Proving the Thermometer

What evaluation evidence can be reused, for which uses, under which assumptions, and what makes it obsolete.

Two economies of measurement

Large language models have cross-cutting leaderboards. Every quarter, a new score reshuffles the rankings. Specialized models that predict toxicity, generate a synthetic cohort or estimate survival have partial common references: public datasets per task, protocols per discipline. They have no leaderboard that holds beyond a single task. They have dossiers.

The difference lies in what each is trying to obtain. A leaderboard produces visibility on a given protocol; a dossier seeks acceptance of a result for a given decision. *Nature Medicine* offered an illustration in 2026, which we analysed in "Same Benchmark Performance, Opposite Care". Vishwanath et al. reported there that three general-purpose language models outperformed two commercial clinical AI tools, themselves built on language models, on medical benchmarks and on one hundred real physician queries drawn from a single deployment. The *Matters Arising* by Beaulieu-Jones and Nemati argued that the limited scope of these evaluations constrains the conclusions, especially when they are stretched toward procurement, reimbursement or regulation. The leaderboard is not without value. It answers a different question from the regulator's.

This text defends one thesis: *the credibility of a model depends on the credibility of the apparatus that evaluates it, and that second credibility is made of demonstrations that do not have the same lifespan*. Some can be established once and reused across many dossiers. Others hold for a family of tasks. Others still must be renewed for each use, or during use. Part of the cost of evidence today comes from the fact that we do not yet know how to organize this conditional reuse: what can be carried over, for which uses, under which assumptions, and which changes make it obsolete.

"Specialized healthcare model" is too broad a category for a uniform regulatory conclusion. Two objects are in scope. The first is the AI-enabled medical device that falls under the AI Act as a high-risk system: under its Article 6, the AI must itself be a product covered by Annex I, or a safety component of such a product, and that product must be subject to third-party conformity assessment. The second is the computational tool used in drug development whose results support a regulatory decision. The intended purpose of the system and the role of its results in the decision govern the reasoning. For the first object, evidence has two moments: qualification before placing on the market, and

monitoring of the system once deployed. A general-purpose model that supports a defined clinical decision shares the same methodological requirements, representativeness, independence of evaluation, relevance of thresholds, without automatically falling under the same regulatory qualification.

Integrity is not validity

The word *benchmark* covers too many things. This text speaks of a *test bench*: the whole formed by the evaluation choices, the test and reference data, the protocol, the tooling that runs it, and the domain outside which it no longer says anything. In 2026, Uberti-Bona Marin and co-authors formalized the four normative choices that accuracy evaluation in the sense of the AI Act presupposes: which metrics, how to trade them off, which representative data, which thresholds.

A first distinction is necessary. Sealing, fingerprinting, versioning and logging establish the *integrity* of the bench: what was executed, on what, in which order. They do not demonstrate its *validity*, that is, that it measures the right property. A protocol can be perfectly reproducible and systematically measure the wrong thing.

Proving the thermometer therefore requires five distinct demonstrations, the first two of which extend the distinction between verification and validation formalized by ASME V&V 40 and, for digital health measurements, by the V3 framework of Goldsack and co-authors. Computational correctness: the software computes the announced metric correctly. Measurement validity: the metric actually informs the claimed property. Quality of the reference: labels, events and comparators are reliable enough to serve as a standard. Discriminating capacity: the bench detects relevant defects, including deliberately introduced ones. Decisional relevance: the result supports the decision at hand, and not a neighbouring one.

The fourth is, in our view, the most often neglected, and the easiest to make enforceable. A bench that has never been confronted with negative controls, known reference cases, injected defects or simple baseline methods has not shown that it can reject a model with a given defect. What it does show remains bounded: detecting injected defects establishes a capacity on those defects, not exhaustive coverage of possible errors. The bench documents its coverage and its limits. For synthetic cohorts the test is concrete: the bench must catch a generator that faithfully reproduces marginal distributions while destroying the relationships between variables, and a generator that memorizes real patients.

The metaphor in the title has its limits. A thermometer measures a defined quantity. A bench combines fidelity, utility, privacy and acceptability. Part of it is measurement; another part is a choice about what matters and about acceptable risk. That choice cannot be demonstrated; it is declared and justified.

The bench that selects cannot judge alone

The second distinction is more delicate, and it goes to the heart of systems that generate their own models. At TweenMe, our implementation ground, several dozen models belong to heterogeneous task families, and some are generated by an optimization agent that hybridizes architectures and then proposes a set of trade-offs between criteria. In such a configuration, the bench's results flow back to whatever builds the models. The bench takes part in the optimization. Without precautions, it is no longer an instrument of final evaluation.

A dataset can remain sealed, timestamped and perfectly unchanged while being progressively overfitted by the decisions taken on the basis of its scores. The phenomenon has long been documented: in 2010, Cawley and Talbot described it in the *Journal of Machine Learning Research* as overfitting to the model selection criterion, which then biases the performance estimate. An exhaustive log makes this adaptation observable; it does not remove it.

Three functions must therefore be separated: the data and criteria used for training, those used to select architectures and arbitrate between trade-offs, and the confirmatory evaluation, shielded from those decisions. Separating functions does not mean cutting a small cohort into three fixed pieces. It means controlling information flows with a strategy suited to the sample size: nested validation, external data, prospective data. What matters is not that a bench predates the model. An existing model can receive an excellent prospective validation on new data; an earlier bench can become a poor judge if it guided the whole construction. *Informational independence takes precedence over chronological precedence.*

Statistical power is also part of qualification. In a cohort of 184 patients with metastatic non-small cell lung cancer carrying a BRAF V600E mutation, the subject of work accepted for presentation at ISPOR Europe 2026, the raw sample size says little: in survival analysis, the number of events and the censoring structure matter as much, and precision collapses as soon as one moves to subgroups. Producing more synthetic patients does not, by itself, create more independent clinical information.

Finally, synthetic cohorts intended for an external control arm require an angle that no better metric can replace. Statistical fidelity is not enough: one must examine the comparability of populations, the definition of time zero, the measurement of events, censoring, and the assumptions that allow a difference to be read as a treatment effect. A generator can faithfully reproduce an observational database, and faithfully reproduce its biases. An honest bench distinguishes what it verifies directly, what it examines through sensitivity analysis, and what it assumes without being able to identify it in the data.

Three levels of reuse

The absence of a universal bench does not mean rebuilding a complete bench for each dossier. Different architectures compare perfectly well on the same task with a common protocol; what separates dossiers is the intended purpose, the population and the decision to be supported. These differences do not rule out common modules. They set their boundaries.

Three levels emerge.

1. *Infrastructure*: execution tooling, sealing and versioning of test data, traceability, access control.
2. *Task family*: survival analysis protocols, handling of censoring, calibration analyses, privacy tests for synthetic data, batteries of negative controls and injected defects.
3. *Specific use*: target population, comparator, thresholds, consequences of errors, deployment conditions.

The first level is demonstrated once and maintained. The second is demonstrated per family and reused within that family. The third is demonstrated for each use, and rightly so: that is where the decision lives.

This breakdown shifts the question. It is no longer a matter of choosing between a common thermometer and one thermometer per dossier, but of knowing which part of the evidence survives into the next dossier.

What regulators already know how to qualify

Regulators know how to qualify tools independently of the products they are used to evaluate, and qualification serves precisely to pool effort: a tool qualified within a context of use has its data accepted in other dossiers without re-justification. In Europe, the EMA issued in March 2025 its first qualification opinion on an AI-based methodology, AIM-NASH, which assists the reading of liver biopsies in clinical trials. It is an AI qualified as a measurement instrument; it is not an instrument that evaluates other AI.

For the systems themselves, Article 15 of the AI Act requires an appropriate level of accuracy and provides, in its paragraph 3, that the levels of accuracy and the relevant metrics be declared in the instructions for use. Its paragraph 2 tasks the Commission with encouraging, together with metrology authorities, the development of benchmarks and measurement methodologies. The actual common instruments provided for by the Regulation are harmonised standards and, failing those, common specifications that the Commission may adopt by implementing acts (Articles 40 and 41). According to the tracking of CEN-CENELEC work published by AI Standardisation Watch, the parts of

standards dealing with accuracy and robustness had not reached public enquiry in September 2026, and no harmonised standard under the AI Act had yet been cited in the Official Journal. For AI embedded in a medical device, the Digital Omnibus Regulation (Regulation (EU) 2026/1744) postponed the deadline to 2 August 2028. In France, the descriptive grid published by the HAS (the French National Authority for Health) in 2020 for machine-learning devices asks what the performance thresholds are, without imposing a common test set.

The US regulator, which has practised tool qualification the longest, gives the measure of throughput. On the drug side, the qualification of drug development tools predated the 21st Century Cures Act; its Section 507 codified it and gave it a statutory procedure. The clinical outcome assessment qualification programme listed 86 tools for 7 qualified, in six years on average, none as a primary endpoint (Yang et al., 2025). On the device side, the voluntary MDDT programme counted in February 2026 21 tools qualified since 2017, several of them computational, none presented as an AI tool. These figures describe the history of programmes, not their capacity, and they mostly concern clinical measurement tools, not instruments for evaluating learned models.

The latter do exist, however, in another form. The FDA's catalogue of *Regulatory Science Tools*, presented as a resource for use where standards and qualified tools are still lacking, includes a domain dedicated to AI and machine learning, with tools for evaluating calibration or the agreement between a segmentation device and a panel of experts. They are not qualified. But they show that the regulator itself produces reusable building blocks at the "task family" level, which supports the breakdown proposed here. Finally, the most directly relevant credibility framework is recent: the draft guidance published by the FDA in January 2025 on the use of AI in regulatory decision-making for drugs sets out seven steps, and scales the requirement to the model's influence on the decision and to the consequences of a wrong decision. In October 2026, it was still at the draft stage.

As a dated benchmark, evaluation practices remained very uneven: between 2015 and 2020, 126 of the 130 AI devices cleared by the FDA had been evaluated on retrospective data only (Wu et al., *Nature Medicine*, 2021).

What gets reinvented with every dossier

A notified body would point out here that ISO 13485, in its clause 7.6, already requires the validation of monitoring and measurement software before first use, then after changes, proportionately to risk. It would be right for part of the infrastructure level: the tooling that runs the bench already has a regulatory home within the quality management system, and its validation covers the first of the five demonstrations, computational correctness. It does not cover measurement validity, nor the quality of the reference. If this tooling is revalidated for every dossier, it is often because vendors have not placed it where the framework expects it.

Our experience as a vendor suggests as much: the cost of evidence for a new model does not decrease appreciably from one dossier to the next when intended purposes differ, and a non-negligible share of that cost lies at the infrastructure level. This share describes a *potentially poolable perimeter*, not a demonstrated saving: shared infrastructure has to be built, maintained and revalidated. What is reinvented with every dossier is part of that perimeter, not all of it.

Other industries have learned to grow confidence in a tool instead of rebuilding it. In automotive, part 8 of ISO 26262 accepts, among the methods for establishing confidence in a software tool, increased confidence from use, when the tool has served the same purpose, in comparable cases, with documented malfunctions and an unchanged specification; it is also the least recommended route at the highest integrity levels, where validation of the tool takes over. In aeronautics, DO-178C together with DO-330 scales the qualification of a verification tool to what its output allows one to stop verifying elsewhere. On the FDA side, the 2023 guidance on the credibility of computational modelling, built on ASME V&V 40, proceeds in the same spirit while excluding learned models. These are methodological analogies, not transposable obligations. They show that evidence about a tool can accumulate, provided its use is tracked. The FDA tried, conversely, to pre-certify the organization itself; the pilot, concluded in 2022, ended with the finding that a legislative change would be needed.

The rest of the cost is not fixed for all that. Protocols per task family, data contracts and reusable statistical methods can reduce part of it. What does not pool automatically is the last level, where a population, a comparator, a threshold and the consequences of an error are set: a threshold justification or a comparator can be carried over between closely related uses, but their transferability must be demonstrated for each new use. *The cost of use-specific evidence follows the diversity of decisions; the infrastructure debt follows only the profession's delay.*

Share principles, not models

There is a precedent in which an ecosystem of heterogeneous models, from different producers, obtained regulatory acceptability without starting from a common bench or a leaderboard. In November 2004, the OECD adopted five principles for the validation of structure-activity relationship models: a defined endpoint, an unambiguous algorithm, a defined domain of applicability, appropriate measures of goodness-of-fit, robustness and predictivity, and a mechanistic interpretation where possible. The REACH Regulation, in 2006, already accepted results of this kind in place of tests, under conditions. The ICH M7 guideline, in 2014, required for mutagenic impurities in medicines two complementary methodologies, one expert rule-based, the other statistical, compliant with those principles and subject to expert review where needed; the absence of alerts from both is sufficient to conclude that there is no mutagenic concern, with no further testing. Each

producer kept its models. What was pooled were criteria for judging whether a model is acceptable. *Toxicology did not share its models; it shared its principles.*

A common bench did exist as well, but it came after the principles. From 2014 onward, Japan's National Institute of Health Sciences submitted, blind, seventeen tools from twelve vendors to the same set of more than 12,000 substances. That programme aimed at the comparative improvement of models; it did not, on its own, constitute a regulatory qualification procedure.

The transposition has its limits. Mutagenicity is a narrow, almost binary endpoint; the models discussed here serve multiple purposes, from classification to cohort generation. What transposes is a form, qualification requirements shared per task family, a rule of method, expert review, not a content. And time must be stated: two years between the OECD principles and a first conditional acceptance under REACH, ten years until an acceptance that waives testing in medicines.

Proposals for principles will come from several places: academic consortia, evaluators, large users, standardization bodies, vendors that produce models at scale. The latter have an obvious interest in seeing their own criteria become the norm, which is all the more reason for their proposals to be published, contestable and examined by a third party whose funding is known, who has access to failed trials and not only to successes, and who can challenge the criteria themselves. Publishing is not enough to prevent capture by dominant players; it at least makes capture visible. Article 15(2) of the AI Act offers a foothold for this pooling; harmonised standards and common specifications are its outlet.

Reusable evidence must be revocable

Pooling has a counterpart: it concentrates failures. A computational error, a preprocessing defect or an inadequate reference in a shared bench can invalidate the conclusions of dozens of models at once. Qualifying a bench once is therefore not enough. One must know which models depend on which version, which evidence becomes invalid when a defect is discovered, which changes require re-evaluation, how to handle a test set that has become too well known, how to detect that the population or practices have changed.

Regulators have begun to equip themselves for this question. The EMA opinion on AIM-NASH qualifies a locked tool, a major modification of which may require a new qualification. The FDA's predetermined change control plans, whose final guidance for AI-enabled device software functions, issued in December 2024, was updated in August 2025, organize in advance which modifications remain covered and which call for a new evaluation. Reusable evidence is not frozen evidence. It is evidence whose dependencies are known, which can be revoked and recomputed. What becomes obsolete is not always

the result: it may be its interpretation, or its right of reuse in a given context. Observations and their traceability are kept; conclusions or reuses that they no longer support are withdrawn when necessary. Versioning the instrument protects its integrity; only the re-examination of its assumptions tells whether its validity holds.

From the factory bench to the in-service bench

The lifespan of evidence is not counted in dossiers alone. It is also counted in months of operation: justifying entry into service is not enough; one must still establish, during use, that the conditions on which that justification rested still hold. We argued in "Same Benchmark Performance, Opposite Care" that validation establishes that a system could be used, and that governance must establish that it may keep being used. The point here is to draw the consequences for the bench and for what a provider delivers.

For a high-risk AI system, and it is indeed the system that the AI Act targets, not the model alone, this second moment rests on shared responsibility. The provider makes the system capable of logging throughout its lifetime (Article 12), declares in the instructions for use its characteristics, its limitations, its level of accuracy and the relevant metrics, as well as the mechanisms that allow the deployer to collect, store and interpret the logs (Article 13), then actively collects and analyses operational data to evaluate continued compliance (Article 72); for a medical device, this monitoring can be integrated into the post-market surveillance required by the MDR. The deployer monitors operation according to the instructions for use, informs the provider and suspends use in the risk situations provided for (Article 26).

These texts do not require delivering a complete bench to the operator, and they do not allow the provider to transfer monitoring by handing over a method. Delivering continuous evaluation methods together with the system is the defensible industrial translation of this allocation, provided the regulatory requirement is distinguished from the technical means chosen to meet it. In practice, each system should come with an in-service bench: the properties to monitor and the data they require, the indicators with their references and their limits of interpretation, the analysis frequencies and sample sizes needed for an analysis to say anything, the criteria that trigger an investigation, a re-evaluation or a suspension, and the allocation of roles together with the arrangements for reporting back to the provider.

Three technical functions support it. *Logging* makes it possible to reconstruct events and execution conditions. *Telemetry* collects the signals needed for monitoring: version in service, input characteristics, anomalies, outputs, human interventions, incidents, as the use requires. *Evaluation* interprets these signals to decide whether performance and validity conditions remain acceptable. Telemetry provides the observations; the evaluation protocol gives them meaning. Input drift justifies an investigation without demonstrating a drop in performance; stable inputs do not guarantee that the system

remains clinically relevant. The AI Act does not require the permanent transmission of all data to the provider: collection can remain local or go through controlled reporting, whose scope, retention period and responsibilities are explicitly defined, without centralizing patient data.

Three blind spots then appear. Continuous monitoring does not mean permanent measurement of clinical performance. Missing data, out-of-domain inputs and technical anomalies can be tracked quickly. Survival calibration or performance on a rare event require hindsight and observed events, sometimes months later. Without this feedback loop that matches predictions against real outcomes, telemetry mainly informs on operation and detectable changes, very little on the correctness of results. Those outcomes must also be accessible and representative: if only some patients are followed up, or if results come back only for cases that triggered an alert, monitoring paints a biased picture of performance. The quality of clinical feedback is part of the qualification of the in-service bench. The absence of an immediate alert does not demonstrate that clinical validity is maintained.

The in-service bench must itself be evaluated. Which deviations does it detect, with what delay, with how many false alarms, and which defects does it let through? The thermometer problem continues after delivery: a dashboard can work perfectly without detecting the degradation that matters. The five demonstrations expected of a bench apply to it, with one more constraint: remaining valid in other hands, on a population its author does not control.

Finally, organizing data collection, obtaining clinical outcomes, investigating alerts and re-assessing conclusions is a legitimate recurring cost, nothing like a debt: it exists because use continues and its environment changes. Monitoring, for its part, follows the duration of use: it can be optimized, but it does not disappear; it has to be funded.

The delivered product is therefore no longer limited to the system. It includes the evidence from its initial evaluation, the instrumentation for its monitoring, and the rules for acting on the signals collected. What must extend into operation is the evaluation capability and the flow of information it requires, not necessarily the bench itself. An evaluation that never leaves the factory proves only the factory exit.

Limitations

The main ground is that of a single vendor. The data underpinning the economic argument are not published, so the reader cannot check it; it should be read as a reasoned hypothesis. Other producers may already experience a decreasing marginal cost, and a cost that does not decrease appreciably between models with different intended purposes does not, by itself, establish an absence of learning. The industrial standards cited are paywalled and paraphrased from summaries produced by parties with an

interest in tool qualification, tool vendors or a notified body. Finally, a harmonised standard or a common specification on accuracy, if adopted, would shift part of the argument toward its fitness for the models concerned.

Organizing what survives into the next dossier

Four requirements follow: separate selection from confirmatory evaluation; demonstrate the validity of the bench, level by level; make all reusable evidence revocable; deliver with the system the means to know when its conclusions are no longer supported.

A leaderboard says who comes first on a given protocol. *A conclusion must state what makes it obsolete*.

Sources cited

- AI Act, Regulation (EU) 2024/1689, Articles 6, 12, 13(3), 15, 26(5)-(6), 40, 41, 72 (OJ L, 12/07/2024).
- Regulation (EU) 2026/1744 of 8 July 2026 (Digital Omnibus on AI), OJ of 24/07/2026, entry into force 27/07/2026, https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng [references verified on EUR-Lex; application date of Annex I on 2 August 2028 verified 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; preprint 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, AIM-NASH qualification opinion, 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, MDDT programme, registry current as of 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, Regulatory Science Tools catalogue (CDRH/OSEL), AI/ML domain, https://cdrh-rst.fda.gov [accessed 12/10/2026 during the rigour review].
- FDA, *Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products*, draft guidance, 07/01/2025 (draft status verified in October 2026).
- FDA, *Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions*, final guidance, 2023; ASME V&V 40-2018.
- FDA, *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions*, final guidance (first version December 2024; current version August 2025, posted 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*, September 2022, https://www.fda.gov/media/161815/download
- HAS (French National Authority for Health), descriptive grid for machine-learning medical devices, 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), edition 16, September 2026, https://krogrules.com/news/edition-16-september-2026; AI Act Insight, status of harmonised standards as of 24/08/2026, https://aiactinsight.eu/en/standards/harmonised-standards/ [concordant secondary sources; outcome of the October 2026 JTC 21 plenary not published at the time of writing].
- ISO 13485:2016, clause 7.6; ISO 26262-8:2018, clause 11; RTCA DO-178C and DO-330 (2011) [paywalled standards, cited via: BSI, *Medical devices validation and verification* (webinar); Synopsys, *Automotive tool confidence, ISO 26262* (white paper); LDRA, *DO-330 Test Tool Qualification Technical Briefing* v2.0].
- OECD, principles for the validation of (Q)SARs, November 2004 (JRC, *The Characterisation of (Q)SARs: Preliminary Guidance*, 2005); Regulation (EC) 1907/2006 (REACH), Annex XI, section 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, accepted for presentation (Vienna, 8-11 November 2026).
- Vetillard J., "Same Benchmark Performance, Opposite Care", 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.