La quatrième génération de la donnée clinique

Pourquoi l'individu observé n'a jamais été l'objet scientifique.

La médecine soigne des individus ; la science qui la fonde n'a jamais étudié que des distributions. Au lit du malade, le clinicien traite une personne. Mais le chercheur, lui, n'a jamais eu accès qu'à des échantillons dont il infère des populations. Cette asymétrie est ancienne, et personne ne la conteste. Ce qui est nouveau, c'est qu'une technologie vient la rendre littérale.

Le débat sur les « données synthétiques » se trompe d'étage. Il oppose des faux patients à des vrais, alors que la question ne porte pas sur les patients. Un modèle génératif ne fabrique pas, fondamentalement, des individus fictifs ; il rend explicite la représentation de la distribution que la recherche clinique poursuivait déjà sans le dire. La thèse de cet article est donc historique avant d'être technique : nous entrons dans la quatrième génération de la donnée clinique, celle où la donnée cesse d'être une chose que l'on conserve pour devenir un modèle que l'on interroge. Et le corollaire est plus inconfortable que la thèse : l'objet scientifique d'une étude n'a jamais été l'individu observé, mais la distribution qu'il permettait d'estimer. L'individu était l'indice, pas la cible.

Une borne, posée d'emblée, parce qu'elle conditionne tout le reste. Cette thèse vaut pour la science de l'inférence : épidémiologie, effets de traitement, structures de risque, trajectoires. Elle ne vaut pas pour la décision singulière, où l'individu redevient la cible et non l'indice. La distinction soigner / étudier n'est pas un détail rhétorique ; c'est la frontière qui empêche cet article de dire une bêtise.

Les trois âges de la donnée, et le quatrième

La donnée clinique a connu trois âges, et chacun a traité la donnée comme une accumulation à conserver puis à fouiller.

- Le dossier papier en fut l'âge archivistique : l'individu, une trace rangée dans une armoire.
- La base relationnelle en fut l'âge transactionnel : l'individu, une ligne requêtable, jointe, agrégée.
- L'entrepôt de vie réelle en fut l'âge analytique : la cohorte, un agrégat sur lequel un biostatisticien écrit un plan d'analyse pour une question, puis un autre pour la suivante.

Trois âges, une même logique : la donnée est un stock, et la connaissance s'en extrait par requête.

Le quatrième âge n'extrait plus ; il apprend. À partir de la cohorte, il construit une représentation de la population, et c'est elle, désormais, que l'on interroge. Le centre de gravité quitte l'archive pour le modèle***. La donnée cesse d'être une archive ; elle devient un modèle***. Cette phrase n'est pas un effet de style. C'est l'énoncé exact de la rupture, et tout le reste de l'article en déroule les conditions et les limites.

L'image qui tient ce déplacement de bout en bout est cartographique, et je la file volontairement, car elle fonctionne partout où le raisonnement doit aller. Les trois premiers âges collectaient le territoire. Le quatrième en dresse la carte. Une carte n'est pas un territoire miniaturisé ; c'est une représentation sélective, qui retient ce qui sert à circuler et écarte le reste. C'est précisément ce que fait un modèle de population, et c'est pourquoi l'objection de l'individu manqué tombe à plat : on ne reproche pas à une carte de ne pas contenir chaque arbre.

L'émergence du modèle

Le mot « modèle » arrive trop vite si on le pose sans le construire. Reprenons donc la chaîne, lentement, car c'est elle qui rend la suite intelligible.

On part d'une cohorte. Cette cohorte est un échantillon, produit par un recrutement particulier.

De cet échantillon, on infère. Ce que l'on infère, ce sont les propriétés d'une distribution sous-jacente, que l'échantillon ne montre qu'en partie.

Cette distribution, on peut se contenter d'en estimer quelques paramètres, comme le fait la statistique classique ; ou on peut en apprendre une représentation complète, et c'est le geste du modèle génératif.

La population synthétique n'apparaît qu'au bout de cette chaîne, comme une réalisation du modèle, jamais comme son point de départ.

Ce détour montre que rien, dans le quatrième âge, ne rompt avec la biostatistique. L'objectif reste d'inférer une structure invisible à partir d'un échantillon limité.

Ce qui change, c'est l'objet appris : non plus quelques nombres résumant la distribution, mais une règle capable de la régénérer.

Et c'est pourquoi l'entrepôt de vie réelle, aussi exhaustif soit-il, n'était déjà plus la réalité. Il était un levé topographique : une mesure du territoire, prise avec un instrument, depuis un point de vue, à une date. Passer du levé à la carte ne fait pas perdre le territoire, puisque le territoire n'était jamais dans le levé.

Ce patient n'existe pas, cette cohorte est représentative : la formule cesse d'être un slogan dès lors qu'on a compris que l'objet fut toujours la distribution, et jamais le point relevé.

La compression : d'une accumulation à une équation

Voici le geste que le vocabulaire des « faux patients » empêche de voir, et qui est probablement l'apport le plus profond de ce changement de génération. Pendant deux siècles, nous avons traité la donnée comme une accumulation : plus on en avait, mieux c'était, et la connaissance était au bout de l'entassement.

Le modèle génératif fait l'inverse. Il cherche la description la plus compacte capable de reproduire les régularités observées. Il ne conserve pas la cohorte ; il la résume en une règle d'où la cohorte peut être régénérée. ***La cohorte cesse d'être un ensemble de lignes ; elle devient une équation.***

Cette opération porte un nom implicite que l'on peut décrire sans le théoriser. Une cohorte contient trois choses mêlées :

- De l'information,
- De la redondance,
- Et du bruit.

La redondance, ce sont les régularités répétées d'un patient à l'autre. Le bruit, c'est l'idiosyncrasie propre à chaque trace, ce qui n'appartient qu'à un individu et à personne d'autre. L'information utile, c'est la structure de population : les dépendances, les corrélations, les formes de trajectoire. Le modèle ne retient que cette information ; il jette le bruit, et il compacte la redondance. On comprend alors, et seulement alors, pourquoi l'individu importe peu dans cette opération : il était, pour une large part, le bruit que la compression écarte, pas le signal qu'elle conserve.

Une précision de rigueur, sans laquelle l'image trahirait. Il s'agit d'une compression d'information, pas nécessairement d'une économie de paramètres : certains modèles profonds comptent plus de paramètres que la cohorte ne contient de valeurs. La compression est celle de la structure pertinente, pas celle du décompte brut. Dire que la cohorte « devient une équation » est une compression rhétorique de cette idée, pas une affirmation littérale sur le nombre de termes. C'est le levé qui devient carte : non pas plus léger en encre, mais sélectif dans ce qu'il retient.

Une représentation, pas des patients

Une fois admis que l'objet est la distribution, le générateur rejoint une famille bien établie.

Un grand modèle de langage ne stocke pas des phrases ; il apprend une représentation de la langue d'où des phrases peuvent être produites.

Un modèle de diffusion ne stocke pas des images ; il apprend une représentation d'où des images peuvent émerger.

Il n'y a aucune raison qu'une population clinique échappe à cette logique : un générateur de population est un modèle de représentation des populations, et le mouvement est exactement le même.

Cette représentation a une géométrie. Plutôt que de la nommer par son implémentation, je la nommerai par sa fonction : c'est une *géométrie probabiliste*, l'espace des configurations de patients que la distribution rend plausibles, avec leurs densités relatives.

Les modèles profonds implémentent souvent cette géométrie sous la forme d'un espace latent ; mais le concept ne dépend pas de cette implémentation, et le confondre avec elle reviendrait à confondre la carte avec la technique d'impression.

Dans cette géométrie, un patient synthétique n'est pas un individu inventé : c'est un point que l'on projette, une position que l'on lit sur la carte. On ne demande pas à ce point d'exister ; on lui demande d'être au bon endroit.

Les zones blanches de la carte

Toute carte a des zones blanches, et c'est là, exactement, que se gagne ou se perd la rigueur. Tant que l'on lit la carte dans une région densément levée, elle est fiable, parce qu'elle interpole entre des mesures réelles.

Aux marges, là où le géomètre n'est jamais passé, la carte n'est plus une mesure du territoire ; elle est une conjecture du cartographe. *Une carte n'est fiable que là où elle a été levée.* La distinction est nette et il faut la tenir : interpoler dans le support relevé n'est pas extrapoler dans la zone blanche.

Trois dangers vivent dans ces zones blanches, et la métaphore les nomme tous.

- Le premier est le biais de levé : ce qui était sur-représenté à l'arpentage le sera sur la carte, avec d'autant plus d'aplomb que le tracé est net ; la littérature documente que les modèles génératifs profonds amplifient les biais de leurs données et produisent alors des populations déséquilibrées et non représentatives [1].
- Le deuxième est la zone non relevée : une région plausible mais peu observée, où le modèle extrapole sans qu'aucune donnée ne le contraigne.

- Le troisième est le rare : *un générateur n'invente pas le rare, il recopie le peu qu'il en a vu*, et propage les particularités de cette poignée comme si elles décrivaient la population. Une carte ne révèle pas une cité enfouie sous un seul tesson ; un modèle ne révèle pas une sous-population sous une douzaine de cas.

L'interface : la donnée devient un itinéraire

Si le modèle est la carte, alors une cohorte synthétique n'en est que l'usage : l'itinéraire qu'un humain demande à la carte de tracer. Et cet usage change de nature. La donnée devient conversationnelle.

Hier, on interrogeait un entrepôt en SQL, et chaque question exigeait sa requête. Demain, on dialogue avec un modèle de population. « Montre-moi des patients comparables à celui-ci, mais sans insuffisance rénale » ne décrit pas une requête sur une table ; cela décrit un itinéraire dans un espace. Le déplacement est de même nature que celui des moteurs de recherche passant de l'index au modèle de langage : on ne cherche plus une ligne, on navigue dans une représentation.

Il faut résister à l'euphorie que cette fluidité inspire, car elle dissimule l'inférence sous la conversation. ***Chaque itinéraire reste un calcul***. « Sans insuffisance rénale » n'est légitime que si la carte a correctement relevé la dépendance entre cette condition et le reste ; sinon l'itinéraire traverse une zone blanche en donnant l'illusion d'une route. L'interface conversationnelle ne supprime pas l'inférence ; elle la glisse sous la surface, ce qui la rend plus facile à oublier, donc plus dangereuse à ne pas valider.

La validation démontre la substituabilité, pas la ressemblance

Valider un modèle de population, c'est mesurer jusqu'où sa carte reste fiable pour un usage donné. Et le critère décisif n'est pas la ressemblance. *Une représentation ne se valide pas sur sa fidélité d'apparence ; elle se valide sur sa substituabilité opérationnelle.* Deux niveaux doivent être distingués, et leur confusion explique la plupart des malentendus.

- La fidélité statistique mesure à quel point la distribution apprise épouse la distribution réelle : distance de Wasserstein faible, log-rank non significatif entre courbes de survie, pMSE proche de l'indiscernabilité [2][3].
- La substituabilité opérationnelle mesure autre chose : un modèle entraîné sur les sorties du générateur, puis testé sur des données réelles, préserve-t-il les conclusions ? C'est le protocole Train-on-Synthetic, Test-on-Real, qui ne demande pas si la carte ressemble au terrain, mais si l'on arrive à destination en s'y fiant.

La distinction n'est pas un raffinement. Une carte peut être fidèle dans ses grandes lignes et trompeuse sur l'itinéraire précis, parce qu'elle a manqué une dépendance que les marges ne montrent pas.

C'est pourquoi la validation se fait toujours contre des mesures réelles externes, jamais contre le seul jugement du modèle sur lui-même, comme l'établissent les travaux qui confrontent les générateurs de référence à des indicateurs cliniques calculés sur données réelles [4].

La génération conditionnelle vers des sous-populations sous-représentées le confirme : augmenter un jeu de données par des cohortes conditionnellement générées peut améliorer la généralisation à ces sous-populations [5], mais à proportion exacte de la fidélité avec laquelle elles avaient été relevées.

La condition n'est pas un détail ; elle est le sujet. Une représentation sans protocole de validation contre données réelles n'est pas un modèle de population : c'est une assertion cartographiée.

TweenMe®, comme terrain d'implémentation, ne déroge pas à cette exigence ; il la rend opérationnelle, et c'est à ce niveau que la chose se juge, pas au niveau de la promesse.

Le principe de représentation clinique

Tout ce qui précède se laisse condenser en un principe, et c'est ce principe, plus que la technologie, qui mérite d'être retenu.

Principe de représentation clinique. Une cohorte n'est jamais étudiée pour elle-même. Sa seule fonction est de fournir une représentation de la distribution qui l'a engendrée, suffisamment fidèle pour soutenir une famille de décisions. Les modèles génératifs ne modifient pas ce principe ; ils rendent cette représentation explicite et manipulable.

L'intérêt de cet énoncé est qu'il ne dépend de rien de ce dont on a parlé. Il ne dépend ni des réseaux de neurones, ni des données synthétiques, ni même de la médecine. Il décrit une propriété générale de l'inférence statistique : l'objet scientifique est la représentation de la distribution, et non les observations qui ont permis de l'estimer. Le mot décisif est *suffisamment* : la fidélité n'est jamais absolue, elle est relative à la famille de décisions visée. C'est exactement le critère de substituabilité, remonté au rang de principe. Les trois premiers âges de la donnée appliquaient ce principe sans le formuler, en laissant la représentation implicite dans la tête du biostatisticien. Le quatrième le rend explicite, et avec lui rend explicites ses conditions de validité, ce qui est un progrès et une exposition à la fois.

La médecine quitte les objets

Reste à refermer là où l'article rejoint une trajectoire plus large que lui. J'ai soutenu ailleurs que, dans un système régulé, la preuve n'est pas un objet que l'on détient mais une relation que l'on établit, valide sous conditions explicitées et nulle au-dehors. Le présent article en est le pendant : la représentation, elle non plus, n'est pas un objet. Trois déplacements le disent, et ils relèvent du même mouvement.

Une cohorte n'est pas un objet ; elle est une observation échantillonnée.

Une population synthétique n'est pas un objet ; elle est une représentation.

Une preuve n'est pas un objet ; elle est une relation.

À chaque fois, la médecine quitte la chose pour ce qui la relie à la distribution dont elle procède. La représentation est une propriété émergente, au même titre que la preuve : elle ne réside dans aucun patient, réel ou synthétique, mais dans la relation qu'une carte entretient avec un territoire qu'elle ne contiendra jamais.

La quatrième génération de la donnée clinique ne consiste donc pas à fabriquer des patients qui n'existent pas. Elle consiste à reconnaître que l'objet n'a jamais été le patient, et à manipuler enfin la distribution comme une carte plutôt que comme une archive. C'est un gain considérable, et c'est pour cette raison même qu'il faut en marquer la limite plutôt que de la taire. La carte devient plus riche ; le territoire ne change pas. Le modèle étend ce que l'on peut demander à une cohorte ; il n'étend pas ce que cette cohorte savait. Confondre les deux, c'est précisément l'erreur que la validation existe pour empêcher.

Références

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. La preuve est une promesse conditionnelle: vers une théorie relationnelle de la validation des preuves computationnelles et des populations synthétiques. 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.