Prouver le thermomètre

Ce qu'une preuve d'évaluation peut réutiliser, pour quels usages, sous quelles hypothèses, et ce qui la rend caduque.

Deux économies de la mesure

Les grands modèles de langage ont des classements transversaux. Chaque trimestre, un nouveau score redistribue les places. Les modèles spécialisés qui prédisent une toxicité, génèrent une cohorte synthétique ou estiment une survie disposent de références communes partielles, de jeux publics par tâche, de protocoles par discipline. Ils n'ont pas de classement qui vaille au-delà d'une tâche. Ils ont des dossiers.

La différence tient à ce que chacun cherche à obtenir. Un classement produit de la visibilité sur un protocole donné ; un dossier cherche à faire accepter un résultat pour une décision donnée. *Nature Medicine* en a donné une illustration en 2026, que nous avons analysée dans « Même performance au benchmark, soin opposé ». Vishwanath et al. y rapportaient que trois modèles de langage généralistes surpassaient deux outils d'IA clinique commerciaux, eux-mêmes fondés sur des modèles de langage, sur des benchmarks médicaux et sur cent requêtes réelles de médecins issues d'un unique déploiement. Le *Matters Arising* de Beaulieu-Jones et Nemati soutenait que la portée limitée de ces évaluations contraint les conclusions, surtout lorsqu'on les étire vers l'achat, le remboursement ou la régulation. Le classement n'est pas sans valeur. Il répond à une autre question que celle du régulateur.

Ce texte soutient une thèse : *la crédibilité d'un modèle dépend de la crédibilité du dispositif qui l'évalue, et cette seconde crédibilité est faite de démonstrations qui n'ont pas la même durée de vie*. Certaines peuvent être établies une fois et reprises dans de nombreux dossiers. D'autres valent pour une famille de tâches. D'autres enfin doivent être renouvelées pour chaque usage, ou pendant l'usage. Une part du coût de la preuve vient aujourd'hui de ce que l'on ne sait pas encore organiser cette réutilisation conditionnelle : ce qui peut être repris, pour quels usages, sous quelles hypothèses, et quels changements le rendent caduc.

« Modèle spécialisé en santé » est une catégorie trop large pour une conclusion réglementaire homogène. Deux objets sont visés. Le premier est le dispositif médical intégrant de l'IA qui relève de l'AI Act comme système à haut risque : au sens de son article 6, l'IA doit être elle-même un produit couvert par l'annexe I, ou un composant de sécurité d'un tel produit, et ce produit doit être soumis à une évaluation de conformité par un tiers. Le second est l'outil computationnel utilisé dans le développement d'un médicament,

dont les résultats soutiennent une décision réglementaire. La destination du système et le rôle de ses résultats dans la décision commandent le raisonnement. Pour le premier objet, la preuve a deux moments : la qualification avant la mise sur le marché, et la surveillance du système une fois déployé. Un modèle généraliste qui soutient une décision clinique définie partage les mêmes exigences méthodologiques, représentativité, indépendance de l'évaluation, pertinence des seuils, sans relever automatiquement de la même qualification réglementaire.

Intégrité n'est pas validité

Le mot *benchmark* recouvre trop de choses. Ce texte parle de *banc d'essai* : l'ensemble formé par les choix d'évaluation, les données de test et de référence, le protocole, l'outillage qui l'exécute et le domaine hors duquel il ne dit plus rien. Uberti-Bona Marin et ses coauteurs ont formalisé en 2026 les quatre choix normatifs que suppose l'évaluation d'exactitude au sens de l'AI Act : quelles métriques, quel arbitrage entre elles, quelles données représentatives, quels seuils.

Une première distinction s'impose. Scellement, empreinte, versionnage et journalisation établissent l'*intégrité* du banc : ce qui a été exécuté, sur quoi, dans quel ordre. Ils ne démontrent pas sa *validité*, c'est-à-dire qu'il mesure la bonne propriété. Un protocole peut être parfaitement reproductible et mesurer systématiquement la mauvaise chose.

Prouver le thermomètre suppose donc cinq démonstrations distinctes, dont les deux premières prolongent la distinction entre vérification et validation que formalisent ASME V&V 40 et, pour les mesures numériques en santé, le cadre V3 de Goldsack et ses coauteurs. L'exactitude du calcul : le logiciel calcule correctement la métrique annoncée. La validité de la mesure : cette métrique renseigne effectivement la propriété revendiquée. La qualité de la référence : labels, événements et comparateurs sont assez fiables pour servir d'étalon. La capacité discriminante : le banc repère des défauts pertinents, y compris des défauts introduits volontairement. La pertinence décisionnelle : le résultat soutient la décision envisagée, et non une décision voisine.

La quatrième est, à notre sens, la plus souvent négligée, et la plus facile à rendre opposable. Un banc qui n'a jamais été confronté à des contrôles négatifs, à des cas de référence connus, à des défauts injectés ou à des méthodes simples de comparaison n'a pas démontré qu'il sait rejeter un modèle présentant un défaut déterminé. Ce qu'il démontre reste borné : détecter des défauts injectés établit une capacité sur ces défauts, pas une couverture exhaustive des erreurs possibles. Le banc documente sa couverture et ses limites. Pour des cohortes synthétiques, le test est concret : le banc doit détecter un générateur qui reproduit fidèlement les distributions marginales tout en détruisant les relations entre variables, et un générateur qui mémorise des patients réels.

La métaphore du titre a ses limites. Un thermomètre mesure une grandeur définie. Un banc combine fidélité, utilité, confidentialité et acceptabilité. Une partie relève de la mesure ; une autre relève d'un choix sur ce qui compte et sur le risque acceptable. Ce choix ne se démontre pas, il se déclare et se justifie.

Le banc qui choisit ne peut pas juger seul

La seconde distinction est plus délicate, et elle touche au cœur des systèmes qui engendrent leurs propres modèles. Chez TweenMe, terrain d'implémentation, plusieurs dizaines de modèles relèvent de familles de tâches hétérogènes, et certains sont engendrés par un agent d'optimisation qui hybride des architectures puis propose un ensemble de compromis entre critères. Dans une telle configuration, les résultats du banc remontent vers ce qui construit les modèles. Le banc participe à l'optimisation. Il n'est plus, sans précaution, un instrument d'évaluation finale.

Un jeu de données peut rester scellé, horodaté et parfaitement inchangé tout en étant progressivement surajusté par les décisions prises à partir de ses scores. Le phénomène est documenté depuis longtemps : Cawley et Talbot l'ont décrit en 2010 dans le *Journal of Machine Learning Research* comme un surajustement au critère de sélection, qui biaise ensuite l'estimation de performance. Un journal exhaustif rend cette adaptation observable ; il ne la supprime pas.

Il faut donc séparer trois fonctions : les données et critères qui servent à entraîner, ceux qui servent à choisir les architectures et à arbitrer entre les compromis, et l'évaluation confirmatoire, protégée de ces décisions. Séparer des fonctions ne veut pas dire découper une petite cohorte en trois morceaux fixes. Cela veut dire maîtriser les flux d'information avec une stratégie adaptée à l'effectif : validation imbriquée, données externes, données prospectives. Ce qui compte n'est pas qu'un banc soit antérieur au modèle. Un modèle existant peut recevoir une excellente validation prospective sur des données nouvelles ; un banc antérieur peut devenir un mauvais juge s'il a guidé toute la construction. *L'indépendance informationnelle prime sur l'antériorité chronologique*.

La puissance appartient aussi à la qualification. Sur une cohorte de 184 patients atteints d'un cancer bronchique non à petites cellules métastatique porteur d'une mutation BRAF V600E, terrain de travaux acceptés pour présentation à ISPOR Europe 2026, l'effectif brut dit peu : en survie, le nombre d'événements et la structure de la censure comptent autant, et la précision s'effondre dès qu'on descend en sous-groupes. Produire davantage de patients synthétiques ne crée pas, à soi seul, davantage d'information clinique indépendante.

Les cohortes synthétiques destinées à un bras de contrôle externe exigent enfin un angle que le meilleur indicateur ne remplace pas. La fidélité statistique n'y suffit pas : il faut examiner la comparabilité des populations, la définition du temps zéro, la mesure des

événements, la censure, et les hypothèses qui autorisent à lire une différence comme un effet du traitement. Un générateur peut reproduire fidèlement une base observationnelle, et reproduire fidèlement ses biais. Un banc honnête distingue ce qu'il vérifie directement, ce qu'il examine par analyse de sensibilité, et ce qu'il suppose sans pouvoir l'identifier dans les données.

Trois niveaux de réutilisation

L'absence de banc universel ne conduit pas à reconstruire un banc complet par dossier. Des architectures différentes se comparent très bien sur une même tâche avec un protocole commun ; ce qui sépare les dossiers, ce sont la finalité, la population et la décision à soutenir. Ces différences n'interdisent pas des modules communs. Elles en fixent les frontières.

Trois niveaux se dessinent.

1. L'*infrastructure* : outillage d'exécution, scellement et versionnage des données de test, traçabilité, contrôle d'accès.
2. La *famille de tâches* : protocoles d'analyse de survie, traitement de la censure, analyses de calibration, tests de confidentialité pour les données synthétiques, batteries de contrôles négatifs et de défauts injectés.
3. L'*usage précis* : population cible, comparateur, seuils, conséquences des erreurs, conditions de déploiement.

Le premier niveau se démontre une fois et se maintient. Le deuxième se démontre par famille et se réutilise au sein de cette famille. Le troisième se démontre à chaque usage, et c'est normal : c'est là que vit la décision.

Ce découpage déplace la question. Il ne s'agit plus de choisir entre un thermomètre commun et un thermomètre par dossier, mais de savoir quelle part de la preuve survit au dossier suivant.

Ce que les régulateurs savent déjà qualifier

Les régulateurs savent qualifier des outils indépendamment des produits qu'ils servent à évaluer, et la qualification sert précisément à mutualiser : un outil qualifié dans un contexte d'usage voit ses données acceptées sans re-justification dans d'autres dossiers. En Europe, l'EMA a rendu en mars 2025 son premier avis de qualification d'une méthodologie fondée sur l'IA, AIM-NASH, qui assiste la lecture de biopsies hépatiques dans les essais cliniques. C'est une IA qualifiée comme instrument de mesure ; ce n'est pas un instrument qui évalue d'autres IA.

Pour les systèmes eux-mêmes, l'article 15 de l'AI Act exige un niveau d'exactitude approprié et prévoit, en son paragraphe 3, que les niveaux d'exactitude et les indicateurs

utilisés sont déclarés dans la notice. Son paragraphe 2 charge la Commission d'encourager, avec les autorités de métrologie, l'élaboration de benchmarks et de méthodes de mesure. Les véritables instruments communs prévus par le règlement sont les normes harmonisées et, à défaut, des spécifications communes que la Commission peut adopter par actes d'exécution (articles 40 et 41). Selon le suivi des travaux du CEN-CENELEC publié par AI Standardisation Watch, les parties de normes consacrées à l'exactitude et à la robustesse n'avaient pas atteint l'enquête publique en septembre 2026, et aucune norme harmonisée au titre de l'AI Act n'était encore citée au Journal officiel. Pour l'IA embarquée dans un dispositif médical, le règlement Digital Omnibus (règlement (UE) 2026/1744) a reporté l'échéance au 2 août 2028. En France, la grille publiée par la HAS en 2020 pour les dispositifs à apprentissage automatique demande quels sont les seuils de performance, sans imposer de jeu de test commun.

Le régulateur américain, qui pratique la qualification d'outils depuis le plus longtemps, donne la mesure du débit. Côté médicament, la qualification des outils de développement préexistait au 21st Century Cures Act ; sa section 507 l'a codifiée et dotée d'une procédure légale. Le programme de qualification des outils de mesure clinique comptait 86 outils inscrits pour 7 qualifiés, en six ans en moyenne, aucun comme critère principal (Yang et al., 2025). Côté dispositif, le programme volontaire MDDT comptait en février 2026 21 outils qualifiés depuis 2017, plusieurs computationnels, aucun présenté comme un outil d'IA. Ces chiffres décrivent l'histoire de programmes, pas leur capacité, et ils portent surtout sur des outils de mesure clinique, pas sur des instruments d'évaluation de modèles appris.

Ces derniers existent pourtant, sous une autre forme. Le catalogue des *Regulatory Science Tools* de la FDA, présenté comme une ressource à utiliser là où normes et outils qualifiés font encore défaut, compte un domaine consacré à l'IA et à l'apprentissage automatique, avec des outils d'évaluation de la calibration ou de l'accord entre un dispositif de segmentation et un panel d'experts. Ils ne sont pas qualifiés. Mais ils montrent que le régulateur produit lui-même des briques réutilisables de niveau « famille de tâches », ce qui va dans le sens du découpage proposé ici. Enfin, le cadre de crédibilité le plus directement pertinent est récent : le projet de guidance publié par la FDA en janvier 2025 sur l'usage de l'IA dans les décisions réglementaires du médicament propose sept étapes, et proportionne l'exigence à l'influence du modèle sur la décision et à la gravité d'une décision erronée. En octobre 2026, il restait au stade de projet.

À titre de repère daté, les pratiques d'évaluation restaient très inégales : entre 2015 et 2020, 126 des 130 dispositifs d'IA autorisés par la FDA n'avaient été évalués que sur données rétrospectives (Wu et al., *Nature Medicine*, 2021).

Ce qui se réinvente à chaque dossier

Un organisme notifié rappellerait ici que la norme ISO 13485, en son paragraphe 7.6, impose déjà de valider les logiciels de surveillance et de mesure avant leur première utilisation, puis après modification, de façon proportionnée au risque. Il aurait raison pour une partie du niveau infrastructure : l'outillage qui exécute le banc a une adresse réglementaire existante, au sein du système qualité, et sa validation couvre la première des cinq démonstrations, l'exactitude du calcul. Elle ne couvre pas la validité de la mesure, ni la qualité de la référence. Si cet outillage est revalidé à chaque dossier, c'est souvent que les éditeurs ne l'ont pas logé là où le cadre l'attend.

Notre expérience d'éditeur le suggère : le coût de preuve d'un nouveau modèle ne décroît pas sensiblement d'un dossier à l'autre lorsque les finalités diffèrent, et une part non négligeable de ce coût tient au niveau infrastructure. Cette part décrit un *périmètre potentiellement mutualisable*, pas une économie démontrée : une infrastructure partagée se construit, se maintient, se revalide. Ce qui est réinventé à chaque dossier, c'est une partie de ce périmètre, pas sa totalité.

D'autres industries ont appris à faire croître la confiance dans un outil au lieu de la reconstruire. Dans l'automobile, la partie 8 de la norme ISO 26262 admet, parmi les méthodes d'établissement de la confiance dans un outil logiciel, la confiance accrue par l'usage, quand l'outil a servi au même usage, dans des cas comparables, avec des défaillances documentées et une spécification inchangée ; c'est aussi la voie la moins recommandée aux niveaux d'intégrité les plus élevés, où la validation de l'outil prend le relais. Dans l'aéronautique, le couple DO-178C et DO-330 proportionne la qualification d'un outil de vérification à ce que sa sortie permet de ne plus vérifier ailleurs. Côté FDA, la guidance de 2023 sur la crédibilité des modèles computationnels, adossée à ASME V&V 40, procède du même esprit tout en excluant les modèles d'apprentissage. Ce sont des analogies de méthode, pas des obligations transposables. Elles montrent qu'une preuve d'outil peut s'accumuler, à condition d'en suivre l'usage. La FDA a tenté, à l'inverse, de pré-certifier l'organisation elle-même ; le pilote, conclu en 2022, a abouti au constat qu'une modification législative serait nécessaire.

Le reste du coût n'est pas pour autant figé. Des protocoles par famille de tâches, des contrats de données, des méthodes statistiques réutilisables peuvent en réduire une partie. Ce qui ne se mutualise pas automatiquement, c'est le dernier niveau, celui où l'on fixe une population, un comparateur, un seuil et les conséquences d'une erreur : une justification de seuil ou un comparateur peuvent être repris entre usages proches, mais leur transférabilité doit être démontrée pour chaque nouvel usage. *Le coût de la preuve d'usage suit la diversité des décisions ; la dette d'infrastructure ne suit que le retard de la profession.*

Partager les principes, pas les modèles

Il existe un précédent où un écosystème de modèles hétérogènes, issus de producteurs différents, a obtenu une recevabilité réglementaire sans commencer par un banc commun ni un classement. En novembre 2004, l'OCDE adoptait cinq principes de validation des modèles de relation structure-activité : un critère d'évaluation défini, un algorithme sans ambiguïté, un domaine d'applicabilité défini, des mesures d'ajustement, de robustesse et de prédictivité, une interprétation mécanistique lorsque c'est possible. Le règlement REACH, en 2006, admettait déjà sous conditions des résultats de ce type à la place d'essais. La ligne directrice ICH M7, en 2014, demandait pour les impuretés mutagènes du médicament deux méthodologies complémentaires, l'une fondée sur des règles d'expert, l'autre statistique, conformes à ces principes et soumises si besoin à revue d'expert ; l'absence d'alerte dans les deux suffit à conclure à l'absence de préoccupation mutagène sans essai complémentaire. Chaque producteur a gardé ses modèles. Ce qui a été mis en commun, ce sont des critères pour juger qu'un modèle est recevable. *La toxicologie n'a pas partagé ses modèles, elle a partagé ses principes.*

Un banc commun a existé aussi, mais il est venu après les principes. À partir de 2014, l'Institut national japonais des sciences de la santé a soumis à l'aveugle dix-sept outils de douze éditeurs à un même jeu de plus de 12 000 substances. Ce programme visait l'amélioration comparative des modèles ; il ne constituait pas, à lui seul, une procédure de qualification réglementaire.

La transposition a des limites. La mutagénicité est un critère étroit, presque binaire ; les modèles dont il est question ici visent des finalités multiples, du classement à la génération de cohortes. Ce qui se transpose est une forme, des exigences de qualification partagées par famille de tâches, une règle de méthode, une revue d'expert, pas un contenu. Et le temps doit être dit : deux ans entre les principes de l'OCDE et une première recevabilité sous conditions dans REACH, dix ans jusqu'à une recevabilité qui dispense d'essai dans le médicament.

Les propositions de principes viendront de plusieurs lieux : consortiums académiques, évaluateurs, grands utilisateurs, organismes de normalisation, éditeurs qui produisent des modèles en série. Ces derniers ont un intérêt évident à ce que leurs propres critères deviennent la norme, ce qui est une raison de plus pour que leurs propositions soient publiées, contestables et instruites par un tiers dont on connaît le financement, qui accède aux essais infructueux et pas seulement aux succès, et qui peut contester les critères eux-mêmes. Publier ne suffit pas à empêcher la captation par les acteurs dominants ; cela permet au moins de la voir. L'article 15, paragraphe 2, de l'AI Act offre à cette mise en commun un point d'appui ; les normes harmonisées et les spécifications communes en sont le débouché.

Une preuve réutilisable doit pouvoir être révoquée

La mutualisation a une contrepartie : elle concentre les défaillances. Une erreur de calcul, un défaut de prétraitement ou une référence inadéquate dans un banc partagé peut invalider les conclusions de dizaines de modèles d'un coup. Qualifier un banc une fois ne suffit donc pas. Il faut savoir quels modèles dépendent de quelle version, quelles preuves deviennent invalides lorsqu'un défaut est découvert, quelles modifications imposent une réévaluation, comment traiter un jeu de test devenu trop connu, comment détecter que la population ou les pratiques ont changé.

Les régulateurs ont commencé à outiller cette question. L'avis de l'EMA sur AIM-NASH qualifie un outil figé, dont une modification majeure peut exiger une nouvelle qualification. Les plans de modification prédéterminés de la FDA, dont la guidance finale pour les fonctions logicielles d'IA, publiée en décembre 2024, a été mise à jour en août 2025, organisent à l'avance quelles modifications restent couvertes et lesquelles appellent une nouvelle évaluation. Une preuve réutilisable n'est pas une preuve figée. C'est une preuve dont on connaît les dépendances, qu'on peut révoquer et recalculer. Ce qui devient caduc n'est pas toujours le résultat : ce peut être son interprétation, ou son droit de réemploi dans un contexte donné. On conserve les observations et leur traçabilité ; on retire, lorsque c'est nécessaire, les conclusions ou les réemplois qu'elles ne soutiennent plus. Versionner l'instrument protège son intégrité ; seule la réévaluation de ses hypothèses dit si sa validité tient.

Du banc de sortie au banc de service

La durée de vie d'une preuve ne se compte pas seulement en dossiers. Elle se compte aussi en mois d'exploitation : justifier une mise en service ne suffit pas, il faut encore établir, pendant l'usage, que les conditions qui fondaient cette justification restent réunies. Nous avons soutenu dans « Même performance au benchmark, soin opposé » que la validation établit qu'un système pouvait être utilisé, et que la gouvernance doit établir qu'il peut continuer à l'être. Il s'agit ici d'en tirer les conséquences pour le banc et pour ce qu'un fournisseur livre.

Pour un système d'IA à haut risque, et c'est bien le système que vise l'AI Act, pas le modèle seul, ce second moment repose sur une responsabilité répartie. Le fournisseur rend le système journalisable tout au long de sa vie (article 12), déclare dans la notice ses caractéristiques, ses limites, son niveau d'exactitude et les indicateurs utilisés, ainsi que les mécanismes qui permettent au déployeur de collecter, stocker et interpréter les journaux (article 13), puis collecte et analyse activement les données d'exploitation pour évaluer le maintien de la conformité (article 72) ; pour un dispositif médical, ce suivi peut s'intégrer à la surveillance après commercialisation du règlement MDR. Le déployeur

surveille le fonctionnement selon la notice, informe le fournisseur et suspend l'usage dans les situations de risque prévues (article 26).

Ces textes n'obligent pas à livrer un banc complet à l'exploitant, et ils ne permettent pas au fournisseur de transférer la surveillance en lui remettant une méthode. Livrer avec le système des méthodes d'évaluation en continu est la traduction industrielle défendable de cette répartition, à condition de distinguer l'exigence réglementaire du moyen technique choisi pour la satisfaire. Concrètement, chaque système devrait être accompagné d'un banc de service : les propriétés à surveiller et les données qu'elles exigent, les indicateurs avec leurs références et leurs limites d'interprétation, les fréquences d'analyse et les effectifs nécessaires pour qu'une analyse dise quelque chose, les critères qui déclenchent une investigation, une réévaluation ou une suspension, et la répartition des rôles avec les modalités de remontée vers le fournisseur.

Trois fonctions techniques le soutiennent. La *journalisation* permet de reconstituer les événements et les conditions d'exécution. La *télémétrie* recueille les signaux nécessaires à la surveillance : version en service, caractéristiques des entrées, anomalies, sorties, interventions humaines, incidents, selon ce que l'usage rend pertinent. L'*évaluation* interprète ces signaux pour décider si les performances et les conditions de validité restent acceptables. La télémétrie apporte les observations ; le protocole d'évaluation leur donne un sens. Une dérive des entrées justifie une investigation sans démontrer une baisse de performance ; des entrées stables ne garantissent pas que le système reste cliniquement pertinent. L'AI Act n'impose pas la transmission permanente de toutes les données au fournisseur : la collecte peut rester locale ou passer par une remontée encadrée, dont le périmètre, la durée de conservation et les responsabilités sont définis explicitement, sans centraliser les données des patients.

Trois angles morts apparaissent alors. Surveiller en continu ne veut pas dire mesurer en permanence la performance clinique. Les données manquantes, les entrées hors domaine, les anomalies techniques se suivent vite. Une calibration de survie ou une performance sur un événement rare exigent du recul et des événements observés, parfois des mois plus tard. Sans cette boucle de retour qui rapproche les prédictions des issues réelles, la télémétrie renseigne surtout le fonctionnement et les changements détectables, très peu la justesse des résultats. Encore faut-il que ces issues soient accessibles et représentatives : si seuls certains patients sont suivis, ou si les résultats ne remontent que pour les cas ayant déclenché une alerte, la surveillance dessine une image biaisée de la performance. La qualité du retour clinique fait partie de la qualification du banc de service. L'absence d'alerte immédiate ne démontre pas le maintien de la validité clinique.

Le banc de service doit lui-même être évalué. Quels écarts détecte-t-il, avec quel délai, avec combien de fausses alertes, et quels défauts laisse-t-il passer ? Le problème du thermomètre se prolonge après la livraison : un tableau de bord peut fonctionner

parfaitement sans détecter la dégradation qui importe. Les cinq démonstrations attendues d'un banc s'y appliquent, avec une contrainte de plus : rester valide entre d'autres mains, sur une population que son auteur ne contrôle pas.

Enfin, organiser la collecte, obtenir les résultats cliniques, instruire les alertes et réévaluer les conclusions est un coût récurrent légitime, qui n'a rien d'une dette : il existe parce que l'usage continue et que son environnement change. La surveillance, elle, suit la durée de l'usage : elle s'optimise, mais elle ne disparaît pas ; elle se finance.

Le produit livré ne se limite donc plus au système. Il comprend les preuves de son évaluation initiale, l'instrumentation de sa surveillance, et les règles qui permettent d'agir sur les signaux recueillis. Ce qui doit se prolonger en exploitation, c'est la capacité d'évaluation et la circulation des informations qu'elle exige, pas nécessairement le banc lui-même. Une évaluation qui ne sort pas de l'usine ne prouve que la sortie d'usine.

Limites

Le terrain principal est celui d'un éditeur. Les données qui fondent l'argument économique ne sont pas publiées, si bien que le lecteur ne peut pas le contrôler ; il doit le lire comme une hypothèse argumentée. D'autres producteurs peuvent déjà connaître un coût marginal décroissant, et un coût qui ne décroît pas sensiblement entre modèles de finalités différentes n'établit pas, à lui seul, une absence d'apprentissage. Les normes industrielles citées sont payantes et paraphrasées à partir de synthèses produites par des acteurs intéressés à la qualification d'outils, éditeurs ou organisme notifié. Enfin, une norme harmonisée ou une spécification commune sur l'exactitude, si elle était adoptée, déplacerait une partie de l'argument vers son adéquation aux modèles concernés.

Organiser ce qui survit au dossier suivant

Quatre exigences en découlent : séparer la sélection de l'évaluation confirmatoire ; démontrer la validité du banc, niveau par niveau ; rendre toute preuve réutilisable révocable ; livrer avec le système de quoi savoir quand ses conclusions cessent d'être étayées.

Un classement dit qui arrive premier sur un protocole donné. *Une conclusion doit dire ce qui la rend caduque.*

Sources citées

- AI Act, règlement (UE) 2024/1689, articles 6, 12, 13(3), 15, 26(5)-(6), 40, 41, 72 (version française officielle, JOUE L du 12/07/2024).
- Règlement (UE) 2026/1744 du 8 juillet 2026 (Digital Omnibus on AI), JOUE du 24/07/2026, entrée en vigueur le 27/07/2026, https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng [références vérifiées sur EUR-Lex ; date d'application de l'annexe I au 2 août 2028 vérifiée via White & Case].
- Vishwanath et al., « General-purpose large language models outperform specialized clinical AI tools on medical benchmarks », *Nature Medicine*, 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*, 2026, DOI 10.1038/s41591-026-04638-6.
- Uberti-Bona Marin L. G., Rijsbosch B., Meding K., Spanakis G., van Dijck G., Kollnig K., « Is your AI Model Accurate Enough? The Difficult Choices Behind Rigorous AI Development and the EU AI Act », DOI 10.1145/3805689.3806436 ; prépublication arXiv 2604.03254.
- Cawley G. C., Talbot N. L. C., « On over-fitting in model selection and subsequent selection bias in performance evaluation », *Journal of Machine Learning Research* 11(70):2079-2107, 2010, https://www.jmlr.org/papers/v11/cawley10a.html
- Wu E., Wu K., Daneshjou R., Ouyang D., Ho D. E., Zou J., *Nature Medicine* 27(4):582-584, 2021, DOI 10.1038/s41591-021-01312-x ; Stanford HAI, *Toward Stronger FDA Approval Standards for AI Medical Devices*, 2022.
- EMA, avis de qualification AIM-NASH, 20/03/2025, https://www.ema.europa.eu/en/news/ema-qualifies-first-artificial-intelligence-tool-diagnose-inflammatory-liver-disease-mash-biopsy-samples
- FDA, Drug Development Tool qualification process (section 507 FD&C Act), https://www.fda.gov/drugs/drug-development-tool-ddt-qualification-programs/drug-development-tool-ddt-qualification-process
- FDA, programme MDDT, registre à jour au 10/02/2026, https://www.fda.gov/medical-devices/medical-device-development-tools-mddt
- Yang F., Bousnina I., Madej A., Kalamegham R., *Therapeutic Innovation & Regulatory Science* 59(4):871-881, 2025, https://link.springer.com/article/10.1007/s43441-025-00790-2

- FDA, catalogue des Regulatory Science Tools (CDRH/OSEL), domaine IA/ML, https://cdrh-rst.fda.gov [consulté le 12/10/2026 lors de la revue de rigueur].
- FDA, *Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products*, projet de guidance, 07/01/2025 (statut projet vérifié en octobre 2026).
- FDA, *Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions*, guidance finale, 2023 ; ASME V&V 40-2018.
- FDA, *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions*, guidance finale (première version décembre 2024 ; version actuelle août 2025, mise en ligne le 18/08/2025), https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial
- FDA, *The Software Precertification (Pre-Cert) Pilot Program: Tailored Total Product Lifecycle Approaches and Key Findings*, septembre 2022, https://www.fda.gov/media/161815/download
- HAS, grille descriptive des dispositifs médicaux à apprentissage automatique, 14/10/2020, https://has-sante.fr/jcms/p_3212876/fr/un-nouvel-outil-pour-l-evaluation-des-dispositifs-medicaux-embarquant-de-l-intelligence-artificielle
- AI Standardisation Watch (G. P. Krog), édition 16, septembre 2026, https://krogrules.com/news/edition-16-september-2026 ; AI Act Insight, état des normes harmonisées au 24/08/2026, https://aiactinsight.eu/en/standards/harmonised-standards/ [sources secondaires concordantes ; issue du plénier JTC 21 d'octobre 2026 non publiée à la date de rédaction].
- ISO 13485:2016, §7.6 ; ISO 26262-8:2018, clause 11 ; RTCA DO-178C et DO-330 (2011) [normes payantes, citées via : BSI, *Medical devices validation and verification* (webinaire) ; Synopsys, *Automotive tool confidence, ISO 26262* (livre blanc) ; LDRA, *DO-330 Test Tool Qualification Technical Briefing* v2.0].
- OCDE, principes de validation des (Q)SAR, novembre 2004 (JRC, *The Characterisation of (Q)SARs: Preliminary Guidance*, 2005) ; règlement (CE) 1907/2006 (REACH), annexe XI, §1.3 ; ICH M7(R2), section 6.
- Honma M. et al., Ames/QSAR International Challenge Project, *Mutagenesis* 34(1), 2019.
- 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, accepté pour présentation (Vienne, 8-11 novembre 2026).

- Vetillard J., « Même performance au benchmark, soin opposé », Twingital Institute, 04/09/2026.
- Goldsack J. C. et al., « Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs) », *npj Digital Medicine* 3:55, 2020.