Représenter n'est pas reproduire

Projection, validation et domaine de substituabilité des cohortes synthétiques.

Du principe à l'instrument

L'article précédent défendait une thèse contre-intuitive : dans la quatrième génération de la donnée clinique, l'objet scientifique n'est plus le patient observé mais la distribution de population que les observations permettent d'estimer. Une cohorte synthétique ne vaut donc jamais parce qu'elle ressemble à des patients réels. Elle vaut si elle préserve les propriétés nécessaires aux analyses pour lesquelles elle a été construite. Cette thèse restait abstraite. Le présent article quitte le terrain doctrinal pour celui de l'instrumentation, et il s'y tient.

Une proposition le résume. Un générateur tabulaire est un instrument de projection d'un levé clinique imparfait. Sa validation consiste à délimiter les zones où cette projection demeure substituable à la distribution observée, et celles où elle devient spéculative.

Le mot projection est le pivot, et il faut le fixer avant qu'il ne se déforme. *Une projection est une transformation qui conserve certaines propriétés d'un objet, en abandonne d'autres, et réorganise le reste.* Les trois opérations ne sont pas équivalentes : conserver n'est pas réorganiser, et une représentation latente peut faire apparaître des régularités que les observations ne montraient pas. Choisir une architecture générative, c'est donc choisir ce que l'on conserve, ce que l'on perd et ce que l'on réorganise. Rien dans cet article n'emploie projection en un autre sens.

Une borne, posée d'emblée. Le présent article s'arrête à la construction et à la validation d'une cohorte synthétique. La modélisation prédictive, la calibration, la décision clinique et l'évaluation médico-économique relèvent d'une couche distincte ; elles supposent une représentation valide sans en constituer le critère, et ne sont pas traitées ici.

Une clarification s'impose avant d'aller plus loin. Le réel n'est jamais directement observé. Les observations permettent d'en estimer une distribution. Le générateur construit une représentation de cette distribution, et non du réel lui-même. La validation ne porte donc jamais sur le réel, mais sur l'adéquation de cette représentation à un domaine d'usage déclaré. Tout le reste de l'article découle de cette hiérarchie.

Reste à justifier que cet objet mérite un traitement séparé, et c'est le meilleur point d'appui du raisonnement. La génération tabulaire n'est pas un cas particulier de la génération d'images. Une image possède une géométrie donnée par le capteur : une grille

de pixels, une invariance par translation, un voisinage naturel entre points proches. Une table clinique n'a rien de tout cela. Ses variables sont hétérogènes, ses marginales souvent multimodales, ses catégories fortement déséquilibrées, et aucune métrique d'adjacence ne préexiste à l'analyse. *Synthétiser une image exploite une géométrie donnée ; synthétiser une table clinique exige d'abord de la fabriquer.* La conséquence est immédiate et mérite d'être dite sans détour : un générateur tabulaire apprend autant une métrique implicite qu'une distribution. La littérature d'évaluation l'a reconnu en constatant que les métriques taillées pour les images diagnostiquent mal les défaillances des générateurs ailleurs (Alaa et al., ICML 2022). C'est cette asymétrie qui interdit de traiter le tabulaire comme une variante du visuel.

Trois mots, et un instrument faillible

Trois définitions fixent le cadre. Un *générateur tabulaire* apprend, à partir d'un échantillon fini, une approximation de la distribution jointe d'un ensemble de variables hétérogènes, puis produit de nouvelles observations compatibles avec elle. Il ne mémorise pas une cohorte ; il apprend une règle. Une *cohorte synthétique* est l'échantillon tiré dans cette distribution estimée ; elle n'est pas une copie de la cohorte d'entraînement, pas davantage qu'un second lancer de dé n'est la copie du premier. L'*espace des plausibles* est l'ensemble des configurations auxquelles le modèle attribue une probabilité non négligeable : le support de la distribution estimée.

Cette dernière notion appelle une correction. L'espace des plausibles n'est pas une propriété intrinsèque de la population ; il est l'estimation, à un instant donné, du support que les données autorisent à inférer. Le support n'est pas découvert, il est estimé. Et l'estimation hérite des conditions du recueil : critères d'inclusion, pratiques, codage, recrutement, période.

Ici se loge un angle mort que la plupart des manuscrits laissent sous le tapis. On traite le levé clinique comme un échantillon imparfait. Il est aussi un instrument imparfait. Les erreurs de codage, la variabilité inter-observateur, les changements de nomenclature, l'évolution des pratiques et les biais des dispositifs de mesure modifient la distribution avant toute modélisation. *Le générateur ne projette pas seulement un échantillon fini ; il projette les traces laissées par un système de mesure lui-même faillible.* La distinction n'est pas rhétorique : un échantillon imparfait se corrige par plus de données, un instrument imparfait ne se corrige que par la connaissance de ses biais. Un cas particulier le rend tangible : le support estimé dépend autant des observations présentes que du mécanisme des absences, et selon que les données manquent de façon complètement aléatoire, conditionnellement aléatoire ou non aléatoire (MCAR, MAR, MNAR), le territoire projeté n'est pas le même.

Le support a une étendue ; il a aussi une résolution, et ce second axe gouvernera tout le reste de l'article. L'étendue dit où la carte existe ; la résolution dit avec quelle finesse elle

informe. Et cette finesse n'est pas une affaire de densité mais d'information locale : mille patients presque identiques résolvent mal une région, quand trente patients suffisamment variés la résolvent bien. *La résolution n'est pas une propriété spatiale mais informationnelle*, fonction de la variance, de l'entropie et de la séparabilité des observations, non de leur simple nombre. De là une formule qui tient tout le reste : *une région peut être couverte sans être suffisamment résolue*. Un profil BRAF V600E peut figurer dans une cohorte sans y déposer assez d'information pour soutenir une décision conditionnelle. La couverture rassure ; la résolution décide. On retrouvera cette paire à chaque étape : dans les métriques, dans l'extrapolation, dans la génération guidée, dans la validation.

Chaque architecture est un biais inductif

Les familles d'architectures sont souvent présentées comme une succession de progrès. La lecture est trompeuse : toutes répondent à la même question à partir d'hypothèses différentes. Cette partie est illustrative, pas thèse. Une seule phrase doit en rester : chaque architecture est un biais inductif différent, c'est-à-dire un choix différent de ce que la projection conserve, abandonne et réorganise. Et ce biais n'est pas un défaut à corriger. *Le biais inductif n'est pas une faiblesse de l'algorithme ; c'est la formalisation mathématique des hypothèses que la projection impose lorsque les observations deviennent insuffisantes*. Un même levé admet d'ailleurs plusieurs représentations compatibles ; le choix d'architecture sélectionne donc une hypothèse d'identification parmi celles que l'échantillon fini sous-détermine.

Les *réseaux bayésiens* factorisent explicitement la jointe selon un graphe acyclique, $P(X) = \prod_i P(X_i \mid \text{parents}(X_i))$: ils conservent des dépendances auditables, ils abandonnent les interactions fortement non linéaires à grande échelle. Les *copules* dissocient, par le théorème de Sklar, les marginales de la dépendance, $F = C(F_1, \ldots, F_d)$: robustes sur petits jeux, lisibles, au prix d'une forme de dépendance imposée.

CTGAN (Xu et al., NeurIPS 2019) résout deux problèmes propres au tabulaire. Contre la multimodalité des variables continues, une normalisation par mélange gaussien variationnel représente chaque valeur par un indicateur de mode et un scalaire normalisé, $\alpha = (c - \eta_k)/(4\,\phi_k)$. Contre le déséquilibre catégoriel, un vecteur conditionnel et un échantillonnage d'entraînement imposent une exposition régulière aux modalités rares, sous une perte de Wasserstein à pénalité de gradient (Gulrajani et al. 2017). *CTGAN n'apprend pas à dessiner des patients ; il apprend à ne pas laisser la moyenne effacer les modes que les sous-populations contiennent*. TVAE optimise une borne inférieure de vraisemblance, $\text{ELBO} = E_q[\log p(x|z)] - \mathrm{KL}(q(z|x) \,\|\, p(z))$, plus stable mais plus lissée : là où le GAN trompe un juge, le VAE comprime puis reconstruit. TabDDPM (Kotelnikov et al., ICML 2023) bruite puis débruite, gaussien sur les numériques, multinomial sur les catégorielles (Hoogeboom et al., 2021) : il n'apprend pas la donnée, il apprend la distance

qui la sépare du bruit. Les générateurs causaux enfin, tel DECAF (van Breugel et al., 2021), préservent les mécanismes et non la seule jointe, car deux générateurs indiscernables sur les distributions peuvent diverger sur l'effet d'une intervention : préserver P(X) ne préserve pas P(Y | do(X)).

Ce que ces familles apprennent n'est d'ailleurs pas le même objet. Un GAN apprend une frontière discriminante, une diffusion un champ de score, un autoencodeur variationnel une variété latente, un réseau bayésien une factorisation conditionnelle, un transformeur un jeu de dépendances contextuelles. Elles convergent pourtant vers une même cible, une distribution. *Ces architectures diffèrent donc moins par ce qu'elles représentent que par l'opérateur avec lequel elles construisent la représentation.* C'est cet opérateur, et non la distribution visée, que le choix d'architecture sélectionne.

Un exemple stylisé fixe l'idée, et il vaut comme illustration, non comme mesure. Soit une même cohorte et une même tâche, confiées à trois générateurs. Le GAN restitue fidèlement les marginales mais peut relâcher la structure jointe d'ordre élevé. La diffusion capture mieux cette jointe, ce que les synthèses récentes confirment. Le générateur causal préserve ce que les deux autres ignorent, l'effet d'une intervention. Aucun n'est meilleur dans l'absolu ; chacun projette selon son opérateur, et le bon choix est celui dont les pertes tombent hors de la tâche visée.

Ces familles ne sont pas des lignées pures, et c'est leur recombinaison qui le montre le mieux. On peut greffer un transformeur ou des blocs d'attention sur un GAN tabulaire : l'opération ne produit pas un meilleur générateur, elle change le biais inductif pour capter les dépendances conditionnelles non locales entre colonnes, là où la normalisation modale et le vecteur conditionnel restaient muets. C'est la voie de TT-GAN en santé (Ryu et al., Sci Rep 2025) qui rapportent sur ce point un avantage mesuré sur CTGAN et copula GAN. On peut de même hybrider diffusion et transformeur, en remplaçant le perceptron débruiteur par un encodeur-décodeur à attention conditionnelle (MTabGen, Villaizán-Vallelado et al., ACM TKDD 2025), ou diffuser dans l'espace latent continu d'un autoencodeur avant de décoder vers la table (TabSyn, Zhang et al., ICLR 2024). Les schémas accélérateurs, où un GAN approxime un débruitage en moins d'étapes pour réduire le coût d'inférence, ou régularisateurs, où un discriminateur pénalise les sorties d'un modèle de diffusion, relèvent du même principe.

Le piège serait d'en conclure que l'hybride est par nature supérieur. Il ne l'est pas. *Une architecture peut être hybridée, mais l'hybridation n'a de valeur scientifique que si chaque composant répond à un mode de défaillance identifié de la projection : une dépendance mal apprise, un mode rare effacé, un support extrapolé, une confidentialité dégradée, un coût d'inférence excessif.* Le mode collapse appelle la diffusion ou PacGAN (Lin et al. 2018) ; les dépendances conditionnelles non locales, l'attention ; l'hétérogénéité des colonnes, des embeddings typés et une normalisation modale ; la rareté conditionnelle, un conditionnement contrôlé ; le coût d'inférence, une distillation ou un accélérateur

adverse ; la confidentialité, la régularisation différentielle et l'audit d'appartenance ; la causalité, une structure explicite et non une couche de plus. Greffer de l'attention sans défaillance à corriger n'ajoute pas d'information : sur les petits effectifs cliniques, l'attention apprend volontiers le bruit, renforce la mémorisation, et produit une matrice qui impressionne les comités sans rien expliquer.

D'où une posture, et elle prolonge la thèse au lieu de s'en écarter. On n'élit pas une architecture reine ; on orchestre une famille de générateurs, chacun rapporté à un profil de données, à un domaine d'applicabilité, à un compromis fidélité-utilité-confidentialité et à un protocole de validation. *Le choix d'un générateur n'est pas une décision de performance brute ; c'est une décision de projection sous contraintes*.

Reste l'objection la plus sérieuse à cette posture, et il faut la formuler à pleine force. Si les modèles de diffusion dominent désormais la plupart des classements, comme le rapportent les synthèses récentes sur un large éventail de benchmarks tabulaires (Diffusion Models for Tabular Data, arXiv:2502.17119, 2025 ; Shi et al. 2025), pourquoi traiter l'architecture comme une décision de second ordre ? La réponse ne conteste pas les benchmarks. Ils mesurent de mieux en mieux plusieurs dimensions, fidélité, diversité, confidentialité, utilité. Ils agrègent néanmoins ces performances sur des régions moyennes, et restent mal équipés pour qualifier une substituabilité conditionnelle dans une marge clinique rare. La domination de la diffusion est réelle au centre et indémontrée aux bords. *Le leaderboard est vrai et hors-sujet*. Ce qui se décide en clinique se décide précisément là où il se tait.

Mesurer une représentation

La question dominante est mal posée : ce générateur est-il meilleur qu'un autre ? Elle suppose une mesure unique de qualité, qui n'existe pas. La qualité d'une cohorte synthétique est multidimensionnelle, et ses dimensions ne sont pas indépendantes : améliorer l'une déplace souvent les autres. Valider ne consiste pas à maximiser un score, mais à caractériser un compromis.

La *fidélité* s'examine à trois échelles. Les marginales, variable par variable, par des tests de type Kolmogorov-Smirnov ou des divergences adaptées aux catégorielles. Les dépendances, par les corrélations, l'information mutuelle, et les contraintes logiques du domaine, une grossesse chez un homme ou un décès antérieur à la naissance étant une violation de structure et non une mauvaise corrélation. La jointe, enfin, par la distance de Wasserstein, la Maximum Mean Discrepancy ou le pMSE, qui entraîne un classifieur à séparer réel et synthétique : plus la distinction est difficile, plus les distributions sont proches. Deux jeux peuvent partager des marginales presque identiques et décrire des populations différentes si les dépendances ont disparu ; préserver chaque colonne ne garantit pas de préserver la cohorte.

Une avancée récente rend ce niveau lisible à l'échelle de l'échantillon et donne à la notion de zones blanches son ancrage métrique. Le triptyque prolonge la lignée précision-rappel pour modèles génératifs (Kynkäänniemi et al. 2019) et la spécialise au niveau de l'échantillon (Alaa et al., ICML 2022). Le triptyque sépare trois échecs que les distances agrégées confondent. L'α-Precision mesure la part d'observations synthétiques tombant hors du support réel : la zone non relevée, où le modèle extrapole. Le β-Recall mesure la part de régions réelles non couvertes : le biais de levé transmis. L'Authenticité mesure si le générateur a généralisé ou recopié : exactement le piège du rare, la poignée d'observations propagée au lieu d'une structure estimée. Trois axes nomment, terme à terme, les trois manières de tromper. Et la résolution s'y lit directement : un β-Recall acceptable au grain grossier peut masquer une région couverte mais sous-résolue, où la décision réclamait un grain fin.

L'*utilité* est une autre dimension, et l'une des observations les plus importantes de la littérature récente. Une représentation peut être remarquablement fidèle et dégrader fortement les modèles entraînés sur elle ; une légère perte de fidélité peut au contraire rester sans conséquence sur l'analyse réellement menée. La raison est structurelle : la fidélité conserve P(X), les modèles prédictifs exploitent P(Y|X), et ces deux objets n'évoluent pas ensemble. La fidélité est nécessaire à l'utilité ; elle n'en est jamais la preuve. Le protocole de référence est le Train-on-Synthetic, Test-on-Real (Esteban et al. 2017), dont le statut exact sera précisé plus loin, car il ne mesure pas une qualité intrinsèque mais une relation.

La *confidentialité* est une troisième dimension, et répond à une autre question : le générateur a-t-il appris une structure ou mémorisé des individus ? La distance au plus proche réel, les attaques par inférence d'appartenance, les formulations de confidentialité différentielle estiment le risque qu'une observation réelle soit retrouvée. Ces mesures ne sont pas des récompenses ; elles quantifient un risque.

Présentées comme trois objectifs indépendants, ces dimensions induisent en erreur. Elles forment un compromis de Pareto : renforcer la confidentialité ajoute du bruit, ce qui peut réduire la fidélité, ce qui peut dégrader l'utilité ; viser la fidélité maximale pousse à mémoriser, ce qui augmente le risque de divulgation. *Il n'existe pas de générateur optimal au sens absolu ; il n'existe que des points de fonctionnement adaptés à un usage*. L'adéquation à l'usage n'est donc pas la dernière question mais la première : une cohorte destinée à entraîner un classifieur diagnostique n'impose pas les mêmes exigences qu'un bras de contrôle synthétique ou qu'un partage inter-établissements. La validation ne demande pas si une cohorte est bonne ; elle demande si cette représentation est suffisamment fidèle, utile et protectrice pour cette famille d'analyses, dans ce domaine déclaré.

Interpoler n'est pas extrapoler

Aucun générateur n'observe la population qu'il prétend représenter ; il observe un échantillon fini et en construit une approximation. Son domaine naturel est l'interpolation à l'intérieur de la région documentée. Dans une zone densément couverte, il combine des dépendances abondamment contraintes, et les estimations sont robustes. À mesure que la densité diminue, ces contraintes s'effacent et les hypothèses de l'architecture prennent le relais. *Plus le modèle s'éloigne du territoire relevé, moins il estime la distribution observée et plus il exprime ses propres hypothèses*. La transition est continue ; elle n'a pas de frontière nette, et toute validation consiste précisément à déterminer où elle devient incompatible avec l'usage revendiqué.

C'est ici que résolution et extrapolation se distinguent, et la confusion entre les deux est coûteuse. L'extrapolation est une défaillance de couverture : on sort du support. La sous-résolution est une défaillance de grain : on reste dans le support mais sans densité suffisante pour le contraindre. Les deux produisent des observations plausibles d'apparence, et c'est précisément ce qui les rend dangereuses. L'interface conversationnelle des outils modernes accentue le risque : la fluidité dissimule la transition entre estimation et extrapolation, et plus la génération paraît naturelle, plus il devient facile d'oublier qu'une région n'a jamais été véritablement observée.

D'où une mise au point. Les générateurs produisent des patients dits nouveaux, ce qui est vrai au sens combinatoire et faux au sens statistique. Les lignes produites sont de nouvelles combinaisons compatibles avec la distribution apprise ; elles ne sont pas la découverte de nouvelles régions. La nouveauté porte sur les réalisations, jamais sur la structure. Et si le recrutement est biaisé ou les dépendances datées, le générateur interpolera fidèlement un levé imparfait, produit d'un instrument imparfait. Aucune architecture ne crée d'information : elle redistribue l'information disponible. C'est une conséquence directe de la théorie de l'information (Shannon 1948 ; Cover & Thomas 2006), pas une limite des architectures actuelles. La nuance est importante, car certains modèles rendent effectivement exploitable une information auparavant inaccessible : ils ne l'inventent pas, ils la réorganisent. La qualité de la représentation dépend donc autant de la qualité du relevé que de la sophistication de l'algorithme.

Fidélité n'est pas substituabilité

Voici le résultat de l'article, et tout ce qui précède y conduit. Une objection paraît naturelle : si la cohorte reproduit fidèlement les distributions, pourquoi ne remplacerait-elle pas la cohorte réelle ? L'intuition est séduisante et fausse. Elle confond la qualité d'une représentation et la validité d'un raisonnement construit sur elle. *La fidélité est une propriété descriptive ; la substituabilité, une propriété opérationnelle*. Les deux sont liées sans être équivalentes.

Deux cartes routières l'illustrent. La première reproduit parfaitement le relief mais omet des routes secondaires. La seconde simplifie le relief mais conserve tout le réseau. La première ressemble davantage au territoire ; la seconde conduit plus sûrement à destination. La fidélité demande si la représentation ressemble aux observations. La substituabilité demande si les analyses qu'on y mène conduisent aux mêmes conclusions. On peut répondre excellemment à la première sans satisfaire la seconde, parce que la fidélité approxime la conservation de P(X) quand les conclusions dépendent de P(Y|X) ; une faible altération de certaines dépendances peut être négligeable sur les distributions globales et lourde sur les performances. Il serait surprenant que fidélité et utilité soient fortement corrélées.

D'où le statut exact du Train-on-Synthetic, Test-on-Real. Son intérêt n'est pas seulement de comparer apprentissage réel et synthétique ; il est de mesurer une propriété relationnelle. Le TSTR n'attribue pas une note au générateur. Il qualifie une représentation relativement à une famille de tâches, et son résultat dépend simultanément de quatre éléments : le générateur, la tâche, le modèle aval et le protocole d'évaluation. Modifier l'un d'eux modifie le résultat sans que le générateur ait changé, et deux générateurs peuvent inverser leur classement selon qu'on entraîne un modèle de survie, un classifieur ou un estimateur causal. *La substituabilité ne caractérise jamais un générateur ; elle caractérise une représentation relativement à une famille de décisions*. Une cohorte construite pour un modèle pronostique peut préserver les structures de risque nécessaires à cette tâche et rester insuffisante pour estimer un effet de traitement, sans contradiction, parce que les tâches n'exploitent pas les mêmes propriétés de la représentation. Cette thèse est falsifiable, et il faut le dire pour qu'elle reste scientifique : une corrélation forte et stable entre les métriques de fidélité et la substituabilité opérationnelle, mesurée à travers des tâches variées, la réfuterait. C'est la même logique que les domaines d'applicabilité des modèles QSAR en toxicologie prédictive : un modèle n'est pas valide parce qu'il prédit bien en général, mais parce que son domaine de validité est connu, documenté et respecté.

Tout cet article se condense alors en un énoncé stable.

Principe de projection clinique. Toute cohorte est un échantillon. Toute projection applique un biais inductif. Toute cohorte synthétique est une estimation issue de cette projection. Toute estimation possède un domaine. Toute validation est conditionnelle à ce domaine. Toute substituabilité disparaît hors de ce domaine.

Ce principe ne répète pas le principe de représentation clinique de l'article parent ; il le prolonge. Le premier établissait que l'objet scientifique est la distribution et non l'individu observé. Celui-ci ajoute que cette distribution, une fois projetée par un générateur, n'est substituable que dans les bornes d'un domaine, et nulle au-dehors. L'un porte sur ce qu'on étudie, l'autre sur ce qu'on peut en faire.

La validation cesse ainsi d'être un certificat. Elle devient une preuve conditionnelle, valable pour un domaine explicitement déclaré, et dont la restriction même fait la valeur scientifique.

La validation est un processus

Une preuve conditionnelle possède quatre composantes : elle vaut pour une population, une période, une famille de tâches, un protocole. Modifier l'une d'elles modifie l'objet validé. Une cohorte validée sur un cancer du sein ne devient pas valide pour un cancer bronchique ; une validation sur des données de 2015 ne survit pas nécessairement à l'introduction d'une nouvelle stratégie thérapeutique ; une représentation substituable pour un modèle pronostique ne l'est pas nécessairement pour un effet causal. La validation n'est jamais transportable sans justification.

La raison profonde est double, et la seconde est rarement nommée. Non seulement la population dérive, mais l'instrument de mesure dérive avec elle. La médecine n'est pas stationnaire : les traitements évoluent, les critères diagnostiques sont révisés, les nomenclatures et les codages se renouvellent. Le levé d'aujourd'hui n'est pas le levé d'hier, et une représentation parfaitement valide à sa date perd progressivement cette propriété sans que le générateur ait jamais cessé de fonctionner. La dérive n'est pas un défaut de validation ; elle est la conséquence normale d'un monde qui change, mesuré par un instrument qui change aussi.

La conséquence est opérationnelle, et l'industrie la sous-estime régulièrement. *Le coût réel d'une plateforme générative n'est pas le coût de production d'une cohorte ; c'est le coût de maintien de son domaine de validité.* Surveiller, revalider, recalibrer, versionner, monitorer : ces opérations, et non l'entraînement initial, décident de la viabilité d'un dispositif en conditions réelles. La logique est connue ailleurs, des domaines d'applicabilité QSAR aux procédures de gestion des modifications des dispositifs médicaux logiciels et au monitoring de dérive des modèles déployés. TweenMe constitue une implémentation de cette doctrine : le produit ne cherche pas à optimiser un générateur, il cherche à industrialiser le maintien du domaine de validité.

Aucune métrique ne franchit cette frontière. Ni une distance de Wasserstein, ni un pMSE, ni un score TSTR, ni un niveau de confidentialité différentielle ne répond à une question qui n'est pas la sienne. Une excellente fidélité ne garantit pas la pertinence d'une décision ; une excellente confidentialité ne garantit pas l'utilité ; un excellent TSTR ne garantit pas une analyse causale. Chercher un score unique résumant la qualité d'une cohorte reviendrait à demander à un seul biomarqueur l'état physiologique complet d'un patient. L'information utile est dans la combinaison des indicateurs, pas dans leur réduction.

Un générateur tabulaire ne fabrique donc pas des patients. Il projette un levé faillible en une représentation interrogeable, et toute sa valeur démontrable tient à la délimitation

honnête des zones où cette projection demeure substituable. La fidélité décrit la proximité de deux distributions ; la substituabilité démontre qu'elles conduisent aux mêmes conclusions pour une famille de tâches. Le reste est affaire de domaine.

Reste une question que cet article n'ouvre pas, et qu'un autre devra porter. Si la distribution observée n'était elle-même que la projection statistique d'une géométrie de contraintes que le générateur apprend d'abord, alors le levé ne serait qu'une ombre, et l'objet véritable se tiendrait un cran plus loin. Ce déplacement, des distributions vers les géométries, est la thèse d'un article à venir, pas la conclusion de celui-ci.

Références

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.