npj | digital medicine

Comment

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02987-7

Embedded transparency in artificial intelligence: a prerequisite for equity and representation in AI-enabled clinical trials

Johanna M. C. Blom & Luca Pani

Artificial intelligence is being embedded in clinical trial infrastructure, shaping who is identified, stratified, and analysed. Opaque models risk amplifying existing disparities in the evidence base. We argue that *embedded transparency*, the structural integration of *ex ante* interpretability, demographic auditability, documented uncertainty handling, and stakeholder-relative explanation, is a necessary, though not sufficient, condition for equitable AI-enabled trials, and propose governance recommendations actionable across regulatory regimes.

The representation crisis persists despite decades of effort

The underrepresentation of racial and ethnic minorities, women, older adults, pregnant individuals, and socioeconomically disadvantaged populations in clinical trials is not a novel observation; it is a chronic epistemic failure with profound consequences for the generalisability and justice of medical evidence. Gross et al. demonstrated that although the geographic scope of registration trials has expanded over the past 25 years, the actual diversity of enrolled populations has not substantively improved[1]. In Alzheimer's disease drug development, Franzen and colleagues found that trial participants remained overwhelmingly White (median 94.7%), with no significant temporal trend toward improvement over nearly two decades[2]. Guerra et al. documented that fewer than 5% of cancer patients enrol in clinical trials, with minorities, elderly individuals, and rural populations disproportionately excluded[3].

It is important to acknowledge that this persistence is *not* for want of effort. Substantial industry initiatives, including Merck's public diversity commitments and the patient-engagement programmes of several large sponsors, together with the academic and consortium work of the Multi-Regional Clinical Trials Center at Brigham and Women's Hospital and Harvard, the Clinical Trials Transformation Initiative, and the FDA's 2024 Diversity Action Plans guidance[4], have materially advanced the field. Eligibility criteria have been broadened: Kaur et al. showed that loosening cut-offs in cancer trials increased the eligible population by 78%, with the most substantial gains among historically marginalised groups[5]. The point is not that the field has stagnated. The point is that, despite genuine and well-resourced efforts, structural representation deficits remain, which is precisely why an architectural, rather than purely behavioural or exhortative, solution is required as AI moves from analytic adjunct to operational substrate of clinical trials. As we have argued elsewhere, equity must be embedded as a foundational governance principle, not appended as a post-hoc corrective[6].

AI as both threat and remedy

Documented harms. AI technologies hold transformative potential for clinical research, yet they simultaneously harbour the capacity to amplify long-standing disparities. Cunningham et al., reviewing the integration of AI across the clinical trial lifecycle in cardiovascular research, cautioned that while AI can automate and streamline trial operations, it poses considerable risks including the generation of inaccurate results and the amplification of biases against underrepresented groups[7]. The global landscape of AI clinical trials, as characterised by Kwon through the WHO International Clinical Trials Registry Platform, reveals extreme geographic concentration, with China and the United States accounting for a disproportionate share of registrations, while formal international collaboration remains critically low at only 8.7% of trials[8]. This fragmented topology directly undermines the global generalisability that equity demands.

The mechanisms through which AI can perpetuate inequity are well documented. Corti et al. elucidated how AI algorithms in oncology can mirror the unconscious biases of their developers and the historical disparities embedded in training datasets[9]. In a striking empirical demonstration, Omar et al. evaluated nine large language models across over 1.7 million outputs and found that cases labelled as Black, unhoused, or LGBTQIA+ were more frequently directed toward urgent care or invasive interventions without clinical justification, while high-income-labelled cases disproportionately received recommendations for advanced imaging[10]. The Obermeyer et al. case is now canonical: a widely deployed health-management algorithm used healthcare spending as a proxy for need, systematically under-prioritising Black patients with equivalent clinical complexity, a bias that remained undetected until external audit because the algorithm was opaque to its users[11]. When such architectures are deployed to identify eligible patients, stratify risk, or adjudicate outcomes in clinical trials, the potential for systematic demographic distortion becomes a matter of both scientific validity and regulatory concern.

Remedies, properly specified. Appropriately governed AI can also *reduce* disparities, and intellectual honesty requires that we say how. Four mechanisms are credibly supported by published evidence. *First*, AI can systematically detect inequitable eligibility cut-offs at protocol-design stage, surfacing exclusion patterns that human reviewers miss[5]. *Second*, language-model-assisted simplification of consent and participant-facing materials can extend access to populations with limited health literacy or

for whom the trial site language is not native, provided that simplifications are validated against the source. *Third*, federated learning architectures allow under-resourced sites and lower- and middle-income-country participants to contribute to evidence generation without ceding data sovereignty, addressing the data-monopoly dynamic that Kwon documents[8]. *Fourth*, algorithmic flagging of underrepresented subgroups during ongoing recruitment enables mid-trial corrective action, a capability that retrospective reporting, by construction, cannot provide. None of these mechanisms is automatic; each depends on the AI system being transparent enough that its operation can be audited and corrected. This is the conceptual bridge to the next section.

Defining embedded transparency

Discussion of AI transparency is bedevilled by terminological imprecision. Three conceptual distinctions matter for what follows.

Disclosure versus transparency. *Disclosure* refers to the technical availability of information about a system: published code, model weights, statistical parameters. *Transparency*, in the sense relevant to equity, requires that information reach the right audience in a usable form. A model card published on a corporate repository discharges disclosure; a clinician's, ethics committee's, or participant's ability to interrogate the decision logic that affects them discharges transparency. The two are routinely conflated, but only the latter creates accountability.

Opacity properly specified. Following Burrell[12], we distinguish three forms of opacity: *algorithmic* (the model's internal representations are not interpretable in domain terms), *epistemic* (the model's behaviour cannot be reliably anticipated by domain experts), and *procedural* (the conditions of the model's deployment, retraining, or update are not visible to those affected). Each demands a different remedy. Embedded transparency targets all three; post-hoc explainability addresses, at best, the first.

***Ex ante* interpretability versus *ex post* explainability**. Rudin has argued forcefully[13], and Lipton's earlier taxonomy systematised[14], that *ex ante* interpretability, building the model so that its decision logic is natively inspectable, is preferable to *ex post* explanation methods (LIME, SHAP, saliency maps) when the stakes are high. Post-hoc methods produce approximations of the model's reasoning that are themselves models, and the approximations are, in general, neither faithful nor stable. They retain a complementary diagnostic role but should not be treated as a discharge of the transparency requirement in clinical trial settings.

We accordingly define *embedded transparency* as the structural integration, at the design stage of an AI system intended for clinical trial use, of four properties: (i) *ex ante* interpretability of the model's decision logic by domain experts using the model's native representations; (ii) demographic auditability of inputs, outputs, and intermediate representations; (iii) documented uncertainty handling, including explicit designation of cases on which the model abstains; and (iv) stakeholder-relative explanation, in which different audiences (clinician, regulator, ethics committee, participant) receive explanations calibrated to their decision-making role. All four properties must be present *by design*, not retrofitted. A six-item assessment checklist operationalising this definition is given in Box 1.

Why embedded transparency matters for equity

Embedded transparency cannot architecturally guarantee equity. The well-known impossibility results in algorithmic fairness, Chouldechova[15] and Kleinberg et al.[16], show that calibration, predictive parity, and balance for the positive and negative classes cannot in general be jointly satisfied when base rates differ across groups. No design property can dissolve these trade-offs. What embedded transparency does is make the trade-offs *visible, contestable, and revisable*, which is the most that can honestly be claimed.

Within that more modest frame, the case rests on three operational claims.

First, transparency enables detection. We accept that bias can be partially detected through black-box output analysis (audit studies, demographic disparity testing); the Obermeyer et al. case is itself such a detection. The point is not that opaque-model bias detection is *impossible*, but that it is incomplete, after-the-fact, and dependent on the willingness of system owners to subject outputs to external scrutiny, none of which can be presumed in commercial trial settings. Embedded transparency makes detection a routine property of the system rather than a discretionary act of its owners. Habib et al. demonstrated the substantial variability and structural complexity of clinical trial eligibility criteria, arguing that standardised, computable frameworks are necessary to improve algorithmic transparency in trial recruitment[17].

Second, transparency enables correction. Identifying a bias is necessary but insufficient; transparency must afford the capacity for remediation. Adegunle et al. reviewed AI decision support tools and found that current regulatory models in both the United Kingdom and the United States lack mechanisms to systematically detect or prevent bias[18]. The Obermeyer et al. case is the most concrete illustration of *why this matters*: the racially biased proxy variable was identified only because external researchers were granted access to outputs and their associated demographic metadata. In the absence of mandated transparency, the same architecture could have continued to operate, at scale, indefinitely. We accept that this is a single case; we do not claim it generalises mechanistically. We claim only that it shows what unmandated transparency makes possible, and what mandated transparency would have prevented.

Third, transparency enables accountability. In the domain of clinical trials, where the results directly inform treatment guidelines and regulatory decisions, accountability is a legal as well as an ethical requirement. Jha et al. proposed a comprehensive ethical framework for foundational models in medical imaging that emphasises systematic bias detection, mitigation strategies, and meaningful human oversight[19]. Tubbs and Vazquez identified inclusive data practices, equity audits, and demographically representative training data as essential for operationalising fairness in AI-enabled trials[20]. Accountability without transparency is a contradiction in terms.

Distributed responsibility for embedded transparency

If embedded transparency is the standard, who is responsible for delivering it? The answer is necessarily distributed across five actor classes.

Regulators (EMA, FDA, NMPA). Regulatory authorities should require an algorithmic transparency annex as part of trial protocol submissions for AI-enabled studies, analogous to the inclusion plan we have proposed for representation[6]. The annex would document the system against the six items in Box 1 and disclose the demographic composition of the development and validation cohorts.

Research ethics committees and IRBs. Local committees should treat algorithmic auditability as a condition of approval for AI-enabled trials. Most committees lack the technical capacity to perform a full audit

Box 1 | Assessing embedded transparency

1. Interpretability by design. Is the model's decision logic inspectable by domain experts using the model's native representations, or does interpretation depend on a separate post-hoc explanation method?

2. Demographic auditability. Can inputs, outputs, and intermediate representations be stratified by demographic and clinical subgroup? Is the necessary metadata captured at training and at inference?

3. Uncertainty handling. Does the model explicitly designate cases on which it abstains, and is the demographic distribution of abstentions reported?

4. Stakeholder-relative explanation. Are explanations available in forms appropriate to clinicians, regulators, ethics committees, and trial participants? Is plain-language documentation of model behaviour produced?

5. Versioning and procedural transparency. Are training data sources, retraining schedules, and update conditions documented and made visible to affected stakeholders?

6. Independent auditability. Is the system constructed so that an external auditor with appropriate credentials can reproduce the audit, including on populations not represented in the original development cohort?

themselves, which argues for a tiered model in which a small number of national or regional algorithmic-review bodies issue conformity assessments that local committees can rely on, analogous to the EU AI Act's conformity-assessment architecture for high-risk healthcare systems.

Sponsors and developers. Sponsors and developers bear the architectural burden. To make this economically viable in a competitive market, regulators and payers can deploy three incentive mechanisms: (i) procurement preferences for transparency-by-design tools in publicly funded research; (ii) conditional reimbursement tied to algorithmic auditability for tools deployed within reimbursed care pathways downstream of trials; and (iii) fast-track regulatory designations or fee waivers for tools meeting embedded-transparency criteria. Without such incentives, transparency-by-design tools face structural disadvantage against opaque competitors.

Publishers. Journals can mandate model cards, equity audits, and demographic stratification of validation cohorts as conditions of publication for AI-enabled clinical trial reports. *npj Digital Medicine* and several Nature Portfolio titles already encourage such disclosures; the next step is to require them, with reviewer expertise commensurate to the audit.

Participants and patient communities. Participant-centric data return initiatives, including the European FACILITATE Consortium, demonstrate that participants can be treated as stakeholders in a learning system rather than passive data sources[6]. Embedded transparency makes this position substantive: participants who receive algorithmic decisions about their trial eligibility or stratification have a legitimate claim to plain-language explanation of how those decisions were reached.

Barriers and pathways to implementation

Embedded transparency is not free. Honest engagement with the agenda requires explicit treatment of four classes of obstacle.

Technical barriers. There is a documented performance–interpretability trade-off in some, though not all, model classes. Embedded transparency may also constrain the use of large pretrained foundation models whose internal representations are not natively inspectable. The trade-off is, however, frequently overstated for the structured, tabular data typical of clinical trial enrolment and stratification: Rudin[13] and Caruana et al.[21] demonstrate that interpretable model classes, generalised additive models, prototype-based methods, sparse rule lists, and explainable boosting machines, match or exceed the performance of black-box alternatives on a wide range of clinical tasks. The technical case for opacity in clinical trial AI is, in our view, weaker than commonly assumed.

Economic barriers. Embedded transparency increases development cost and slows time-to-deployment, creating competitive disadvantage in the absence of regulatory parity. The incentive mechanisms described above (procurement preferences, conditional reimbursement, fast-track designations) are intended to internalise this externality. Crucially, the cost of *not* embedding transparency, measured in withdrawn approvals, reputational harm, and downstream litigation, is rarely accounted for in the development business case. Several recent enforcement actions against opaque healthcare algorithms suggest this calculus is shifting.

Institutional barriers. Most IRBs and ethics committees lack the technical capacity to audit AI systems, and the asymmetry of expertise between sponsors and committees is a structural impediment. The tiered-conformity-assessment model proposed above is the most plausible response: concentrate scarce expertise at national or regional review bodies whose assessments local committees can rely on. The EU AI Act's notified-body architecture provides one template; the FDA's Software Pre-Cert programme, though paused, provides another.

Epistemic barriers. Even technically transparent systems can produce explanations that are not useful to their intended audience. Recent work on the audience-relativity of explanation suggests that transparency requirements should be specified *relative to the decision-maker*, clinician, regulator, participant, rather than as a single technical property of the model. Embedded transparency, on the definition we adopt, treats this as a design constraint rather than a downstream concern.

Three regulatory geographies

The global equity argument requires engagement with the three regulatory regimes that, together, govern most AI-enabled clinical trial activity.

European Union. The EU AI Act (Regulation 2024/1689) classifies AI systems used in healthcare as high-risk and imposes specific requirements under Article 9 (risk management), Article 10 (data governance and bias

mitigation), Article 13 (transparency), and Article 14 (human oversight)[22]. Article 10 requires that training, validation, and testing datasets be "relevant, representative, and to the best extent possible, free of errors and complete", language that maps directly onto the demographic auditability item in Box 1. The conformity-assessment regime for high-risk systems provides an institutional architecture into which embedded transparency requirements can be incorporated. The European Health Data Space[23] adds an interoperable substrate for the demographic metadata that auditability requires.

United States. The FDA's Good Machine Learning Practice guiding principles, the 2024 Diversity Action Plans guidance[4], and the Predetermined Change Control Plan framework together establish the most operationally specific regulatory regime to date for AI in clinical settings. Where the EU regime is principle-based, the US regime is procedurally articulated. The two are interoperable in principle but require deliberate harmonisation of demographic categories and audit standards if they are to support cross-jurisdictional evidence.

China. The National Medical Products Administration (NMPA), through its Center for Medical Device Evaluation, has issued AI-specific guidance covering technical review of AI-enabled medical devices and the use of real-world evidence[24]. China accounts for a disproportionate share of registered AI clinical trials[8], which makes the interoperability of NMPA, EU, and FDA standards a precondition for global generalisability of AI-derived clinical evidence. The current state is one of partial alignment: convergence on technical performance standards, divergence on demographic representation and auditability requirements. Embedded transparency offers a substrate on which cross-jurisdictional auditability can be built without requiring full regulatory harmonisation up front.

Embedded transparency in practice: three illustrations

Three recent contributions illustrate what embedded transparency looks like in operation.

Data governance. Okun et al. developed a set of co-created commitments for the ethical sourcing, use, and reuse of patient data in the digital age, prioritising inclusivity, diversity, and equity alongside informed consent and open communication[25]. The substantive contribution is procedural: a transparent commitment architecture against which data practices can be audited by participants and external reviewers, not only by sponsors.

Clinical validation. Hong et al. showed that stroke risk prediction models, including those employing novel machine learning techniques, exhibited systematically worse discrimination in Black individuals than in White individuals, indicating the need to expand both the pool of risk factors and the modelling approaches to address observed racial disparities[26]. The relevant feature for our argument is not the disparity itself but the fact that it was *detectable*: the analysis was possible because demographic stratification of model outputs was permitted. Pham et al., conversely, found that only 7.1% of 141 AI-based diabetes interventions reported participants' ethnic or racial backgrounds, with the average racial distribution being 69.5% White[27]. The absence of demographic auditability rendered equity shortcomings invisible until explicitly investigated, the negative case for embedded transparency.

Subpopulation discovery (NetraAI). Geraci et al. recently reported NetraAI, an explainable AI platform that integrates dynamical-systems modelling, evolutionary feature selection, and large-language-model-generated insights to discover high-effect-size patient subpopulations from high-dimensional clinical data[28]. Applied to a Phase II ketamine trial for treatment-resistant depression (n = 63; retrospective analysis), the platform identified interpretable subpopulation profiles ("Personas") defined by minimal variable combinations (typically 2–4 features with explicit value ranges). Patients whose feature profiles do not permit stable classification are explicitly designated as "unexplainable," rather than being silently absorbed into the dominant class. The performance metrics reported (including up to 100% accuracy on an 8-MRI-feature model) reflect the platform's ability to isolate a coherent subpopulation within a small retrospective cohort, not global predictive accuracy across all patients; the authors are explicit on this point and describe their findings as hypothesis-generating, requiring independent prospective replication[28].

For our argument, NetraAI is illustrative on three of the six embedded-transparency criteria: *ex ante* interpretability of subpopulation criteria (item 1), explicit uncertainty handling through the unexplainable designation (item 3), and plain-language documentation suitable for ethics-committee review through the LLM-mediated translation of Personas into inclusion/exclusion criteria (item 4). Three further criteria, full demographic auditability (item 2), independent reproducibility on under-represented populations (item 6), and documented versioning of training and update procedures (item 5), remain open and would need to be demonstrated before any equity claim could be sustained. We make this scoping explicit because the platform's equity benefits are *potential*, contingent on prospective validation, and not yet established. The same scrutiny should be applied to any platform, including platforms in which the authors have a financial interest, that claims equity contributions from architectural transparency. The financial relationship between L.P. and NetraMark is disclosed in the Competing Interests section.

Trust as a systemic property

The deployment of opaque AI in clinical trials does not merely risk technical bias; it strains the social alliance between researchers and participants on which clinical research depends. As we have argued previously[29], trust and trustworthiness in clinical research are emergent properties of complex systems, developing through transparent communication, participant empowerment, and ethical governance. The introduction of algorithmically opaque decision-making into this ecosystem, particularly in communities whose well-founded distrust derives from documented histories of unethical research practice, threatens to fracture whatever fragile trust has been established. Embedded transparency is therefore not only an algorithmic property but a precondition of the social licence on which clinical research operates.

Conclusion

The integration of AI into clinical trial infrastructure is a structural development, not a methodological refinement. If it occurs without embedded transparency, the resulting systems will inherit, encode, and scale the very disparities that decades of representation work have failed to dissolve. If embedded transparency is structurally integrated, through *ex ante* interpretable architectures, demographic auditability, documented uncertainty handling, stakeholder-relative explanation, distributed responsibility across the actor classes outlined above, and incentives that internalise the cost of opacity, AI can become an instrument for the democratisation of clinical evidence rather than its further concentration. Embedded transparency does not architecturally guarantee equity; no design property does. What it guarantees is that the trade-offs on which equity rests are visible, contestable,

Box 2 | Six governance recommendations

1. Algorithmic transparency annex. Regulators (EMA, FDA, NMPA) should require an algorithmic transparency annex as part of trial protocol submissions for AI-enabled studies, modelled on the inclusion plan annex proposed for representation[6].

2. Mandatory pre-deployment fairness audits. AI systems used in trial eligibility, stratification, or analysis should undergo demographic fairness audits before deployment, with results disclosed in the protocol and reviewed at the conformity-assessment stage.

3. Periodic post-deployment equity monitoring. Aggregate inclusion performance, abstention rates, and demographic stratification of model outputs should be published through privacy-preserving regulatory dashboards, by therapeutic area and geography.

4. Standardised model cards. All AI systems deployed in regulated trials should publish model cards covering training data composition, validation cohorts, performance by subgroup, abstention behaviour, and update history.

5. Tiered conformity assessment. National or regional algorithmic-review bodies should issue conformity assessments that local IRBs and ethics committees can rely on, addressing the asymmetry between sponsor and committee technical capacity.

6. Participant-facing plain-language summaries. Trial participants whose eligibility or stratification is determined by AI should receive plain-language documentation of the decision logic that affects them, on request.

and revisable. That is the most that can honestly be claimed, and it is precisely what the current generation of opaque AI in clinical trials does not provide.

The choice is not between AI and no AI. It is between opaque AI that deepens disparities and embedded-transparent AI that makes equity at least *auditable*. We end with six governance recommendations in Box 2.

Data availability

No datasets were generated or analysed during the current study.

Code availability

Code sharing is not applicable to this Comment, as no code was generated or used to derive new results.

Johanna M. C. Blom[1,2] ✉ **& Luca Pani**[1,2,3]

[1]Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia (UNIMORE), Modena, Italy. [2]Center for Neuroscience and Neurotechnology, University of Modena and Reggio Emilia (UNIMORE), Modena, Italy. [3]Department of Psychiatry and Behavioral Sciences, University of Miami, Miami, FL, USA.

✉e-mail: joan.blom@unimore.it

Received: 12 March 2026; Accepted: 1 July 2026;
Published online: 11 July 2026

References

1. Gross, A. S., Harry, A. C., Clifton, C. S. & Della Pasqua, O. Clinical trial diversity: an opportunity for improved insight into the determinants of variability in drug response. *Br. J. Clin. Pharmacol.* **88**, 2700–2717 (2022).
2. Franzen, S. et al. Diversity in Alzheimer's disease drug trials: the importance of eligibility criteria. *Alzheimers Dement.* **18**, 810–823 (2021).
3. Guerra, C. E., Fleury, M. E., Byatt, L. P., Lian, T. & Pierce, L. Strategies to advance equity in cancer clinical trials. *Am. Soc. Clin. Oncol. Educ. Book* **42**, 1–11 (2022).
4. U.S. Food and Drug Administration. Diversity Action Plans to Improve Enrollment of Participants from Underrepresented Populations in Clinical Studies: Draft Guidance for Industry (2024).
5. Kaur, M. et al. Broadening eligibility criteria and diversity among patients for cancer clinical trials. *NEJM Evid.* **3**, EVIDoa2300236 (2024).
6. Blom, J. M. C. et al. Embedding equity in clinical research governance. *Nat. Med*. https://doi.org/10.1038/s41591-026-04232-w (2026).
7. Cunningham, J. W. et al. Artificial intelligence in cardiovascular clinical trials. *J. Am. Coll. Cardiol.* **84**, 2051–2062 (2024).
8. Kwon, C. Y. Beyond the growth: a registry-based analysis of global imbalances in artificial intelligence clinical trials. *Healthc. (Basel)* **13**, 2018 (2025).
9. Corti, C. et al. Artificial intelligence in cancer research and precision medicine: applications, limitations and priorities to drive transformation in the delivery of equitable and unbiased care. *Cancer Treat. Rev.* **112**, 102498 (2022).
10. Omar, M. et al. Sociodemographic biases in medical decision making by large language models. *Nat. Med.* **31**, 1873–1881 (2025).
11. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
12. Burrell, J. How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data Soc.* **3**, 1–12 (2016).
13. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
14. Lipton, Z. C. The mythos of model interpretability. *Commun. ACM* **61**, 36–43 (2018).
15. Chouldechova, A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* **5**, 153–163 (2017).
16. Kleinberg, J., Mullainathan, S. & Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *Proc. 8th Innov. Theor. Comput. Sci. Conf. (ITCS 2017)* **43**, 1–23 (2017).
17. Habib, D. R. S., Mahajan, I., Evancha, B., Micheel, C. & Fabbri, D. Computational framework for structuring and analyzing clinical trial criteria for AI-guided fine-grained matching. *J. Med. Syst.* **49**, 168 (2025).
18. Adegunle, F. et al. Bias and oversight in clinical AI: a review of decision support tools and equity frameworks. *J. Gen. Intern. Med*. https://doi.org/10.1007/s11606-026-10229-5 (2026).
19. Jha, D. et al. Ethical framework for responsible foundational models in medical imaging. *Front. Med. (Lausanne)* **12**, 1544501 (2025).
20. Tubbs, A. & Vazquez, E. A. Digital twins in increasing diversity in clinical trials: a systematic review. *J. Biomed. Inform.* **169**, 104879 (2025).
21. Caruana, R. et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In *Proc. 21st ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* 1721–1730 (2015).
22. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* **L1689** (2024).
23. European Union. Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847. *Off. J. Eur. Union L*, 2025/327 (2025).
24. National Medical Products Administration, Center for Medical Device Evaluation. *Guidelines for the Technical Review of Artificial Intelligence Medical Devices* (NMPA-CMDE, Beijing, 2022).
25. Okun, S. et al. Commitments for ethically responsible sourcing, use, and reuse of patient data in the digital age: cocreation process. *J. Med. Internet Res.* **25**, e41095 (2023).
26. Hong, C. et al. Predictive accuracy of stroke risk prediction models across Black and White race, sex, and age groups. *JAMA* **329**, 306–317 (2023).
27. Pham, Q., Gamble, A., Hearn, J. & Cafazzo, J. A. The need for ethnoracial equity in artificial intelligence for diabetes management: review and recommendations. *J. Med. Internet Res.* **23**, e22320 (2021).
28. Geraci, J. et al. Explainable AI-driven precision clinical trial enrichment: demonstration of the NetraAI platform with a phase II depression trial. *npj Digit. Med.* **8**, 749 (2025).
29. Blom, J. M. C., Rivi, V., Tascedda, F. & Pani, L. Building trust in clinical research: a systems approach to ethical engagement and sustainable outcomes. *Front. Pharmacol.* **16**, 1570899 (2025).

Acknowledgements
This work has been shaped by ongoing discussions within the broader research and clinical community. No specific funding was received for this Comment.

Author contributions
J.M.C.B. conceived the Comment, developed the conceptual and policy framing, and led the writing of the manuscript. L.P. contributed the regulatory-science and translational perspective and provided critical revision for intellectual content. Both authors approved the final version and take responsibility for its content.

Competing interests
L.P. has the following competing financial interests relevant to this Comment: L.P. is a co-author on Geraci et al.[28], which is cited as one of three illustrations of embedded transparency in practice in this Comment; L.P. holds shares in NetraMark Holdings and has a consultancy relationship with NetraMark Corp., the developer of the NetraAI platform discussed in that reference. L.P. additional financial disclosures (last 24 months): AbbVie, USA; Acumen, USA; Aicure, USA; Alexion, Italy; BCG, Switzerland; Astra-Zeneca, Italy; Boehringer Ingelheim International GmbH, Germany; EDRA-LSWR Publishing Company, Italy; EnZeta Immunotherapies, USA; GH-Pharma, Ireland; GLG-Institute, USA; Immunogen, USA; Johnson & Johnson, USA; LB-Pharmaceuticals, USA; Magdalena BioSciences, USA; MSD, Italy; Sanofi-Aventis-Genzyme, France and USA; Lundbeck, Denmark and Italy; Napo-Pharma, USA and EU; NetraMark, Canada; Pfizer Global, USA; Relmada Therapeutics, USA; Takeda, USA. Shares/options: EnZeta, Relmada, NetraMark. L.P. has no non-financial competing interests to declare. J.M.C.B. declares no competing financial or non-financial interests. Neither author holds an editorial role at *npj Digital Medicine* or at any other Nature Portfolio journal.

Additional information
Correspondence and requests for materials should be addressed to Johanna M. C. Blom.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.