J, Yang Z, van Stein N, et al. Diffusion and flow matching models for tabular data: A survey. arXiv. 2025. Available from: https://arxiv.org/abs/2502.17119

16. Shi R, et al. A comprehensive survey of synthetic tabular data generation. arXiv. 2025. Available from: https://arxiv.org/abs/2504.16506

17. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27:379-423,623-656. doi:10.1002/j.1538-7305.1948.tb01338.x.

18. Cover TM, Thomas JA. Elements of Information Theory. 2nd ed. Hoboken (NJ): Wiley; 2006.

19. Sklar A. Fonctions de répartition à n dimensions et leurs marges. Publ Inst Stat Univ Paris. 1959;8:229-231.

20. Organisation for Economic Co-operation and Development (OECD). Guidance document on the validation of (Q)SAR models. OECD Series on Testing and Assessment No. 69. Paris: OECD Publishing; 2007.