La preuve est une promesse conditionnelle

Pour une théorie relationnelle et émergente de la validation des preuves computationnelles en santé, et au-delà

I. Trois contradictions, et pourquoi ce sont des impossibilités

Commençons par renoncer à une habitude. L'habitude consiste à demander si une preuve est valide, comme on demande si un pont est solide ou si un théorème est exact. La question paraît saine. Elle est mal posée, et trois situations ordinaires suffisent à le montrer, à condition d'aller jusqu'au bout de ce qu'elles montrent. Ce bout, on le verra, mène à concevoir la preuve comme une promesse, non comme un objet.

1. Première situation. Deux études cliniques, conduites avec une rigueur irréprochable, portant sur la même intervention, aboutissent à deux recommandations incompatibles. Aucune n'est entachée d'erreur ; protocoles publiés, analyses pré-enregistrées, populations décrites. Pourtant elles ne peuvent pas être *toutes deux* la base d'une même décision. Si la validité était une propriété intrinsèque de chaque étude, deux objets également valides ne devraient pas se contredire.
2. Deuxième situation. Un essai contrôlé randomisé excellent, l'étalon-or, devient strictement inutilisable hors de sa population cible. Le même essai, identique au mot près, est une preuve dans un contexte et un bruit dans un autre. Sa validité ne voyage pas avec lui.
3. Troisième situation. Un modèle prédictif parfaitement calibré (ses probabilités prédites coïncident avec les fréquences observées) conduit à une mauvaise décision clinique, parce que la décision dépend d'un arbitrage entre erreurs que la calibration ignore. Un test bien calibré qui fixe son seuil sans tenir compte du coût asymétrique d'un faux négatif en oncologie est calibré *et* dangereux.

Jusqu'ici, ces situations sont des paradoxes. Un paradoxe se répare, et la communauté connaît les réparations ; il faut donc montrer qu'aucune ne tient, faute de quoi le lecteur conclura, légitimement, que ces exemples révèlent l'insuffisance de *certaines* notions de validité, non la nécessité d'en changer.

Reprenons les réparations une à une. *Se restreindre à la validité interne ?* La première contradiction subsiste : deux études internement valides divergent encore, parce que

l'incompatibilité ne naît pas d'un défaut interne mais de l'usage qu'on en fait. *Ajouter la validité externe ?* Externe par rapport à quoi : à une population, à une décision, à un contexte. On ne définit la validité externe qu'en nommant ce à quoi la preuve doit s'appliquer ; on a donc déjà concédé que la validité est relative à un usage. *Stratifier, méta-analyser ?* On déplace la décision d'un cran (quelle population, quel sous-groupe, quel modèle d'effets) sans supprimer la relativité, seulement en la remontant. *S'en remettre à une propriété statistique pure, la calibration ?* La troisième contradiction l'interdit : calibré n'est pas valide.

Le constat est alors d'une autre nature. Ce n'est pas que telle notion usuelle de validité soit insuffisante ; c'est qu'*aucune notion de validité non relationnelle ne peut dissoudre les trois contradictions*, parce que chaque réparation, pour fonctionner, réintroduit en contrebande une décision, une perte ou un domaine. La validité intrinsèque n'est pas une bonne idée mal exécutée. C'est une impossibilité. Et de cette impossibilité découle une idée plus modeste et plus exigeante : *la validité n'est pas dans la preuve, elle est dans la relation entre une preuve et l'usage qu'on en fait*. Le reste de ce texte tire les conséquences de ce déplacement, jusqu'à un point qui surprendra peut-être : une preuve computationnelle n'est ni un objet à certifier, ni même une relation statique : c'est une *promesse conditionnelle*, dont le contrat d'usage n'est que la formalisation sociale.

Je n'affirme pas cette thèse. Je la laisse émerger, parce qu'une doctrine qui s'ouvre sur son axiome demande qu'on la croie, là où une doctrine qui s'ouvre sur une impossibilité force qu'on la suive.

II. La structure minimale, et l'inversion de la charge de la preuve

Si la validité est relationnelle, alors « valide » est un prédicat incomplet, comme « à gauche de » ou « plus grand que ». Il appelle des arguments. La question n'est plus *cette preuve est-elle valide ?* mais *valide pour quoi, sous quel critère, dans quelles limites, jusqu'à quand ?* Reste à savoir combien d'arguments sont nécessaires, et la réponse n'est pas affaire de goût, car chaque argument se justifie par le fait que son absence ressuscite l'une des trois impossibilités.

Il faut une **décision**. Sans décision à servir, « valide » n'a aucun contenu : c'est la première impossibilité. Notons-la *D*. ***Au sens strict de la théorie statistique de la décision, fondée par Abraham Wald (Statistical Decision Functions, 1950), une décision n'est pas une croyance mais une règle : une fonction δ qui, à une observation, associe une action***. On ne validera jamais assez ce point. Une décision n'est pas une distribution de probabilité ; c'est une application qui transforme de l'information en acte. On ne valide pas une distribution. On valide la qualité d'une règle d'action, et, on le verra, la valeur de ce qu'elle nous apprend.

J'appellerai désormais *source de preuve* tout dispositif produisant les données qui alimentent δ (cohorte réelle, population synthétique, jumeau numérique, simulation), sans préjuger de son statut : c'est elle, et non la donnée brute, qu'on cherchera à rendre substituable.

Il faut un critère pour ordonner les conséquences de la règle : une **fonction de perte**, notée **L**. C'est la troisième impossibilité : le modèle calibré échoue parce qu'il ignore que se tromper dans un sens ne coûte pas ce que coûte se tromper dans l'autre. Wald nomme *risque* l'espérance de cette perte et en fait le seul juge d'une règle. J'ajoute d'emblée deux corrections que la médecine impose et que la théorie classique élude, et que je développe plus loin : cette perte n'est pas un scalaire, et elle n'est pas seulement le coût d'une action : elle inclut la valeur de ce qu'une preuve nous fait *savoir* avant d'agir. La perte porte donc, dès sa première apparition, deux fonctions : décider, et réduire l'incertitude.

Il faut un **domaine** où la preuve est réputée fiable, noté *Δ*. C'est ici la deuxième impossibilité : l'essai randomisé ne transfère pas parce que sa validité était locale et qu'on l'a crue universelle. La notion est codifiée : le troisième des cinq principes de l'OCDE pour la validation des modèles structure-activité (*Guidance Document on the Validation of (Q)SAR Models*, 2007) exige qu'un modèle déclare son domaine d'applicabilité. Hors de *Δ*, une preuve n'est pas fausse ; elle est hors sujet, ce qui est plus dangereux, car une preuve hors sujet a l'air d'une preuve.

Il faut enfin un **temps**, noté *T* : une preuve validée aujourd'hui ne l'est pas pour toujours. Je lui consacre la section VI.

On obtient une structure (*D*, **L**, *Δ*, *T*) que je revendique comme minimale. *Minimale* ne veut pas dire élégante ; cela veut dire qu'on ne peut rien retirer sans casse : ôtez *D* et le prédicat se vide, ôtez **L** et la première impossibilité revient, ôtez *Δ* et c'est la deuxième, ôtez *T* et c'est l'expiration des preuves. La structure n'est pas choisie ; elle est ce qui reste quand on a retiré tout ce dont l'absence se paie d'une contradiction.

Quant à la suffisance, je ne procède pas par aveu : « je ne connais pas de cinquième composante » serait une faiblesse, non un argument. Je pose une règle, et j'inverse la charge de la preuve : *toute composante candidate doit démontrer qu'elle n'est réductible ni à D, ni à L, ni à Δ, ni à T*. À défaut d'une telle démonstration, elle est déjà contenue dans l'une des quatre. Ce n'est pas à la théorie de prouver qu'aucune cinquième composante n'existe ; c'est à qui en propose une de prouver son irréductibilité. La minimalité cesse alors d'être une conjecture timide pour devenir une contrainte opposable.

III. La théorie se juge elle-même

L'objection la plus sérieuse se lève ici. Si la validité n'est jamais qu'une relation à un usage, n'a-t-on pas dissous toute notion de qualité ? Tout ne se vaut-il pas, dès lors que chacun invoque son propre triplet ? Le relativisme guette, et avec lui la ruine de l'argument : une doctrine qui rend tout relatif se rend elle-même indéfendable.

La parade n'est pas de réintroduire en douce une validité absolue. Elle est d'exiger d'une théorie de la validation qu'elle satisfasse des critères eux-mêmes explicites. *Une théorie de validation est admissible si et seulement si elle est cohérente, composable, transférable et réfutable.*

- Cohérente : elle ne se contredit pas,
- Composable : les validations partielles s'assemblent le long d'une chaîne de décisions sans que la garantie se perde en route,
- Transférable : une validation se transporte d'un contexte à un autre sous des conditions déclarées, et non par espoir,
- Réfutable : chacun de ses énoncés peut être démenti par une observation.

La relativité cesse d'être arbitraire : elle est gouvernée par une conjonction de quatre conditions. On ne dit pas « tout se vaut » ; on dit « la valeur d'une preuve se mesure relativement à un usage, et la mesure elle-même obéit à des règles ».

Une théorie qui impose ces critères aux autres doit s'y soumettre, sous peine d'hypocrisie. Passons-la donc à son propre tamis, critère par critère.

- *Cohérence* : la thèse selon laquelle toute preuve est relative à une décision est elle-même relative à une décision (celle de valider des preuves), ce qui la met en accord avec elle-même plutôt qu'en abyme ; elle ne s'auto-réfute pas,
- *Composabilité* : la section X le démontre sur un cas, où une cohorte validée pour une sous-décision alimente une décision plus large sans rupture de garantie, parce que la substituabilité, étant définie relativement à une décision, se compose comme se composent les décisions,
- *Transférabilité* : la section XI transporte la théorie hors de la santé en déclarant ses conditions de transfert, ce qui est exactement ce que le critère réclame,
- *Réfutabilité* : la section II en a donné les conditions, et la règle d'inversion de la charge en fournit une de plus : qu'on exhibe une composante irréductible, et la structure devra céder.

Une théorie qui survit à ses propres critères n'est pas vraie pour autant ; elle est admissible. C'est tout ce qu'une théorie peut prétendre, et c'est déjà plus que ce que l'idée de « validité en soi » a jamais offert.

Reste à désamorcer un dernier malentendu. Cette théorie ne prétend pas remplacer les cadres établis de la preuve : ni GRADE pour la gradation des recommandations, ni CONSORT pour le compte rendu des essais, ni TRIPOD pour les modèles pronostiques. Elle prétend décrire le *niveau* auquel ces cadres deviennent comparables : chacun, à sa manière, déclare une décision, un critère, un domaine et une fenêtre de validité. La théorie relationnelle n'est pas un concurrent de plus dans la liste ; c'est la grammaire commune qui explique pourquoi ces cadres font ce qu'ils font, et ce qu'ils ont en partage. Une théorie qui se sait métathéorie ne menace pas les théories ; elle les situe.

IV. La perte est contextuelle avant d'être vectorielle

J'ai dit que la perte n'est pas un scalaire. Il faut le démontrer, car toute la théorie de la décision classique repose sur l'idée qu'on peut résumer les conséquences d'une action par un nombre unique et minimiser son espérance.

Or une décision clinique met en balance, simultanément, des grandeurs irréductibles l'une à l'autre : l'efficacité, la sécurité, l'équité entre sous-populations, le coût, l'acceptabilité par le patient, la conformité réglementaire. Ce ne sont pas des nuances d'une même quantité ; ce sont des dimensions distinctes, parfois antagonistes. La perte est un *vecteur* **L**, et la décision optimale n'est pas un minimum mais un front, ce que l'optimisation multi-objectif nomme un front de Pareto, l'ensemble des arbitrages dont aucun ne domine les autres sur toutes les dimensions à la fois.

Mais il faut aller plus loin, car s'arrêter au vecteur laisserait croire que les dimensions et leurs poids sont fixes. Ils ne le sont pas. Les préférences qui pondèrent l'efficacité contre la toxicité dépendent du patient, du stade, de l'âge, du projet de soin ; ce qui est un bon arbitrage pour l'un est inacceptable pour l'autre. *La perte est contextuelle avant d'être vectorielle* : elle n'est pas un vecteur donné une fois pour toutes, mais un vecteur dont les composantes et les pondérations sont elles-mêmes fonction du contexte de décision. C'est une exigence de plus, pas une complication gratuite : elle interdit de transporter un arbitrage d'un contexte à un autre comme s'il était neutre.

La conséquence est plus politique que technique. Réduire ce vecteur contextuel à un scalaire (fixer un taux d'échange entre une année de vie et un euro, entre la sécurité d'un groupe et l'efficacité moyenne) n'est pas une opération de calcul ; c'est un *jugement de valeur encodé*. Il peut être légitime ; il ne peut pas être clandestin. Une théorie de la preuve qui laisse la «scalarisation» se faire en silence, dans les pondérations d'un score composite ou dans le choix d'un seuil, laisse la gouvernance se faire à son insu. Valider une preuve sans expliciter le vecteur de perte qu'elle sert, et le contexte qui en fixe les poids, c'est valider on ne sait quoi pour on ne sait qui.

V. Le domaine : interne fois externe

Le domaine d'applicabilité, je l'ai emprunté à la validation des modèles structure-activité, où il est bien outillé : on sait y construire la région de fiabilité d'un modèle par l'enveloppe des covariables, l'enveloppe convexe du support, les distances de Mahalanobis ou les valeurs de levier, les zones de forte densité (voir la comparaison de Sahigara et collègues, *Molecules*, 2012). Ces méthodes sont solides. Elles ont un défaut : elles ne mesurent qu'une face du domaine. Un domaine d'applicabilité clinique en a deux, et chacune se dédouble.

Il y a une face *interne*, propre à la relation modélisée. Elle comprend le domaine *statistique* (la région de l'espace des covariables où la distribution observée soutient l'inférence) et le domaine *clinique* (l'ensemble des situations où la relation a un sens physiopathologique, qui peut être plus étroit que le support statistique, car une corrélation apprise n'a pas toujours de portée hors d'un mécanisme).

Il y a une face *externe*, propre au contexte qui entoure la décision. Elle comprend le domaine *temporel* (la fenêtre pendant laquelle les conditions de génération restent comparables à celles de l'usage) et le domaine *organisationnel* (le cadre de pratiques, de recommandations et de contraintes dans lequel la décision est prise). On peut donc écrire le domaine comme un produit : $\Delta = \textit{interne} \times \textit{externe}$, l'interne réunissant le statistique et le clinique, l'externe réunissant le temporel et l'organisationnel. La lecture y gagne : les deux faces internes décrivent ce que la preuve *est*, les deux faces externes décrivent le monde dans lequel on veut s'en servir.

Un exemple tranche la question. Soit une cohorte parfaitement dans son domaine interne (mêmes covariables, même distribution, même mécanisme). Même population, même traitement. Entre-temps, la Haute Autorité de Santé modifie ses recommandations sur l'indication concernée. La décision optimale change, alors qu'aucune covariable n'a bougé. La preuve est restée dans son domaine interne et est sortie de son domaine externe, et cette sortie suffit à l'invalider pour la décision visée. Réduire le domaine à sa face statistique, c'est croire qu'on a fermé la porte parce qu'on a tiré le verrou du haut.

VI. La preuve est un processus

Les deux faces externes du domaine ont un point commun : elles bougent. Les fenêtres se ferment, les recommandations se révisent. C'est pourquoi le paramètre T n'est pas un raffinement mais une nécessité, et c'est lui qui achève de transformer la nature de l'objet que nous manipulons.

Une preuve n'est pas un état stable. Elle est valide à une date, sous des conditions qui ont une durée de vie. Un modèle déployé voit la population qu'il sert s'éloigner lentement de celle sur laquelle il a été ajusté : le phénomène est si bien identifié que les régulateurs

ont cessé de l'ignorer, j'y reviens. Une identification causale jugée impossible hier peut devenir possible demain avec une nouvelle source, rouvrant une question close. Une recommandation se renverse. Dans chacun de ces cas, la preuve ne devient pas *fausse* : elle *expire*. La distinction est cruciale. Une preuve fausse était mal établie ; une preuve expirée était bien établie et a cessé de s'appliquer. Les confondre, c'est soit jeter ce qui valait, soit conserver ce qui ne vaut plus.

D'où une reformulation abrupte : *hors de son domaine, une preuve n'est pas fausse : elle est expirée*. Et une preuve qui peut expirer n'est pas un objet ; c'est un processus, avec un début de validité, une fenêtre, et une condition de fin. On ne certifie pas un processus une fois pour toutes. On l'engage, on le surveille, on le renouvelle. Le vocabulaire du certificat (délivré, acquis, définitif) est inadapté. Celui qui convient relève de l'engagement à terme, et c'est vers lui que tout ce qui précède conduit.

VII. Ce qu'un générateur ne peut rendre inférable

Avant d'y arriver, il faut traiter une question que la montée des modèles génératifs rend brûlante, et le faire avec une précision dont dépend la crédibilité de tout l'édifice. Un générateur (une population synthétique, un jumeau numérique, une cohorte simulée) peut-il produire du contenu qui n'était pas dans ses données ? La réponse naïve est non, et elle est juste pour de mauvaises raisons, ce qui la rend dangereuse.

J'éviterai ici le mot *information*, parce qu'il est si chargé qu'il invite la querelle plutôt qu'il ne la tranche. Parlons de *contenu identifiable* : ce qu'un ensemble de contraintes (données plus hypothèses structurelles) permet de distinguer.

- *Premier lemme* : tout contenu absent des contraintes du problème est non identifiable. C'est le cœur de la théorie de l'identification, telle que l'inférence causale moderne l'a formalisée, de l'émulation d'essai cible de Hernán et Robins (*American Journal of Epidemiology*, 2016) au calcul d'identifiabilité de Pearl. On ne peut estimer que ce que les données, sous des hypothèses déclarées, permettent de distinguer.
- *Second lemme* : ce qui n'est pas identifiable n'est pas récupérable par génération, car un générateur n'a accès à rien d'autre que ces mêmes contraintes. *Théorème de conservation* : aucun générateur ne rend identifiable un contenu absent des contraintes.

Cette formulation résiste là où la précédente prêtait le flanc. Un théoricien pouvait objecter qu'une représentation latente fait apparaître une structure jamais observée explicitement, qu'un modèle rend exploitable une régularité que personne n'avait su nommer, et il avait raison, tant qu'on parlait d'« information ». Mais ce que la représentation latente exhibe n'est pas un contenu *nouveau* ; c'est un contenu déjà

identifiable en principe à partir des contraintes, et que le générateur *rend inférable*, c'est-à-dire opérationnel. **Le verbe juste n'est donc pas *créer* mais *rendre inférable***.

Un générateur ne crée pas de contenu ; il transforme un contenu implicite, déjà identifiable dans les contraintes et le biais inductif du modèle, en contenu mobilisable pour une décision.

La conséquence pour la pratique est à la fois exigeante et libératrice. Exigeante, parce qu'elle interdit la promesse vendeuse selon laquelle on génère des patients là où il n'y a pas de données : on ne génère jamais que dans l'ombre portée de ce qu'on a déjà. Libératrice, parce qu'elle assigne au générateur une valeur réelle et défendable : rendre calculable une structure latente, explorer la région des plausibles que les contraintes autorisent, à un coût sans commune mesure avec celui d'un recrutement. Confondre les deux, prendre la mise-en-inférabilité pour une création de contenu, est l'exacte symétrie de l'erreur du sceptique qui prend l'échantillon observé pour la réalité. Le naïf de la donnée et le naïf du générateur se tiennent par la main, et ni l'un ni l'autre ne le sait.

VIII. Décider et réduire l'incertitude

J'ai annoncé, en introduisant la perte, qu'une preuve a deux fonctions, non une. Si une décision est une règle qui transforme de l'information en acte, alors une preuve sert à deux choses distinctes :

- Choisir l'acte,
- Et juger s'il vaut la peine d'en savoir davantage avant de choisir.

La littérature de l'évaluation des technologies de santé a depuis longtemps identifié qu'une preuve remplit au moins deux fonctions distinctes. La première consiste à soutenir une décision présente. La seconde consiste à déterminer si l'incertitude restante justifie encore de produire de nouvelles connaissances. Cette seconde fonction est précisément ce que quantifie l'analyse de la valeur de l'information (Value of Information, VOI). La valeur attendue d'une information parfaite (EVPI) mesure le bénéfice maximal que procurerait la disparition complète de l'incertitude avant de décider ; la valeur attendue d'une information d'échantillon (EVSI) estime le gain attendu d'une étude réaliste de taille donnée. Depuis les travaux fondateurs de Claxton jusqu'aux recommandations méthodologiques de l'ISPOR, la question n'est donc plus seulement « quelle décision est optimale aujourd'hui ? », mais également « combien vaut-il la peine d'investir pour réduire l'incertitude avant de confirmer cette décision ? ».

Cette distinction révèle une propriété souvent ignorée des cohortes synthétiques. Une cohorte peut parfaitement préserver la décision optimale sous la fonction de perte retenue tout en déformant la structure d'incertitude qui fonde la valeur d'une recherche future. Autrement dit, elle peut conduire au bon choix aujourd'hui tout en suggérant, à

tort, qu'une étude supplémentaire est inutile, ou inversement qu'elle est indispensable. Elle est alors substituable pour décider, mais non pour apprendre.

La différence est fondamentale. La décision optimale dépend uniquement de la position du minimum de perte attendu. La valeur de l'information dépend, elle, de la possibilité que ce minimum se déplace lorsque de nouvelles observations deviennent disponibles. Deux distributions peuvent ainsi conduire exactement à la même décision tout en induisant des EVPI ou des EVSI très différentes. Préserver la décision n'implique donc pas préserver la valeur de l'information. Les deux propriétés relèvent de contraintes mathématiques distinctes.

Nous proposons ainsi de distinguer deux niveaux de substituabilité.

1. La première, que l'on peut qualifier de *substituabilité décisionnelle*, exige que les données synthétiques conduisent à la même règle de décision optimale que les données réelles sous un contexte décisionnel donné.
2. La seconde, plus exigeante, correspond à une *substituabilité informationnelle* : les données synthétiques doivent également préserver la valeur des informations susceptibles de modifier cette décision. Une preuve qui satisfait uniquement la première propriété demeure intrinsèquement myope. Elle permet de choisir correctement aujourd'hui, mais ne permet plus d'évaluer correctement si l'on devrait continuer à chercher demain.

Cette distinction ouvre un prolongement naturel des cadres actuels de validation. Il ne suffit plus de démontrer que les décisions produites à partir des données synthétiques reproduisent celles obtenues sur les données réelles. Il devient également nécessaire de vérifier que les analyses de valeur de l'information conduisent aux mêmes priorités de recherche, aux mêmes arbitrages entre décision immédiate et acquisition de nouvelles données, et aux mêmes conclusions sur l'utilité marginale d'études complémentaires. Une cohorte synthétique pleinement substituable devrait préserver non seulement l'action optimale, mais également l'économie de l'apprentissage qui entoure cette action.

IX. Substituable est un degré, et une date

Tout ce qui précède a manipulé la substituabilité comme un état : une source remplace une autre, ou ne la remplace pas. C'est une commodité qu'il faut abandonner, car elle est fausse.

La substituabilité est graduelle. Une cohorte préserve certaines propriétés mieux que d'autres ; elle soutient certaines décisions et trahit les autres. Le bon objet n'est pas un booléen mais un degré (appelons-le σ, entre zéro et un), ou, plus utilement, un *profil* : substituable pour telle famille de décisions, pas pour telle autre. Dire d'une population synthétique qu'elle est « substituable à 0,82 » n'a de sens qu'assorti de sa décision, de

son vecteur de perte, de son domaine et de l'intervalle d'incertitude qui entoure ce chiffre, incertitude double, car elle combine l'aléa d'échantillonnage, qui affecte toute cohorte finie, et l'incertitude de génération, qui tient à ce que le générateur est lui-même estimé sur des données finies. La prédiction conforme (Vovk, Gammerman et Shafer, 2005 ; pour une introduction, Angelopoulos et Bates, 2021) offre un cadre sans hypothèse distributionnelle pour produire de tels intervalles avec des garanties non asymptotiques.

Le passage du booléen au degré n'est pas une concession de modestie. C'est ce qui rend la substituabilité gouvernable. Un état binaire ne se négocie pas ; un degré assorti d'un profil et d'une incertitude se discute, se «seuille», s'audite. On peut décider qu'une substituabilité de tel niveau suffit pour une décision exploratoire et pas pour une décision confirmatoire. *Substituable n'est pas un état : c'est un degré, et une date*. Et un degré daté, attaché à un usage déclaré, porte déjà tous les traits d'un engagement à terme.

X. Une décision oncologique : dérouler la structure

Le moment est venu de cesser d'illustrer et de dérouler la structure sur un cas. Toute la théorie repose sur la contextualisation ; elle doit donc se laisser dérouler sur une décision réelle, du haut jusqu'au degré de substituabilité. Prenons une décision oncologique précise, et suivons la chaîne sans en sauter un maillon. Les magnitudes empiriques qui suivent sont marquées comme à documenter : la théorie fixe la *structure* de la démonstration, non les chiffres, qui relèvent de la mesure et non de l'argument.

La décision. Chez une patiente atteinte d'un cancer du sein hormonodépendant, sans surexpression de HER2, faut-il intensifier l'hormonothérapie adjuvante ? La décision est une règle δ qui, à un profil de patiente, associe une action dans un ensemble fini d'options thérapeutiques. Ce profil n'est pas un point dans un texte : c'est un vecteur sur plusieurs centaines de variables tabulaires, et c'est pourquoi on ne raisonne pas sur lui comme sur du langage, pas plus qu'on ne raisonne sur un patient porteur d'une mutation comme BRAF V600E en lui appliquant l'intuition d'un modèle de mots (LLM).

Le vecteur de perte, contextuel. **L** met en balance le bénéfice attendu sur la survie sans récidive, la toxicité, l'équité d'accès, le coût, l'acceptabilité par la patiente et la conformité réglementaire. Les pondérations ne sont pas universelles : pour une patiente âgée avec comorbidités, le poids de la toxicité monte ; pour une patiente jeune, celui de la survie sans récidive domine. La perte est donc fixée *dans* le contexte de la décision, pas avant lui.

Le domaine, interne fois externe. En interne : la région statistique des profils effectivement représentés dans les données mobilisées, et la pertinence clinique du sous-type hormonodépendant HER2-négatif. En externe : la fenêtre temporelle des protocoles considérés, et le cadre organisationnel des recommandations en vigueur au moment de la décision. Une patiente dont le profil tombe hors de la région statistique sort

du domaine interne ; un changement de recommandation sort du domaine externe. Les deux invalident, pour des raisons opposées.

Le temps. La validation est datée. Elle tient tant que les distributions ne dérivent pas et que les recommandations ne sont pas révisées ; elle expire sinon. La preuve mobilisée porte donc une condition de fin, pas seulement une condition d'emploi.

La validation. On mobilise une cohorte (réelle, synthétique, ou un jumeau visant par exemple le profil de toxicité, terrain qu'explore un dispositif comme ToxTwin) pour estimer ce dont δ a besoin. La question n'est pas « cette cohorte est-elle réelle ? » mais « préserve-t-elle, sur le domaine déclaré, les propriétés nécessaires à δ sous **L** ? » On la teste par *concordance de décision* : produit-elle, là où on l'emploie, les mêmes actions que la cohorte de référence, pondérées par la perte, et non par simple proximité distributionnelle, qui peut être excellente sans que la décision soit préservée. On la soumet aussi à une revue par des cliniciens experts, car la vraisemblance maximale lisse les anomalies, et l'anomalie (le répondeur atypique, la signature rare) est souvent l'objet le plus précieux, celui qu'aucune divergence de Kullback-Leibler ne signale.

Le degré de substituabilité. Le résultat n'est pas un verdict mais une mesure graduée. Nous proposons de représenter la substituabilité par un indice continu σ, défini relativement à une règle de décision δ, une fonction de perte L et un domaine d'applicabilité explicitement déclarés. Une formalisation naturelle consiste à définir σ comme un regret décisionnel normalisé :

$$\delta = 1 - \frac{E_\theta . [L(\delta_S, \theta\ ) - L(\delta_R, \theta\ )\ ]}{L_{max}}$$

Avec :

- δ_R: règle de décision obtenue à partir des données réelles,
- δ_S: règle de décision obtenue à partir de la cohorte synthétique,
- $L(\delta, \theta)$: perte associée à la décision δ lorsque l'état réel est θ,
- E_θ : espérance sur la distribution des états du monde (ou des patients),
- L_{max} : Regret maximal retenu comme référence de normalisation

Si on utilise la définition du risque bayésien comme étant :

$$R(\delta) = E_\theta[L(\delta, \theta)]$$

Alors :

$$\delta = 1 - \frac{R(\delta_S) - R(\delta_R)}{L_{max}}$$

Ainsi, ($\delta = 1$) signifie que la substitution n'induit aucune perte décisionnelle moyenne ; à mesure que le coût clinique, économique ou organisationnel des divergences augmente, σ décroît vers zéro. L'indice mesure donc non la proximité statistique entre deux distributions, mais le coût attendu de remplacer l'une par l'autre pour la décision considérée.

Cette définition rend la substituabilité intrinsèquement contextuelle. Une même cohorte peut présenter une valeur élevée de σ pour une décision d'intensification thérapeutique et une valeur beaucoup plus faible pour une décision de désescalade, parce que les conséquences des erreurs et donc la fonction de perte diffèrent. Il n'existe dès lors aucune certification universelle d'une cohorte, seulement des degrés de substituabilité relatifs à un usage explicitement défini. Toute estimation de σ devrait être rapportée avec son intervalle d'incertitude, son domaine de validité et son profil décisionnel. Une extension naturelle consiste à vérifier également que la cohorte préserve la valeur de l'information (EVPI, EVSI) associée à cette décision, afin d'évaluer non seulement sa capacité à reproduire les choix présents, mais aussi sa capacité à préserver la valeur des recherches futures.

Que la même chaîne s'applique mot pour mot à une cohorte de données de vie réelle (qu'on déclare sa décision, sa perte contextuelle, son domaine à deux faces, sa date, et qu'on mesure son degré de substituabilité) rend l'essentiel explicite, et vérifie au passage le critère de composabilité de la section III : la validation d'une sous-décision (estimer la toxicité) s'assemble avec celle de la décision englobante (intensifier ou non) parce que toutes deux sont indexées par le même appareil. La doctrine n'est pas une apologie du synthétique. Le synthétique n'en est que la première démonstration, parce qu'il rend visible, par contraste, ce que la donnée réelle dissimulait derrière son évidence.

XI. La santé n'est que le premier terrain

Une théorie se juge aussi à ce qu'elle éclaire au-delà de son point de départ. Rien dans la structure *D*, **L**, *Δ*, *T* n'est propre à la médecine. Partout où une preuve computationnelle informe une décision sous contrainte, les mêmes impossibilités guettent et la même résolution s'impose.

En finance, un modèle de risque calibré sur un régime de marché expire quand le régime change : son domaine temporel est court, et le confondre avec une propriété stable est une source classique de déconvenue. En aéronautique, une simulation de certification n'est valide que pour l'enveloppe de vol déclarée : le domaine y est une notion d'ingénierie avant d'être statistique. En cybersécurité, un détecteur entraîné sur des attaques connues n'a pas de garantie hors de leur distribution, et l'extrapolation y est précisément l'adversaire. Dans chacun de ces champs, la preuve est relative à une décision, sous une perte contextuelle, dans un domaine à deux faces, à une date.

Ce transfert n'est pas une extrapolation gratuite : il satisfait le critère de transférabilité que la métathéorie exigeait, à condition d'en déclarer les conditions, ce que je viens de faire pour trois domaines. Ces transferts restent, à ce stade, des conjectures de portée : la théorie en réclame l'épreuve domaine par domaine, elle ne la fournit pas ici. Présenter la santé comme le premier terrain plutôt que comme le périmètre n'est donc pas une ambition de façade ; c'est la conséquence directe d'une théorie qui, si elle est juste, ne saurait s'arrêter à la frontière d'une discipline. La santé en offre seulement la version la plus dense, parce que les enjeux y sont vitaux, les sous-populations nombreuses et les arbitrages de perte les plus moralement chargés.

XII. Pourquoi cette théorie était difficile à voir

Un relecteur ne posera plus, à ce stade, de question méthodologique. Il posera une question d'histoire : *qui a déjà dit quelque chose de semblable ?* C'est une bonne question, et la pire réponse serait de feindre l'originalité absolue. La vérité est que rien ici n'est neuf pris isolément. Tout est neuf dans l'assemblage.

Chacun des champs convoqués détenait déjà un fragment du problème, et ne voyait que le sien.

- Wald avait la décision et le risque, mais sous une perte scalaire, et sans un mot sur le domaine ni sur le temps.
- La famille GRADE avait la qualité graduée de la preuve, mais traitée comme une propriété d'un corpus, là où nous la lisons comme un degré relatif à une décision.
- La validation des modèles structure-activité, codifiée par l'OCDE, avait le domaine d'applicabilité, mais géométrique et statique.
- Pearl, Hernán et Robins avaient l'identifiabilité (la frontière de ce que les données peuvent soutenir), mais pour des estimandes causales, non comme une loi générale de conservation de la preuve.
- La tradition de la valeur de l'information avait la valeur de réduire l'incertitude, mais comme outil de priorisation de la recherche, non comme seconde fonction de toute preuve.
- La réglementation récente, enfin, avec les plans de contrôle des changements prédéterminés de la FDA (autorité statutaire inscrite à la section 515C du Federal Food, Drug, and Cosmetic Act par le FDORA de 2022, précisée par la guidance finale de décembre 2024 et le projet de gestion du cycle de vie de janvier 2025).
- L'*intended purpose* du règlement européen sur l'intelligence artificielle, avait la validation en cycle de vie, mais comme pratique sans la théorie qui l'explique.

L'apport n'est donc pas une création ; c'est une intégration. La théorie hérite la décision de Wald, le domaine de la tradition QSAR, la frontière d'identifiabilité de l'inférence causale, la valeur de l'incertitude de l'évaluation des technologies de santé, l'unité

d'usage en cycle de vie de la réglementation, la gradation de la qualité de GRADE. Elle conserve le cœur de chaque fragment, relu relationnellement. Elle abandonne l'hypothèse que ces fragments partageaient sans le savoir, celle qui fait de la validité, de la qualité ou du domaine des propriétés d'un objet. Et elle rend enfin explicable une coïncidence troublante : pourquoi ces cadres, développés indépendamment, n'ont cessé de refaire les mêmes gestes : déclarer un usage, un critère, une limite, une fenêtre. Ils le faisaient parce qu'ils touchaient, chacun, une face différente d'une même relation.

Si la théorie était difficile à voir, c'est qu'elle habitait l'entre-deux des disciplines, et qu'aucune n'avait de raison de lever les yeux de son propre fragment.

XIII. De la propriété émergente à la promesse, et de la promesse au contrat

Reste à nommer ce qui émerge du parcours, et à le nommer dans le bon ordre, car l'ordre est ici la pensée même.

Le fait premier, celui dont tout le reste découle, est que la validation n'appartient à aucun des termes en présence. Elle n'est pas une propriété de la donnée, ni du modèle, ni même de la décision. Elle est une *propriété émergente d'une relation* (une source de preuve, une décision, un contexte), indexée par le temps et révisable. Elle naît de leur rencontre et meurt avec elle ; aucun des trois ne la possède, comme aucun des deux versants d'une vallée ne possède la rivière. C'est l'énoncé réellement neuf, et c'est de lui qu'il faut partir, car les deux notions qui suivent n'en sont que des conséquences. Que la validation soit dite *émergente* n'est pas une affirmation métaphysique invérifiable : c'est un cadre interprétatif dont la charge réfutable repose tout entière sur les thèses qui précèdent. Qu'on établisse une seule validité indépendante de la décision, de la perte et du domaine, et l'émergence tombe avec elles.

1. Première conséquence : si la validité émerge d'une relation indexée par le temps, alors une preuve n'affirme jamais un fait inconditionnel. Elle énonce une *promesse conditionnelle* : « si la décision est celle-ci, si la perte est celle-là, si l'on reste dans ce domaine, et tant que dure cette fenêtre, alors la source soutient l'action à ce degré ». La preuve, correctement comprise, a la forme logique d'une implication gardée, pas celle d'un constat. Et la promesse conditionnelle n'est pas propre à la santé : elle est la forme de toute garantie en mathématiques, où un théorème vaut sous ses hypothèses ; en physique, où une loi vaut dans son régime ; en apprentissage automatique, où une borne vaut sous sa distribution ; en réglementation, où une autorisation vaut pour un usage. Voir la preuve comme une promesse conditionnelle, c'est la rendre comparable d'un domaine à l'autre.
2. Seconde conséquence : le contrat. Lorsqu'une promesse conditionnelle doit être tenue entre des parties (un fabricant et un régulateur, un fournisseur de preuve et

un décideur), elle se formalise socialement. Elle devient un *contrat d'usage* : un engagement de substituabilité, gradué, daté, à clauses testables et à domaine déclaré, honoré tant que ses conditions tiennent et renégocié dès qu'elles cèdent. Le contrat n'est donc pas le concept fondamental ; c'est le cas particulier où la promesse conditionnelle devient juridiquement ou organisationnellement explicite. Et c'est, on l'a vu parmi les fragments, la direction que la réglementation a déjà prise : le contrôle des changements en cycle de vie chez la FDA, l'*intended purpose* du règlement européen, font déjà de l'usage déclaré, et non de l'objet seul, l'unité de certification. Le monde réglementaire formalise déjà des promesses conditionnelles sans avoir la théorie qui le dit. Cette note propose cette théorie.

Il faut conclure sur la symétrie qui aura traversé tout le texte, parce qu'elle en est la leçon.

- ***Il y a une naïveté qui prend l'échantillon observé pour la réalité***,
- Il y a une naïveté qui prend le générateur pour une source de contenu,
- Et il y a, plus profonde et plus partagée, une naïveté qui prend la preuve pour un objet, quand elle n'a jamais été qu'une relation, et la relation pour un état, quand elle n'a jamais été qu'une promesse.

Renoncer à la première sans renoncer aux suivantes, c'est changer d'idole. *Une preuve n'est pas un objet qu'on certifie ; c'est une promesse conditionnelle qu'on tient*, et un contrat, quand il advient, n'est que le nom social de cette promesse, pour des parties, pour un objet, et pour un temps.

Références

1. Wald A. *Statistical decision functions*. New York (NY): John Wiley & Sons; 1950.
2. Organisation for Economic Co-operation and Development. *Guidance document on the validation of (Q)SAR models*. OECD Series on Testing and Assessment No. 69. Paris: OECD Publishing; 2007.
3. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-926. doi:10.1136/bmj.39489.470347.AD.
4. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332. doi:10.1136/bmj.c332.
5. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. BMJ. 2015;350:g7594. doi:10.1136/bmj.g7594.
6. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of different approaches to define the applicability domain of QSAR models. Molecules. 2012;17(5):4791-4810. doi:10.3390/molecules17054791.
7. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183(8):758-764. doi:10.1093/aje/kwv254.
8. Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669-688. doi:10.1093/biomet/82.4.669.
9. Vovk V, Gammerman A, Shafer G. *Algorithmic learning in a random world*. New York (NY): Springer; 2005.
10. Angelopoulos AN, Bates S. *A gentle introduction to conformal prediction and distribution-free uncertainty quantification*. arXiv [Preprint]. 2021. arXiv:2107.07511.
11. Rothery C, Strong M, Koffijberg HE, Basu A, Ghabri S, Knies S, et al. Value of information analytical methods: Report 2 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force. Value Health. 2020;23(3):277-286. doi:10.1016/j.jval.2020.01.004.
12. U.S. Food and Drug Administration. *Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device*

software functions: guidance for industry and Food and Drug Administration staff. Silver Spring (MD): U.S. Food and Drug Administration; 2024.