



Le jumeau numérique du patient : anatomie d'un malentendu épistémologique

"**Digital twin patient**" Trois mots, aucune définition précise (terme hautement polysémique). Des milliards de valorisation.

La HealthTech a importé ce concept d'un monde industriel (Manufacturing 4.0) où les données s'échantillonnent à la milliseconde sur des milliers de capteurs calibrés.

Elle l'a appliqué à un patient qu'on voit quatre fois par an, dans un régime de données structurellement pauvre, hétérogène, et majoritairement manquant.

Personne ne parle de ce gouffre. C'est un problème majeur. Mais il y a une conclusion que la logique impose : un jumeau numérique cliniquement valide est nécessairement un modèle étroit (une échelle, une question, un domaine de validité).

Si c'est vrai (et nous allons le démontrer), alors la réponse n'est pas un meilleur jumeau. C'est un générateur de jumeaux numériques, seul capable de pouvoir embrasser la variété des cas d'usage, la variabilité des données, et de réduire le coût marginal de génération d'un nouveau jumeau numérique. Et ça change tout à l'architecture des solutions.

I. Le spectre ignoré : modèle, shadow, thread, twin

Quatre niveaux/définitions que l'industrie Healthtech confond sous un seul terme.

Pourtant, le concept de jumeau numérique a une origine précise. Michael Grieves, en 2002, travaillant avec la NASA sur la maintenance prédictive de systèmes aérospatiaux,

formalise l'idée d'un *modèle computationnel couplé en temps réel à son référent physique* (et ce de façon bidirectionnelle).

Dans ses travaux ultérieurs, Grieves lui-même distingue des niveaux d'abstraction que le marketing HealthTech amalgame avec une constance remarquable.

Le modèle statistique. Représentation abstraite d'une relation entre variables, construite sur des données historiques. Pas de couplage avec le référent. Pas de mise à jour dynamique. Une photographie calculée, parfois de grande qualité, parfois (cliniquement) utile mais structurellement et essentiellement statique.

C'est ce que produit 90% de ce qui se vend aujourd'hui comme "jumeau numérique patient". Appeler cela un jumeau numérique, c'est appeler une radiographie un hologramme.

Le Digital Shadow. Le référent physique alimente le modèle en données temps réel. Le flux d'information est unidirectionnel : réel → numérique. Le modèle s'enrichit, se recalibre, suit la trajectoire de son référent sans agir sur lui. Un dossier patient enrichi en continu par des capteurs connectés est un shadow et non pas un twin. C'est une distinction non triviale : le shadow est atteignable dans un périmètre clinique délimité, et réglementairement défendable. Le qualifier de jumeau est une sur-promesse qui prépare une déception.

Le Digital Thread. Niveau souvent absent de la discussion HealthTech, pourtant fondamental dans la conceptualisation industrielle originale : l'infrastructure de données qui relie les artefacts numériques à *travers le cycle de vie complet du référent physique*. Dans l'industrie manufacturière, le twin n'existe pas sans thread : le jumeau d'une turbine suppose une continuité de données depuis la conception jusqu'à la maintenance.

En santé, ce thread correspondrait à l'infrastructure longitudinale continue (EHR interopérables, capteurs, imagerie, données d'essais cliniques, comptes-rendus structurés) couvrant la trajectoire de vie du patient. Cette infrastructure n'existe pas, ou n'existe qu'en fragments incompatibles, dans la quasi-totalité des systèmes de soin actuels.

L'absence du thread explique structurellement pourquoi le twin reste inaccessible : on ne peut pas construire un jumeau sans continuité de données, et la continuité de données en santé est précisément ce que les systèmes d'information hospitaliers ne garantissent pas.

Le Digital Twin au sens strict. Couplage bidirectionnel entre le modèle et son référent. Mais une précision s'impose ici, qui renforce plutôt qu'elle n'affaiblit l'argument : en santé, ce couplage n'est jamais direct modèle → patient. Il est nécessairement médié

par une boucle de décision clinique. Le système prédit ou recommande ; le clinicien évalue, décide, et agit ; l'action modifie l'état du patient, qui rétroalimente le modèle.

En Santé, il s'agit d'une boucle socio-technique, pas purement cyber-physique. Le clinicien n'est pas un simple relais dans la boucle, il doit être un agent décisionnel autonome, porteur de responsabilité légale et éthique, dont le jugement ne peut ni ne doit être contourné (garanties humaines).

Cette architecture socio-technique est plus riche que le twin industriel mais aussi plus complexe à modéliser, à valider, et à réguler.

Les questions à poser à tout vendeur de "digital twin patient" est donc : à quel niveau opère votre système ? Dispose-t-il du thread de données nécessaire ? La boucle de décision clinique est-elle explicitement intégrée dans l'architecture ? Et la question qui tue, votre solution a t'elle fait l'objet d'une validation médico-économique ? Neuf fois sur dix, la réponse honnête est : non, non, et non, et... non.

II. Le biais réductionniste

Une tension historique que la HealthTech redécouvre naïvement est qu'il existe probablement une raison profonde pour laquelle ces confusions persistent (autre que pour faire du buzz et impressionner les VCs).

La médecine computationnelle a hérité d'une intuition épistémologique dominante issue de la biologie moléculaire : pour comprendre un phénomène biologique, décomposez-le jusqu'à ses constituants élémentaires (biais analytique "occidental cartésianiste).

La vérité est dans les détails moléculaires. Cette intuition a une puissance opératoire réelle à l'échelle à laquelle elle s'applique, le docking ligand-récepteur, la biologie structurale, la conception rationnelle de médicaments.

Elle devient un biais analytique majeur quand on suppose qu'elle traverse les échelles sans rupture (bah quelles échelles ?).

Ce n'est pas une observation nouvelle. En biologie des systèmes, la question a été formalisée avec une rigueur croissante. Denis Noble, dans *The Music of Life* (2006), établit que le gène n'a pas de sens indépendamment du contexte cellulaire, tissulaire et organismal dans lequel il opère. Il définit ce qu'il appelle la "relativité causale" entre niveaux d'organisation. Chaque niveau génère des propriétés émergentes irréductibles au niveau inférieur, non par mysticisme, mais par la structure même des équations qui gouvernent les dynamiques à cette échelle.

Cette thèse est formellement ancrée dans la tradition de la biologie multi-échelles, dont les fondements théoriques remontent aux travaux de Mesarovic et Takahara sur les systèmes hiérarchiques (*General Systems Theory*, 1975) : un système à plusieurs

niveaux ne peut pas être décrit par les seules équations de son niveau le plus bas, parce que les contraintes imposées par les niveaux supérieurs modifient le comportement du niveau inférieur de façon irréductible à ce niveau.

Philip Anderson l'avait formulé en 1972 dans *More is Different* pour la physique de la matière condensée : à chaque niveau d'organisation, de nouvelles lois apparaissent, non déductibles des équations du niveau inférieur. Richard Levins et Richard Lewontin en développaient les implications biologiques en 1985 dans *The Dialectical Biologist*. Hiroaki Kitano posait en 2002 les fondements de la biologie des systèmes précisément pour traiter ces ruptures d'échelle comme des objets de recherche à part entière et non comme des détails d'implémentation.

Ce que la HealthTech redécouvre naïvement en 2024, la pharmacologie quantitative l'a formalisé depuis trente ans sous la forme progressive : PK/PD → modèles semi-mécanistiques → Quantitative Systems Pharmacology (QSP).

Le passage de la concentration plasmatique à la fixation d'un ligand sur son récepteur en compétition avec d'autres molécules et avec des effets d'encombrement stérique... l'effet de l'activation de voies de communication intra/inter-cellulaire modifiant le comportement de la cellule, de l'effet cellulaire à la réponse tissulaire, de la réponse tissulaire à l'effet sur un organe, puis un organisme, jusqu'à l'effet clinique pose un sérieux problème de puissance de calcul et de modélisation de systèmes complexes avec leurs propriétés d'émergence propre à de multiples échelles d'appréhension.

La pharmacologie l'a su et a construit des approches méthodologiques pour y répondre tenter de modéliser ces différents mécanismes. La HealthTech, pressée par les cycles de levée de fonds, saute ces étapes et appelle "jumeau numérique" le résultat.

Suivons concrètement le chemin de la rupture :

- Un ligand se fixe sur un récepteur avec une affinité nanomolaire modélisée avec précision cristallographique. Cette affinité ne prédit pas l'effet pharmacodynamique in vivo : biodisponibilité, variabilité d'expression tissulaire du récepteur, interactions allostériques non modélisées.
- L'effet cellulaire in vitro ne prédit pas l'effet tissulaire (le tissu est un système organisé avec une architecture spatiale et des interactions paracrines absentes du puits de culture).
- L'effet tissulaire ne prédit pas l'effet organe.
- L'effet organe ne prédit pas l'effet organisme, régulé par des boucles homéostatiques dont le modèle moléculaire ne sait rien.

- L'effet organisme ne prédit pas l'effet clinique parce que le patient vit dans un contexte irréductible à ses biomarqueurs : histoire, comorbidités, observance, décisions non encore prises.

Le problème n'est pas le réductionnisme lui-même c'est son passage illégitime d'échelle. Et ce passage illégitime est précisément ce que la majorité des architectures de "jumeaux numériques" en santé commettent implicitement, sans le nommer.

III. Le gouffre innommé des données : IoT Industriel versus données HDLSS cliniques

Le concept de jumeau numérique industriel a été développé dans un monde de données denses, continues, objectives et exhaustives. L'écart avec la réalité clinique mérite d'être mesuré sans euphémisme.

Dans l'industrie, un jumeau de turbine opère sur des centaines à milliers de capteurs calibrés, un échantillonnage à la milliseconde, une couverture temporelle de 100% du fonctionnement, un bruit de mesure connu et modélisable, une ontologie fixe et normalisée, et des lois de comportement connues (thermodynamique, mécanique des fluides...) qui contraignent le modèle et permettent une validation continue par comparaison mesure/prédiction.

Un patient suivi pour une pathologie chronique stable en médecine de ville produit quatre consultations par an, dix à cinquante variables mesurées selon le protocole, des données hétérogènes dont la nature varie selon le prescripteur, le laboratoire et l'heure de prélèvement, une couverture temporelle effective inférieure à 1% du temps de vie réel, un bruit de mesure non caractérisé, et des ontologies fragmentées entre SNOMED, CIM-10, LOINC et terminologies locales ... voire aucune ontologie formelle.

Cette situation a un nom dans la littérature bioinformatique et statistique : Données HDLSS (High Dimension, Low Sample Size). Des milliers de variables potentielles pour un nombre d'observations longitudinales par patient qui se compte en dizaines. Dans ce régime, la malédiction de la dimensionnalité n'est pas un détail technique, c'est une contrainte fondamentale sur ce qu'il est possible d'inférer, quelle que soit la puissance de calcul.

Ce problème n'est pas nouveau. La pharmacologie de population l'a "résolu", ou du moins formalisé, depuis les années 1980. Sheiner et Beal, à l'Université de Californie San Francisco, ont développé dès 1980 les modèles à effets mixtes non-linéaires (NLME), implémentés dans NONMEM (NON-linear Mixed Effects Modeling), précisément pour estimer des paramètres pharmacocinétiques individuels dans des cohortes où les observations par patient sont peu nombreuses, irrégulièrement espacées, et structurellement hétérogènes. Le *framework* NLME décompose la variabilité en variabilité inter-individuelle (paramètres aléatoires de niveau 1) et variabilité résiduelle intra-individuelle, avec une structure de covariance explicite. Ce

corpus enrichi par Davidian et Giltinan (1995) et par les développements SAEM de Kuhn et Lavielle (2005) constitue le précédent méthodologique le plus direct à la problématique HDLSS clinique.

La HealthTech, pressée par les cycles de levée de fonds, l'ignore et réinvente des roues de qualité inférieure.

Trois conséquences structurelles méritent d'être formulées explicitement :

- L'interpolation entre observations est une fiction assumée ou non,
- Entre deux consultations espacées de trois mois, la trajectoire réelle du patient est une zone blanche non couverte,
- Un modèle qui n'assume pas explicitement cette ignorance la dissimule (ce qui est pire).

A ce déficit de fréquence d'échantillonnage s'ajoute la grande hétérogénéité des données et ce de façon structurelle :

- Une glycémie au laboratoire certifié et une glycémie capillaire à domicile ne sont pas la même variable même sous le même libellé dans le dossier (il existe aussi bien entendu des variabilités liées aux techniques d'analyse biologique et aux fournisseurs de "réactifs" de chaque laboratoire d'analyse).
- Le signal manquant est presque jamais aléatoire. La grande majorité des données manquantes en clinique relèvent du mécanisme Missing Not At Random (MNAR) : elles manquent parce que le patient n'est pas venu, parce que la valeur était anormale, parce que le clinicien a renoncé à saisir dans un champ structuré. Imputer sans modéliser ce mécanisme introduit un biais systématique de sous-estimation du risque réel.
- À cela s'ajoute un point statistique fondamental que les approches data-driven ignorent régulièrement : la dépendance intra-patient. Les observations longitudinales d'un même individu ne sont pas indépendantes. Les méthodes doivent gérer simultanément les corrélations intra-individuelles, les trajectoires non stationnaires, et les censures informatives, situations où la fin d'observation est elle-même informative sur l'état du patient (patient décédé avant la fin de l'étude par exemple).

Ignorer ces dépendances dans le design du modèle, c'est produire des estimateurs biaisés avec des intervalles de confiance incorrects et une illusion de précision clinique, et donc *In Fine* une médecine de précision de pacotille.

IV. L'loMT et les wearables ?

Peut-on enrichir les données cliniques et notamment résoudre la problématique de la faiblesse d'échantillonnage ?

Les dispositifs IoMT et wearables semblent offrir une réponse directe au problème HDLSS.

Avant d'examiner leurs limites, il faut reconnaître leurs contributions réelles sans quoi l'argument devient caricature. L'IoMT a démontré une utilité clinique documentée dans plusieurs domaines :

- La détection précoce d'événements arythmiques par photopléthysmographie et ECG à la demande. L'étude Apple Heart Study (2019, NEJM) a démontré la capacité de détection de la fibrillation auriculaire à une échelle de population inédite. Toutefois, il convient de noter que la valeur prédictive positive était de 0,84 dans la sous-population notifiée, et non dans l'ensemble de la population screenée, ce qui modère substantiellement l'enthousiasme sur le taux de détection global.
- Le phénotypage comportemental longitudinal (activité physique, qualité de sommeil, variabilité circadienne) qui enrichit qualitativement la représentation du patient au-delà des biomarqueurs biologiques classiques.
- La surveillance thérapeutique en boucle fermée avec le CGM (contrôle de la glycémie en continu) en diabétologie,
- Le monitoring hémodynamique implantable en insuffisance cardiaque avancée où le couplage données/décision est cliniquement validé et réglementairement encadré.

Dans ces cas contraints, l'IoMT produit un digital shadow légitime : des biomarqueurs validés, une question clinique précise, un couplage données/décision documenté. C'est réel et significatif. Mais l'IoMT transforme le régime HDLSS sans l'abolir et introduit de nouvelles pathologies épistémiques propres.

La validité clinique des biomarqueurs wearables n'est pas transitive. La fréquence cardiaque par PPG (photopléthysmographie) sur bracelet n'est pas la fréquence cardiaque par ECG. La précision dans certaines conditions (mouvement, pigmentation cutanée, arythmies) et surtout la valeur prédictive clinique sont structurellement différentes.

Utiliser l'une comme proxy de l'autre sans validation croisée rigoureuse, c'est réintroduire de l'hétérogénéité habillée en densité plus dangereuse que l'hétérogénéité déclarée, parce qu'invisible.

Le bruit à haute fréquence exige une chaîne de traitement dont les choix ne sont pas neutres. Fenêtrage, filtrage, agrégation temporelle : deux équipes traitant le même flux brut peuvent produire des features incomparables.

La reproductibilité du prétraitement est une condition de la reproductibilité du modèle, rarement documentée dans les publications. Le biais de sélection comportementale est structurel. Les patients qui portent des wearables en continu pendant six mois ne représentent pas la population générale des malades chroniques. Ce sont des patients engagés, souvent moins sévèrement atteints ou au contraire très motivés. Le modèle entraîné sur leurs trajectoires ne s'applique pas nécessairement aux patients en situation de précarité ou de comorbidités sévères (précisément ceux que la médecine prédictive cherche le plus à atteindre) comme dans le programme PREDICARE ([https://www.linkedin.com/posts/jeromev_santaez-praezvention-maladieschroniques-activity-7420487123444604929-
ui5i?utm_source=share&utm_medium=member_desktop&rcm=ACoAAChnekBGR
F-CabUoQQTa69kbB9dT2pmuMI](https://www.linkedin.com/posts/jeromev_santaez-praezvention-maladieschroniques-activity-7420487123444604929-
ui5i?utm_source=share&utm_medium=member_desktop&rcm=ACoAAChnekBGR
F-CabUoQQTa69kbB9dT2pmuMI)) que nous avons imaginé. L'loMT risque de produire des modèles excellents pour les patients qui en ont le moins besoin. J'avais déjà discuté de ce biais de sélection avec les équipes de Withings : les propriétaires de montres connectées ne sont pas ceux qui font du sport devant leur télé en mangeant une glace.

V. Le hiatus algorithmique :

Il existe trois incompatibilités fondamentales au cœur des architectures actuelles qui visent à simuler des essais cliniques "*In Silico*".

C'est le problème le moins discuté dans la littérature appliquée et cependant, probablement le plus important pour quiconque construit ou évalue sérieusement une architecture de jumeau numérique patient.

Les architectures les plus sophistiquées du secteur (modélisation d'essais cliniques *In Silico*) reposent, implicitement ou explicitement, sur une pile de trois couches de nature fondamentalement différente :

- Couche 1 : Génération de la population synthétique (IA générative). Modèles génératifs de type VAE, GAN, modèles de diffusion, transformers entraînés sur des cohortes réelles pour produire des patients synthétiques statistiquement plausibles. L'objectif est de capturer la distribution jointe des caractéristiques démographiques, biologiques et cliniques à $t=0$: un instantané de population. C'est une modélisation fondamentalement statique : elle capture la variabilité inter-individuelle à un instant donné, sans dynamique temporelle ni dépendances causales dans le temps. C'est une photo "instantanée". C'est joli parfois, mais ne permet pas de capturer la dynamique temporelle d'un essai clinique.

- Couche 2 : Modèle pathologique avec introduction de la variable temporelle. Pour modéliser l'évolution de la pathologie : modèles de survie (Cox proportionnel, modèles à risques compétitifs) ou systèmes d'équations différentielles ordinaires (ODE) représentant la dynamique des compartiments biologiques. Modèles mécanistiques ou semi-mécanistiques, ancrés dans des hypothèses causales explicites, avec des paramètres interprétables (hazard ratios, constantes cinétiques). Associé à la couche 1, cela permet de produire une population synthétique pathologique avec une pathologie qui va progresser dans le temps : peut servir à créer un bras contrôle dans un essai qui n'en dispose plus ou pas.
- Couche 3 : Modèle thérapeutique PK/PD. La réponse au traitement modélisée par des ODE décrivant l'évolution de la concentration du médicament et sa traduction en effet biologique. Couche mécanistique supplémentaire, dont les paramètres sont estimés en population dans le cadre des modèles non-linéaires à effets mixtes (NLME).

Le hiatus entre ces couches peut être décomposé en trois incompatibilités fondamentales :

- Incompatibilité ontologique : corrélation versus causalité. La couche générative capture des corrélations dans l'espace des features à $t=0$. Elle ne pose pas de question causale et ne distingue pas conséquence de confondant. Les couches 2 et 3 supposent une structure causale explicite. Un patient synthétique produit par la couche 1 (construit sur des corrélations statistiques) est passé dans des modèles qui supposent des dépendances causales. *Cette transition est effectuée sans justification.*
- Incompatibilité temporelle : statique versus dynamique. La couche 1 produit un instantané sans histoire causale. Les couches 2 et 3 opèrent sur des trajectoires temporelles continues ou semi-continues, avec des hypothèses sur la forme de la dynamique (proportionnalité des risques pour Cox, continuité différentielle pour les ODE). Comment un patient sans passé dynamique est-il passé dans un modèle dont les hypothèses présupposent une trajectoire antérieure ? Cette question est résolue dans la plupart des implémentations par une heuristique d'initialisation qui est rarement documentée, jamais validée.
- Incompatibilité probabiliste : distribution jointe versus processus génératif. La couche générative produit des samples d'une distribution jointe. Les couches mécanistiques propagent une incertitude paramétrique structurée (intervalles de confiance sur les hazard ratios, variabilité des paramètres NLME). Comment l'incertitude de la distribution jointe se compose-t-elle avec l'incertitude paramétrique des couches suivantes ? Cette composition n'est pas triviale, et ses propriétés ne sont pas celles d'un simple produit de convolutions.

Ces trois incompatibilités sont résolues, dans la plupart des implémentations actuelles, par des heuristiques d'ingénierie : passage d'un point de sortie de la couche i à l'entrée de la couche $i+1$, sans traitement explicite des discontinuités.

Le modèle global est cohérent numériquement. Il n'est pas cohérent épistémologiquement. Et dans un contexte de décision clinique, cette distinction n'est pas académique.

VI. Pourquoi l'unification end-to-end ne résout pas le hiatus

Face à ce problème de cohérence inter-couches, l'intuition industrielle est compréhensible : si trois modèles hétérogènes ne se parlent pas, remplaçons-les par un seul modèle capable de tout apprendre depuis les données : LSTM, xLSTM, Temporal Fusion Transformer, architectures Mamba ou S4 pour les séquences longues.

Cette intuition (qui fut la mienne) mérite d'être challengée sur trois fronts, avec une nuance préalable : les architectures séquentielles profondes ne sont pas uniquement prédictives. Des travaux récents explorent la causalité dans les séries temporelles (Neural Controlled Differential Equations (Kidger et al., 2020), Deep State Space Models, modèles causaux structurés sur séquences).

Ces avancées sont réelles et ne sauraient être ignorées. Elles ne résolvent cependant pas les trois problèmes suivants :

- L'interprétabilité causale disparaît à un coût réglementaire et clinique élevé. Un modèle de Cox produit des hazard ratios interprétables, testables, contestables par un clinicien. Un système d'ODE produit des paramètres physiologiques ayant un sens biologique direct. Un Temporal Transformer produit des poids d'attention puissants pour la prédiction, opaques pour la compréhension mécanistique. Dans un contexte MDR Classe IIa, où le dispositif doit être expliqué au patient, justifié devant un comité éthique, et défendu devant un auditeur réglementaire, l'opacité n'est pas un détail technique. C'est une contrainte de conception que l'architecture end-to-end ignore structurellement.
- L'exigence de données explose précisément là où elles sont rares. Les architectures profondes séquentielles nécessitent des trajectoires longues, denses et relativement homogènes pour apprendre des représentations stables. Dans le régime HDLSS clinique, elles sont sous-alimentées. Un modèle de Cox estimé sur 200 patients avec 4 observations longitudinales chacun est statistiquement défendable et ses propriétés asymptotiques sont connues. Un Temporal Transformer entraîné sur les mêmes données sur-apprend, produit des intervalles de confiance illusoire, et généralise mal hors de la distribution d'entraînement. La connaissance mécanistique a priori est abandonnée sans bénéfice démontré. Les ODE et les modèles PK/PD incorporent des décennies de connaissances biologiques validées, ils contraignent l'espace des solutions vers

des régions biologiquement plausibles, réduisant dramatiquement le risque de sur-ajustement.

- Un modèle purement data-driven peut converger vers des solutions numériquement performantes sur la cohorte d'entraînement et biologiquement absurdes en extrapolation. Dans des domaines où l'extrapolation est précisément l'enjeu clinique (prédire la trajectoire future d'un patient) **c'est une propriété rédhibitoire.**

La voie la plus prometteuse est celle des architectures hybrides. Les Neural ODEs (Chen et al., 2018), les Universal Differential Equations (Rackauckas et al., 2020), et les Physics-Informed Neural Networks (Raissi et al., 2019) tentent cette réconciliation : laisser le réseau apprendre les termes de l'équation différentielle non connus a priori, en contraignant la dynamique globale par les lois mécanistiques connues. Dans le domaine pharmaceutique, les modèles hybrides mécanistiques/ML en Quantitative Systems Pharmacology (QSP) représentent une direction active et crédible précisément parce qu'ils préservent l'interprétabilité des paramètres physiologiques tout en gagnant en flexibilité là où la connaissance mécanistique est incomplète.

Les PINNs supposent une connaissance mécanistique suffisante pour formuler les termes de l'équation différentielle comme contraintes. Dans le régime HDLSS clinique, cette connaissance est souvent partielle. Les SINNs que nous développons chez Qualees proposent de substituer à la contrainte mécanistique déterministe une contrainte distributionnelle : le réseau est entraîné sous contrainte de cohérence avec des intervalles statistiques estimés sur les données disponibles (intervalles de prédiction issus de modèles de référence, quantiles de trajectoires observées, ou distributions a priori informées par la littérature clinique). La contrainte reste différentiable et intégrable dans la fonction de perte. Ce n'est pas de la régularisation au sens classique, c'est une inférence sous contrainte épistémique partielle.

Ces approches sont encore largement en phase de recherche. Elles ne sont pas en déploiement clinique opérationnel généralisé. Les présenter comme telles serait reproduire exactement le biais que cet article dénonce.

VII. Ce qui reste défendable

La rigueur de ce diagnostic ne disqualifie pas la modélisation en santé. Elle la recadre dans ses conditions de validité qui sont réelles et cliniquement significatives.

George Box l'a formulé définitivement : "All models are wrong, but some are useful." Un modèle n'est pas jugé sur sa complétude. Il est jugé sur son adéquation à une question spécifiée, dans un contexte délimité, pour une décision identifiée.

Se satisfaire d'un modèle partiel n'est pas une capitulation scientifique, c'est une vertu épistémologique. La parcimonie est une garantie de robustesse. Ce principe a un

équivalent réglementaire déjà formalisé dans l'évaluation des dispositifs médicaux logiciels : le Domaine de Validité (la définition explicite des conditions dans lesquelles un modèle est valide, et au-delà desquelles ses prédictions ne peuvent pas être extrapolées). C'est une exigence réglementaire dans certains contextes MDR.

Ce devrait être une exigence scientifique systématique bien avant que la réglementation ne l'impose.

Cela conduit à un principe opérationnel central : l'IA étroite n'est pas une limitation, c'est une condition de validité. Un modèle étroit, construit à une échelle explicitement choisie, entraîné sur une cohorte homogène, pour prédire un événement clinique défini avec un horizon temporel spécifié, une incertitude quantifiée, et un domaine de validité documenté est défendable devant un comité scientifique, une autorité réglementaire, et un clinicien qui doit décider comment intégrer cette information dans son propre jugement.

Quatre critères de validité minimale pour tout système revendiquant la qualification de "jumeau numérique patient" :

- 1. Échelle explicite. À quel niveau opère le modèle ? Moléculaire, cellulaire, organe, patient, population ? Les transitions d'échelle sont-elles modélisées ou simplement ignorées ?
- 2. Question clinique définie. Le modèle répond-il à une question de diagnostic, de pronostic, ou d'aide à la décision thérapeutique ? Pour quelle population cible ? Sur quel horizon temporel ?
- 3. Régime de données documenté. HDLSS, IoMT, essai clinique, données de vie réelle ? Le mécanisme de données manquantes est-il modélisé ? La dépendance intra-patient est-elle traitée ?
- 4. Architecture causale explicite. Le modèle est-il statistique, mécanistique, ou hybride ? Le hiatus entre ses couches est-il documenté ? La propagation de l'incertitude à travers les couches est-elle quantifiée ?

Sans ces quatre éléments, le terme "digital twin patient" n'a pas de contenu scientifique défendable. Il n'est que du marketing.

VIII. Pourquoi ça compte maintenant

Nous sommes à un point d'inflexion. Les premiers déploiements à grande échelle d'IA clinique produisent des données de vie réelle.

Les déceptions existent. Les retraits de solutions surviennent silencieusement. Les institutions hospitalières deviennent prudentes parfois au point de bloquer des outils qui fonctionneraient, par méfiance légitime vis-à-vis d'un secteur qui a trop promis.

Dans ce contexte, le surdimensionnement des promesses ne fait pas que décevoir ponctuellement. Il détruit la crédibilité systémique de technologies qui pourraient réellement améliorer des trajectoires de soin.

Chaque pitch deck qui vend un "jumeau numérique complet du patient" sans définir le terme, sans préciser l'échelle, sans nommer le hiatus algorithmique entre ses couches, contribue à un climat de méfiance qui ralentit l'adoption des outils qui tiennent leurs promesses (parce qu'ils ont eu l'intelligence et l'honnêteté de les formuler correctement).

Les questions à se poser (si vous construisez, évaluez, ou financez une solution de ce type) ne sont pas "notre benchmark est-il satisfaisant ?" Elles sont :

- à quelle échelle opère le modèle ?
- Quelle question clinique précise répond-il ?
- Comment l'incertitude se propage-t-elle à travers ses couches ?
- Que se passe-t-il quand les données d'entrée s'écartent de la distribution d'entraînement et comment le modèle le signale-t-il ?
- Le domaine de validité est-il documenté ?
- Qui est responsable quand le modèle se trompe hors de ce domaine ?

Ces questions ne sont pas des obstacles au déploiement. Ce sont les conditions minimales d'un déploiement responsable.

Le jumeau numérique du patient ne sera jamais un miroir. C'est une carte. Partielle, construite à une échelle choisie, depuis un point de vue assumé, avec des zones blanches documentées et un domaine de validité explicité, pour répondre à une question que son concepteur a eu le soin de formuler avant d'écrire une seule ligne de code. Et comme toutes les cartes, elle n'est pas le territoire. Mais entre les mains d'un navigateur qui connaît l'échelle à laquelle elle a été dessinée, qui sait ce qu'elle représente et ce qu'elle tait, qui respecte ses limites et sait à quel moment en changer : elle peut éviter bien des naufrages. Dans un secteur où les décisions influencées par des modèles algorithmiques affectent des trajectoires de soin réelles, la rigueur épistémologique n'est pas un luxe académique : **C'est la seule forme d'honnêteté à la hauteur de l'enjeu.**

Conclusion architecturale : non pas un meilleur jumeau mais un générateur

Il existe une conclusion que la logique de cette démonstration impose indépendamment de tout choix technologique. Si un jumeau numérique cliniquement valide est nécessairement étroit (une échelle explicite, une question clinique définie/un

cas d'usage spécifique, un domaine de validité documenté) alors aucun jumeau unique ne peut couvrir la diversité des cas d'usage cliniques :

- Les pathologies diffèrent,
- Les horizons temporels diffèrent,
- Les questions diffèrent,
- Les régimes de données disponibles diffèrent,

Un modèle étroit valide pour la prédiction de rechute en oncologie hématologique n'est pas valide pour la détection précoce de décompensation cardiaque. Ce ne sont pas des insuffisances techniques à corriger, ce sont des contraintes épistémologiques structurelles.

La réponse architecturale n'est donc pas de construire un jumeau plus grand, plus général, plus puissant. C'est de construire un générateur de jumeaux numériques : une méta-architecture capable de produire le bon modèle étroit pour chaque question clinique spécifique, à partir d'une infrastructure commune de données et d'ontologies, en instanciant à la demande l'échelle pertinente, l'architecture causale appropriée, et le régime de données documenté.

C'est une différence de nature, pas de degré. La majorité des acteurs HealthTech construisent un jumeau. Mais un générateur opère à un niveau d'abstraction supérieur, celui de la fabrique, pas du produit.

C'est précisément le problème que TweenMe a choisi de résoudre. Non pas en construisant le jumeau numérique du patient (cette promesse, cet article l'a montré, n'a pas de contenu scientifique défendable), mais en construisant l'infrastructure qui génère le bon jumeau pour la bonne question, avec les bonnes contraintes épistémologiques, dans le bon domaine de validité. La critique devient le cahier des charges. L'honnêteté épistémologique devient l'architecture produit.