Un jumeau numérique en santé ne se valide pas en miroir

La déployabilité d'un jumeau numérique clinique ne découle ni de son réalisme ni des seules performances de son générateur. Elle repose sur une architecture démontrant que, pour une tâche et un domaine déclarés, les décisions produites restent substituables au réel, et sur la capacité du système à refuser quand il en sort.

Les jumeaux numériques arrivent dans la santé régulée par la porte de la rareté. Les données cliniques sont coûteuses à produire, difficiles à partager, parfois inexistantes pour certaines populations ou certains événements rares. Les cohortes synthétiques promettent d'élargir cet espace : entraîner des modèles, explorer des scénarios thérapeutiques, tester des protocoles, construire des bras de contrôle virtuels ou simuler des décisions impossibles à expérimenter directement chez l'humain.

La promesse est crédible. Le critère par lequel un COMEX la défendra devant un régulateur l'est beaucoup moins, parce qu'il porte encore trop souvent sur la mauvaise propriété.

La communauté méthodologique a, pour l'essentiel, déjà tranché : la fidélité statistique d'une cohorte synthétique ne mesure pas son utilité pour une tâche. Le déplacement persiste ailleurs, à la couche qui compte vraiment pour un dossier industriel : celle du déploiement, de la gouvernance et de la responsabilité réglementaire, où l'on continue de juger un jumeau à son réalisme apparent. C'est là que se loge le piège, et c'est là que cette note se situe.

Le piège du réalisme

La validation d'un jumeau repose encore fréquemment, dans les arbitrages de déploiement, sur des mesures de ressemblance : comparaison des distributions, inspection par des experts, capacité d'un discriminateur à distinguer données réelles et synthétiques. Ces mesures sont localement utiles et deviennent trompeuses dès qu'on les promeut en critère général de validation.

Car la question pertinente n'est jamais : « Le jumeau ressemble-t-il au réel ? »

Elle est toujours : « Pour quel usage souhaite-t-on remplacer le réel, et sous quelles garanties ? »

Un jumeau n'est jamais déployé "en général". Il est toujours mobilisé pour une finalité précise : entraîner un modèle, explorer un scénario thérapeutique, estimer une stratégie de recrutement, construire un bras de contrôle, calibrer un seuil ou tester une politique clinique. Chaque usage possède sa propre notion de fidélité.

Dans certains cas, préserver fidèlement une distribution statistique constitue précisément l'objectif. Dans d'autres, ce qui importe est la conservation d'une frontière de décision, d'une calibration ou d'une capacité prédictive. Le débat n'oppose donc pas la ressemblance à son absence. Il oppose deux *niveaux* de ressemblance :

1. La ressemblance distributionnelle globale, souvent mesurée,
2. Et la ressemblance de la structure pertinente pour la tâche, rarement isolée.

Le réalisme n'est pas un mauvais indicateur. C'est un indicateur spécialisé, pertinent uniquement lorsque le niveau qu'il capture correspond à la tâche revendiquée.

Du portrait statistique à l'instrument décisionnel

Tous les jumeaux commencent comme des portraits statistiques. Certains deviennent ensuite des instruments de décision. La différence est structurante.

Un portrait cherche à représenter une population. Un instrument cherche à produire une décision suffisamment fidèle pour remplacer, dans un contexte donné, une décision qui aurait été prise à partir de données réelles.

À partir du moment où un jumeau intervient dans une décision clinique, réglementaire ou industrielle, la question change de nature. La validation ne porte plus principalement sur la qualité de l'imitation. Elle porte sur la robustesse de la substitution. Ce n'est plus le portrait qui est évalué, mais la boucle décisionnelle dans laquelle il intervient.

Trois preuves deviennent alors nécessaires

1. La première consiste à démontrer l'absence de fuite. La cohorte réelle utilisée pour l'évaluation ne doit jamais avoir contribué à la génération des données synthétiques. Sans séparation stricte entre génération et validation, la performance mesurée peut n'être que le reflet d'une contamination des données.
2. La deuxième consiste à démontrer la substituabilité opérationnelle. La méthodologie Train-on-Synthetic / Test-on-Real, formalisée par Esteban, Hyland et Rätsch (2017) pour les séries temporelles médicales, fournit ici un cadre pertinent : le modèle est entraîné exclusivement sur les données synthétiques puis évalué sur une cohorte réelle indépendante. Mais la seule discrimination ne suffit pas. Une substitution crédible doit préserver les propriétés conditionnant l'usage : discrimination lorsque celle-ci est recherchée, mais également

calibration, bénéfice décisionnel et stabilité lorsque ces dimensions gouvernent la décision clinique.

3. La troisième consiste à déclarer explicitement le domaine d'applicabilité. Ce domaine ne décrit pas seulement une population. Il définit l'espace de validité du jumeau : caractéristiques des patients, modalités de collecte, contexte clinique, période temporelle, pratiques thérapeutiques et conditions techniques sous lesquelles les garanties précédentes restent démontrées. Cette déclaration n'est pas une garantie : c'est une hypothèse réfutable, qui doit pouvoir être mise en défaut par la surveillance et révisée en conséquence.

En dehors de cet espace, un jumeau ne doit pas produire une réponse confiante. Il doit produire un refus. Ce refus n'est pas une propriété spontanée du générateur ; c'est une exigence d'architecture, qui suppose un mécanisme explicite de détection des situations hors domaine. La gouvernance commence précisément là où le système reconnaît qu'il sort de son espace de validité (le port de promotion et la taxonomie du refus, développés ailleurs dans cette série, en fournissent l'ossature).

L'objection de la rareté

Une objection se présente ici, et elle est sérieuse. La preuve de substituabilité exige une cohorte réelle indépendante. Or le jumeau est précisément convoqué là où le réel manque : populations sous-représentées, événements rares, maladies orphelines, situations jamais observées.

Pour ces cas, la validation Train-on-Synthetic / Test-on-Real est indisponible par construction. On ne teste pas sur un réel qui n'existe pas.

Il faut tenir cette objection plutôt que la contourner. Elle impose de distinguer deux régimes.

- Dans le *régime ancré*, une cohorte réelle de validation existe. La substituabilité s'y *démontre*, au sens fort, et c'est à ce régime que la thèse de cette note s'applique pleinement.
- Dans le *régime extrapolé*, aucune vérité terrain n'est disponible dans la zone d'usage. La substituabilité ne s'y démontre pas ; elle se *borne et se surveille*. Sa valeur repose alors sur une hypothèse de généralisation du générateur, que seul le domaine d'applicabilité déclaré encadre, et que la détection hors domaine doit pouvoir invalider en temps réel.

La conséquence est inconfortable et doit être assumée : un jumeau gouvernable refusera parfois de répondre exactement là où on l'espérait le plus. Ce n'est pas un échec de l'architecture. C'est la forme honnête de la gouvernabilité, opposée à la

confiance silencieuse d'un système qui répond partout sans jamais savoir où il cesse d'être valide.

Le générateur ne suffit pas

Cette distinction conduit à une autre confusion fréquente. La déployabilité d'un jumeau ne peut pas être déduite des seules performances de son générateur.

Un excellent générateur peut produire des données présentant des fuites de confidentialité, un effondrement des modes rares, une mauvaise préservation des dépendances pertinentes ou une incapacité à extrapoler hors du domaine observé. Inversement, un générateur dont les données restent facilement distinguables du réel peut néanmoins permettre une excellente substituabilité pour une tâche opérationnelle donnée.

Les propriétés du générateur demeurent importantes. Elles ne suffisent simplement pas à établir que les décisions construites à partir du jumeau resteront valides.

Ce que montre réellement ToxTwin

ToxTwin illustre cette distinction, et il convient de cadrer ce qu'il établit exactement.

Le résultat : les données synthétiques générées pour modéliser des relations exposition-réponse demeurent aisément séparables des essais réels par un discriminateur.

Le protocole : un classifieur toxicologique est entraîné sur ces seules données synthétiques, puis évalué selon Train-on-Synthetic / Test-on-Real sur des composés jamais observés pendant l'apprentissage, à l'intérieur du domaine de validité déclaré.

Ce que cela prouve : dans ce domaine, la performance de discrimination reste compatible avec celle obtenue à partir des données réelles.

Ce que cela ne prouve pas : l'indistinguabilité statistique générale, ni, en l'état de cette démonstration, la préservation de la calibration et du bénéfice.

La ressemblance distributionnelle globale échoue. La substitution, pour la tâche considérée, réussit.

L'exemple ne démontre évidemment pas que tous les jumeaux deviennent substituables. Il montre, plus modestement, que l'indistinguabilité n'est ni une condition nécessaire ni une preuve suffisante de déployabilité.

PREDICARE éclaire le problème symétrique. Un jumeau destiné au triage clinique n'a pas seulement l'obligation de produire une décision. Il doit aussi identifier les situations pour lesquelles cette décision n'est plus garantie. Là encore, la gouvernance commence quand le système reconnaît sa propre sortie de domaine.

Ce que le COMEX porte réellement devant le régulateur

L'erreur stratégique serait d'attendre d'une génération de modèles toujours plus réalistes qu'elle résolve une question qui relève de l'architecture. La propriété recherchée n'est pas le réalisme maximal. C'est la gouvernabilité de la substitution.

Cette gouvernabilité ne reste pas une formule si on l'arrime à l'instrument réglementaire qui l'encode déjà. Un dossier ne se dépose pas comme un score de ressemblance. Il se dépose comme une chaîne de garanties : une tâche explicitement définie, une séparation stricte entre génération et validation, une démonstration de substituabilité sur données réelles indépendantes, un domaine d'applicabilité déclaré, des mécanismes de détection des situations hors domaine, et une surveillance continue vérifiant que ces garanties demeurent valides lorsque les pratiques cliniques, les populations ou les traitements évoluent.

C'est précisément la logique des plans de modification prédéterminés que les régulateurs commencent à reconnaître pour les dispositifs fondés sur l'apprentissage (le Predetermined Change Control Plan de la FDA en est l'expression la plus aboutie) : autoriser à l'avance une enveloppe de changements bornée, sous surveillance, plutôt que de figer un modèle. La doctrine du port de promotion en propose la primitive d'architecture.

Devant un régulateur, la responsabilité ne porte donc ni sur la qualité esthétique des données synthétiques, ni sur les performances isolées du générateur. Elle porte sur la capacité de l'ensemble de la chaîne décisionnelle à produire, dans un domaine explicitement déclaré, des décisions dont la substitution au réel reste démontrée et continuellement surveillée.

Un jumeau numérique n'est donc pas un objet autonome. C'est un composant d'une architecture de décision. Et comme toute architecture critique, il ne se valide pas en miroir. Il se valide par les garanties qu'il apporte aux décisions qu'il remplace.

Un jumeau qui ressemble se montre. Un jumeau qui remplace se gouverne.

[Série : Digital Twin en santé - 9/12 – Article dominical de clôture de la série hebdomadaire]