Même performance au benchmark, soin opposé

La validation montre qu'un système pouvait être utilisé. La gouvernance établit qu'il peut continuer à l'être

En juin 2026, une étude parue dans Nature Medicine rapportait que trois grands modèles de langage généralistes de frontière, GPT-5.2, Gemini 3.1 Pro et Claude Opus 4.6, surpassaient deux outils d'IA clinique spécialisés, OpenEvidence et UpToDate Expert AI, sur des benchmarks médicaux et sur un ensemble de 100 requêtes réelles de médecins recueillies dans un unique déploiement clinique en conditions réelles (Vishwanath et al., Nature Medicine, 12 juin 2026). Le titre est puissant, ce qui le rend dangereux, et il devrait être lu avec la discipline que défend le reste de cet article : avant de vouloir dire quoi que ce soit, il faut connaître les tâches, les systèmes, les modèles, le critère d'évaluation, la définition de la « vie réelle », et savoir si les usages prévus étaient seulement commensurables. Les outils comparés étaient des systèmes commerciaux spécialisés, non des dispositifs porteurs d'une autorisation réglementaire, de sorte que la glose populaire selon laquelle « l'IA généraliste bat l'IA clinique » introduit déjà en fraude un saut inférentiel que cet article existe pour refuser.

Le champ a fait la démonstration lui-même dans le trimestre : le 3 septembre 2026, Nature Medicine a publié un Matters Arising (Beaulieu-Jones et Nemati) soutenant que le caractère limité des benchmarks contraint les conclusions de l'étude, précisément lorsqu'on les étire vers l'achat, le remboursement ou la régulation, trois décisions différentes prises par trois autorités différentes. Les auteurs de l'étude ont répliqué le même mois, en maintenant leur conclusion dans leurs propres conditions de déploiement, ce qui est justement le point : le désaccord ne porte pas sur les chiffres, il porte sur la distance qu'un benchmark parcourt vers une décision. Prenons donc la conclusion au sens étroit. Elle ne montre pas que l'IA clinique échoue. Elle montre que des systèmes classés à l'identique sur un benchmark peuvent se comporter très différemment une fois qu'une décision réelle, et un décideur réel, entrent dans la pièce. Que deux systèmes de même performance au benchmark puissent produire un soin opposé, voilà la porte. La question intéressante n'est pas pourquoi. C'est ce que cela nous apprend sur la manière dont l'IA devrait être gouvernée.

Voici la réponse en une ligne, avant l'argument qui la mérite : la validation établit qu'un système pouvait être utilisé au moment où il a été évalué ; la gouvernance doit établir qu'il peut continuer à l'être. Une autorisation ne certifie pas un nombre, elle accepte un passage d'une performance observée vers une revendication, et d'une revendication vers une décision, et ce passage doit être entretenu, pas seulement accordé. Trois choses doivent être tenues distinctes, parce que le discours sur l'IA les confond et le paie. La

crédibilité épistémique est une propriété d'une revendication au regard de ses preuves : que pouvons-nous raisonnablement croire. L'admissibilité décisionnelle n'est pas du tout une propriété du modèle, c'est une relation : une revendication n'est jamais admissible dans l'abstrait, seulement pour une décision spécifiée, sous une autorité, une perte et une conséquence spécifiées. L'autorisation institutionnelle est encore une troisième chose, l'acte par lequel un régulateur, un payeur ou un service clinique permet l'usage. Trois objets, trois autorités, trois standards de preuve. La performance n'implique pas la validité, la validité n'implique pas l'utilité, l'utilité n'implique pas une décision, et aucune d'elles, établie une fois, n'implique l'admissibilité plus tard. Cette dernière relation est le sujet de cet article, et c'est celle que ni un benchmark ni une validation initiale ne peuvent atteindre.

Quatre choses qu'un score n'est pas

Séparons, sans détour, quatre revendications que le mot « performance » amalgame. La discrimination n'est pas la calibration. La calibration n'est pas l'utilité clinique. L'utilité clinique n'est pas le bénéfice pour le patient. Un modèle peut être excellent sur la première et sans valeur sur la dernière, et chaque marche exige sa propre justification. La plupart des disputes sur l'IA clinique sont des disputes sur la marche où l'on se tient sans le dire.

L'étude de Nature Medicine illustre cette compression directement. Son verbe-titre, « surpasser », rassemble quatre dimensions graduées que les cliniciens ont en réalité notées séparément, exactitude, complétude, sécurité et clarté, et les systèmes ne se sont pas séparés de la même façon sur chacune : les modèles de frontière ont davantage creusé l'écart sur la clarté que sur l'exactitude clinique, un W de Kendall de 0,292 contre 0,141, et aucune différence significative n'a été trouvée sur le contenu nuisible ou l'hallucination. « Meilleur », ici, veut surtout dire « plus clair », ce qui n'est pas rien et n'est pas la même revendication. Un outil a par ailleurs refusé près d'un cinquième des requêtes là où les modèles de frontière n'en refusaient presque aucune, si bien qu'une politique de refus, un choix de conception, se replie silencieusement dans le nombre qui se lit comme une performance. Un score est rarement la mesure d'une seule chose.

Commençons alors par l'estimation, puisqu'elle règle la question sans aucun recours au déplacement de distribution. Une estimation ponctuelle n'est pas une distribution a posteriori : le même 0,87 peut se loger dans un intervalle étroit ou large, et l'incertitude n'est pas d'une seule espèce. L'incertitude d'échantillonnage peut souvent être réduite par des observations plus représentatives ; l'incertitude de mesure, de transport et structurelle exige en général une preuve différente, pas simplement davantage de preuve. C'est cette seconde famille qu'un chiffre unique masque le plus complètement, et dans le régime à haute dimension et faible effectif, courant en apprentissage automatique clinique, cette distinction devient décisive.

Nos travaux ISPOR sur le CBNPC métastatique rare BRAF V600E fournissent une illustration utile. À partir d'une cohorte en vie réelle de seulement 184 patients, la génération par copule gaussienne a atteint un score de fidélité agrégé de 0,954 (IC bootstrap à 95 % de 0,941 à 0,957) et 99,7 % d'utilité pour l'apprentissage automatique. Pourtant, ces chiffres globaux rassurants n'impliquaient pas l'identité au niveau des estimations cliniquement significatives : la survie globale à 12 mois différait de 8,9 points de pourcentage, et le rapport de risque associé à un indice de performance ECOG de 2 ou plus était de 2,75 dans la cohorte réelle contre 1,21 dans la cohorte synthétique. La direction des effets pronostiques est restée stable sur 100 réplications synthétiques, mais leur magnitude, non. Un score agrégé élevé peut donc coexister avec une incertitude substantielle précisément au niveau où une décision clinique peut dépendre de l'estimation.

La calibration rend le point canonique. Prenons deux modèles de discrimination identique au benchmark, calibrés dans la population d'évaluation. Pour un patient, le premier renvoie une probabilité de 0,32 et le second de 0,18. À un seuil thérapeutique de 0,25, le premier dit traiter et le second s'abstenir. Deux sens de « score » sont en jeu, et le titre tient à leur différence : la métrique agrégée est la même, la prédiction individuelle ne l'est pas. Même performance au benchmark, soin opposé, pour le même patient. La calibration est aussi dépendante de la distribution, si bien qu'elle s'établit en validité interne et doit être préservée dans le transport, ce qui est déjà un premier signe qu'une propriété prouvée une fois n'est pas possédée pour toujours.

Une distinction de plus, la plus souvent sautée. Un risque prédit n'est pas un effet du traitement : le patient au risque le plus élevé n'est pas nécessairement celui qui gagne le plus à agir. C'est là que la crédibilité et l'admissibilité se séparent. La revendication « ce patient a 38 pour cent de risque de progression » peut être parfaitement crédible et ne rien dire encore sur l'opportunité de traiter. La crédibilité se gagne contre des preuves ; l'admissibilité se gagne contre une décision.

Ce qui transforme une preuve en décision, et ce qui peut la défaire

Si un score n'est pas la preuve, qu'est-ce qui l'est ? Le passage des données à la décision traverse des opérations, mesure, identification, estimation, transport, action, et le mot utile pour ce qui autorise chaque étape est garantie, au sens de Toulmin. Une garantie est simplement la raison pour laquelle une preuve est autorisée à soutenir une revendication : la preuve, plus les hypothèses requises pour faire l'inférence. Preuve et hypothèses ne sont pas la même chose, ce qui compte dans un article sur la charge des hypothèses : la preuve soutient une revendication, une hypothèse autorise une inférence que la preuve seule ne porte pas.

Trois garanties couvrent la chaîne. La garantie interne conduit de la mesure à une estimation fiable : construit, mesure saine, identification, incertitude d'estimation, calibration dans la population d'évaluation. Elle peut échouer sans aucun déplacement, par un critère mal mesuré, une fuite de données, une multiplicité non corrigée, ou une cohorte trop petite pour calibrer. La garantie de transport conduit l'estimation vers la population et le système cibles, préservation de la calibration comprise. La garantie d'actionnabilité, d'ordinaire laissée tacite, conduit d'une estimation transportée à une action : elle établit que l'usage de l'estimation change l'action préférée sous la règle de décision, l'utilité, les contraintes et le décideur pertinents. Le patient individuel, pour qui agir doit l'emporter sur ne pas agir, en est le cas le plus simple ; un payeur ou un système de santé optimise sous contrainte de rareté et de bénéfice populationnel.

Chaque hypothèse critique devrait être interrogée sur la manière dont elle peut être contestée, et toutes les hypothèses ne se contestent pas de la même façon. Certaines portent un réfutateur observable, une constatation qui les invaliderait directement, et peuvent devenir un signal de surveillance. Certaines n'admettent qu'un défi indirect, certaines un test de sensibilité, certaines une preuve externe, et certaines aucun test empirique du tout, seulement l'argument. La fragilité épistémique d'une revendication se lit à la part de son architecture qui repose sur des hypothèses faiblement testables, et prétendre que toute incertitude épistémique se convertit en une métrique de tableau de bord serait une nouvelle version de l'illusion que cet article combat. Une étude de validation, donc, ne s'oppose pas à la justification, elle en fait partie : une étude prospective et externe apporte de la preuve, elle n'abolit jamais les hypothèses de transport, elle les rend plus ou moins défendables. L'audit de cinquante-trois benchmarks médicaux à l'ACL 2026 montre à quelle fréquence la garantie interne échoue selon ses propres termes, avant même que le transport soit en question.

Non une checklist mais un argument, et pourquoi une même preuve autorise des décisions opposées

La manière honnête de présenter cela est comme une unification, non comme une invention, puisque trois traditions détiennent déjà des parts du terrain. Le contexte d'usage (Context of Use), central pour la doctrine de l'usage prévu et pour le cadre de crédibilité fondé sur le risque que la FDA a avancé pour les décisions relatives aux médicaments et aux produits biologiques, spécifie le domaine cible. L'analyse de courbe de décision (Decision Curve Analysis) intègre la perte et le seuil et transforme la performance en bénéfice net (Vickers et Elkin). La validation externe traite le déplacement de distribution. Aucune, seule, ne délivre la suffisance décisionnelle. Le contexte d'usage dit où la performance doit tenir ; la justification dit pourquoi la performance observée est une preuve crédible qu'elle tiendra. L'analyse de courbe de décision est un pont, non un rival : elle calcule le bénéfice net après avoir supposé que

ses probabilités sont valides et transportables, elle arrive donc après l'opération qu'elle présuppose. Le bénéfice net n'est pas la transportabilité.

Ce que l'architecture ajoute est un seuil dont la forme compte. Une justification n'a pas besoin d'être vraie, seulement assez crédible pour la conséquence, et la crédibilité d'un ensemble d'hypothèses n'est pas un scalaire à moyenner, parce que certaines sont non compensatoires : une hypothèse critique fausse n'est pas rachetée par cinq bien étayées. Les hypothèses ne sont pas non plus indépendantes ; certaines sont substituables, certaines complémentaires, certaines critiques seulement conditionnellement à une autre. L'architecture n'est donc pas une checklist mais un argument, à lire au mieux comme un graphe à deux types d'arêtes, celles qui soutiennent une revendication et celles qui l'attaquent, puisqu'un réfutateur s'oppose à une garantie. Cela renverse la posture de la surveillance. La gouvernance n'accumule pas seulement des preuves que le système fonctionne ; elle cherche des preuves capables de défaire la justification en vigueur. La question passe de « montrer qu'il fonctionne encore » à « chercher la raison pour laquelle l'autorisation ne devrait plus tenir », et lire le graphe est ce qui révèle les points uniques de défaillance épistémique, les hypothèses dont trop de revendications dépendent en silence.

C'est aussi pourquoi une même preuve peut autoriser des décisions opposées, et la forme générale importe plus que le slogan clinique. Une décision est une fonction de la preuve, de l'utilité, des contraintes et de l'autorité, non de la preuve seule. Une preuve épistémique identique avec des fonctions de perte différentes n'est pas le même problème de décision, si bien que deux acteurs rationnels peuvent choisir des actions opposées à partir d'un seul nombre. Le régulateur, le payeur et le clinicien en sont l'exemple type et ils ne décident pas de la même chose : le régulateur fixe la disponibilité, le payeur fixe l'accessibilité, le clinicien choisit le soin. Le Matters Arising nommait le même trio, mettant en garde contre le fait de porter un même benchmark vers l'achat, le remboursement et la régulation comme s'il s'agissait d'une seule question. Même preuve, décisions opposées, chacune admissible pour sa propre décision, autorité et perte. Cela se généralise aussitôt hors de la médecine, à un modèle de crédit lu par un directeur des risques et par un régulateur, à un outil de tri lu par un recruteur et par un tribunal, à un détecteur lu par un analyste et par un auditeur. Des issues opposées à partir d'un seul modèle ne sont pas toujours une erreur ; souvent ce sont des réponses correctes à des questions différentes posées à la même estimation.

Contrats de transport, autorité, et gouvernance dans le temps

Le transport a besoin d'un objet assez précis pour être falsifiable, sans quoi il se dissout dans le truisme que le contexte compte. Cet objet est l'invariant, écrit comme un contrat

au sens de l'ingénierie, un ensemble explicite de conditions sous lesquelles une revendication reste admissible, non un instrument juridique. Le transport devient opérationnel en nommant un ensemble minimal de propriétés dont la préservation est nécessaire, sous les hypothèses causales et de mesure de la revendication, pour que la preuve reste pertinente dans le cadre cible. Nécessaire, non suffisante : même si les invariants nommés tiennent, une variable omise peut briser la revendication, et c'est pourquoi il s'agit d'ingénierie et non d'un théorème. Le contrat énonce la propriété, la tolérance, la preuve qu'elle tient, le moniteur qui la surveille, l'autorité qui détient la décision d'agir, et la cadence à laquelle la preuve elle-même expire. Un seuil franchi, de surcroît, n'est pas une décision mais le début d'une réponse graduée, alerte, investigation, usage restreint, revalidation, suspension, chaque étape détenue par quelqu'un. C'est le point porteur pour quiconque doit faire tourner un tel système : un seuil de surveillance sans autorité ni politique d'action, c'est de la télémétrie, pas de la gouvernance. Et deux systèmes également justifiés aujourd'hui peuvent différer nettement par ce qu'il en coûte de les maintenir justifiés, l'un reposant sur des hypothèses observables et vite testables, l'autre sur des hypothèses structurelles révisables seulement par étude externe ; le second porte davantage de dette épistémique, et son coût de cycle de vie est plus élevé pour la même revendication.

Le cadre réglementaire illustre le point multi-autorités, et c'est une illustration, non un fondement, puisque la doctrine tient quoi que fasse le calendrier. Le fait stable et structurel est celui-ci : lorsqu'un système d'IA est, ou est un composant de sécurité de, un dispositif médical exigeant l'évaluation d'un organisme notifié, ses obligations au titre de l'AI Act sont repliées dans l'évaluation de conformité MDR ou IVDR existante plutôt que menées comme une procédure distincte, et plusieurs autorités lisent alors la même preuve pour des pertes différentes. Le calendrier est contingent : les obligations générales haut-risque étaient fixées pour août 2026, tandis que pour l'IA embarquée dans des dispositifs médicaux régulés le Conseil a convenu en mars 2026 de repousser l'échéance plus loin, vers 2028. Les Guiding Principles of Good AI Practice in Drug Development conjoints de la FDA et de l'EMA, publiés en janvier 2026, sont ici un allié à demi juste, nommant la surveillance sur le cycle de vie et une approche fondée sur le risque, tout en s'arrêtant en deçà d'une justification inspectable, détenue et réfutable.

D'où la définition qu'il vaut la peine de retenir. La gouvernance de l'IA est le maintien contrôlé de l'admissibilité décisionnelle sous des preuves, des hypothèses, un contexte et des conséquences changeants. Sa fonction n'est pas conservatrice. Un bon système de gouvernance ne maintient pas toujours une garantie en vie ; parfois son devoir est d'y mettre fin. Il ne maintient pas les garanties en vie, il les garde réfutables.

Coda : le problème plus difficile, et ce que signifie une décision gouvernée

Un cas se tient au-delà même de celui-ci, et il fait l'objet d'un article compagnon, une phrase doit donc suffire ici. La doctrine, jusqu'ici, suppose que surveiller signifie observer si le monde a changé autour du modèle. Il existe un cas plus difficile : un système qui agit sur le processus qu'il observe change ce processus, la prédiction menant à l'action, l'action à l'issue, l'issue au jeu de données suivant. Dès lors que la prédiction change le traitement, la surveillance post-commercialisation d'une IA couplée à l'intervention devient un problème d'inférence causale, non un simple problème de suivi de performance, et il en va de même, plus vivement encore, lorsque la quantité décisive est un contrefactuel qui n'est jamais observé, comme avec les bras de contrôle synthétiques et les jumeaux numériques qui entrent aujourd'hui dans la preuve réglementaire. C'est là que la charge des hypothèses non testables pertinentes pour la décision est la plus grande, et c'est là que cet argument se poursuit ensuite.

Ce qui nous ramène à Elsa, comme illustration et non comme preuve. Le régulateur qui déploie un outil génératif enclin à inventer des études rencontre, dans ses propres instruments, la confusion qu'il doit surveiller dans ceux qu'il évalue : une sortie qui se lit comme une preuve n'est pas pour autant une preuve, et la fluidité n'est pas une garantie. Elsa ne valide pas la doctrine, elle la met en scène. Les modèles produisent des estimations. Les garanties rendent les revendications crédibles. Les décisions les rendent conséquentes. La gouvernance maintient la justification réfutable. La validation demande si le système était justifié au moment où il a été évalué. La gouvernance demande si nous pouvons encore justifier de l'utiliser maintenant, qui détient l'autorité d'en décider autrement, et quelle preuve nous forcerait à arrêter. Une décision gouvernée n'est pas celle dont le modèle a été validé, mais celle dont la justification reste contestable dans le temps, ce qui est aussi pourquoi, à la légère déception de ceux qui aiment les tableaux de bord, elle ne se réduira pas à un unique score de gouvernance.

Bibliographie

- Vishwanath K, Alyakin A, Ghosh M, et al. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. Nature Medicine. 2026;32:2405-2409. Online 12 June 2026. DOI: 10.1038/s41591-026-04431-5.
- Beaulieu-Jones B, Nemati S. Limited benchmarks constrain the conclusions of a general-purpose versus clinical AI comparison (Matters Arising). Nature Medicine. Published 3 September 2026. DOI: 10.1038/s41591-026-04638-6.
- Vishwanath K, Alyakin A, Ghosh M, et al. Reply to: Limited benchmarks constrain the conclusions of a general-purpose versus clinical AI comparison. Nature Medicine. Published September 2026. DOI: 10.1038/s41591-026-04637-7.

- « Beyond the Leaderboard: Rethinking Medical Benchmarks for Large Language Models. » ArXiv 2508.04325 (ACL 2026).
- Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making. 2006;26(6):565-574.
- FDA and EMA. Guiding Principles of Good AI Practice in Drug Development. 14 January 2026.
- Tadmouri A, Barkaoui S, Bennani M, Vetillard J, Mejbri H. Assessing Reproducibility of Real-World Evidence Analyses Using Synthetic Data: A Validation Study in Rare BRAF V600E Metastatic Non-Small Cell Lung Cancer. ISPOR Europe 2026 (Top 5% abstracts).