OPEN ACCESS

Digital Medicine REPORT

The Innovation Medicine

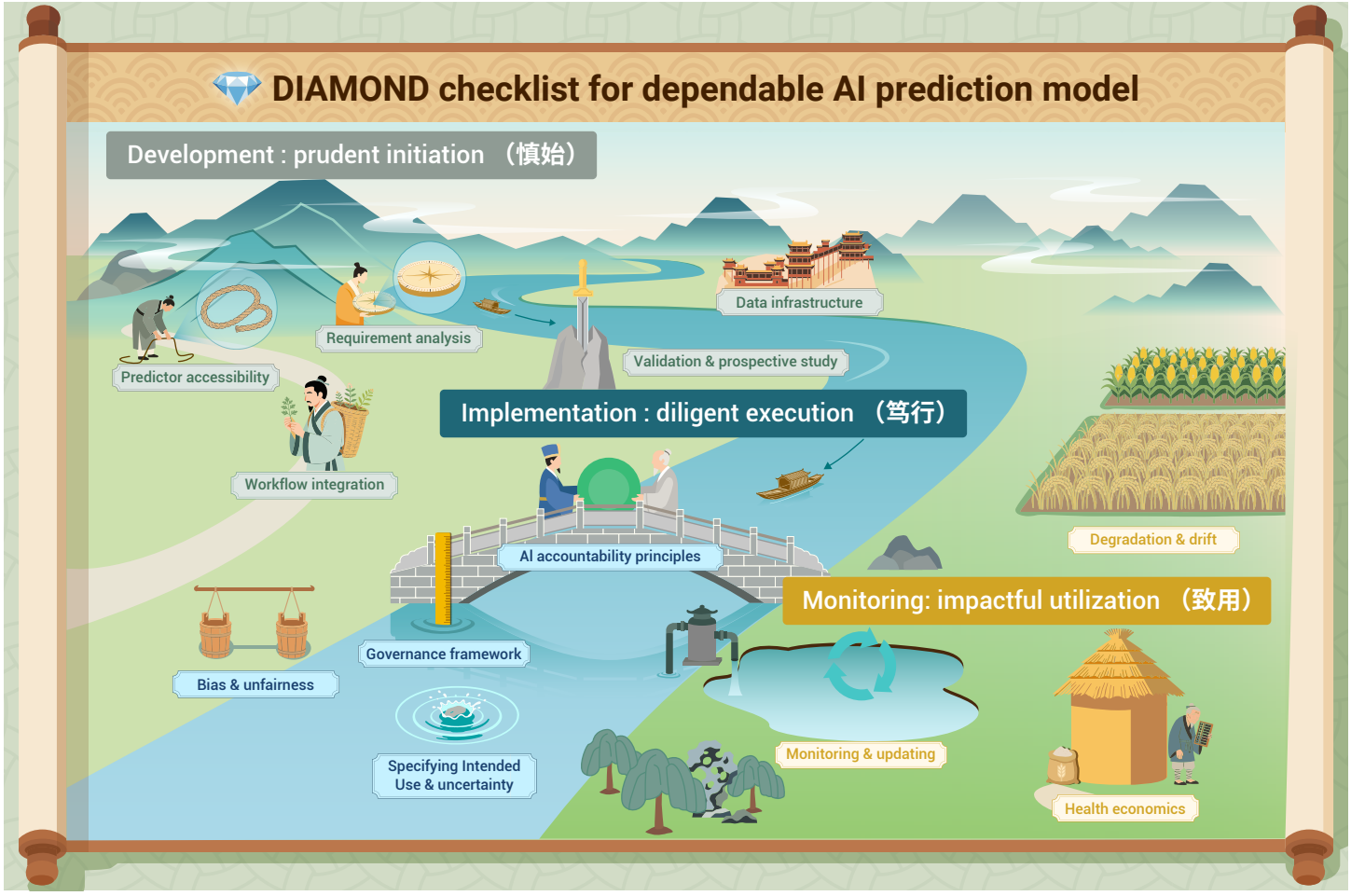
From code to continuous care: Key considerations for developing and implementing dependable AI prediction models in healthcare delivery

Qiaochu Wei,[1,2] Miao Cui,[1,2] Yujuan Qi,[3] Qiang Zhang,[3] Zhonghua Liu,[4] Canqing Yu,[1,2,5] F.D. Richard Hobbs,[6,7] David C. Christiani,[8] Yi Li,[9] Ruyang Zhang,[10] Feng Chen,[10] Guoshuang Feng,[11,*] Yuantao Hao,[1,2,5,*] and Yongyue Wei[1,2,5,*]

*Correspondence: ywei@pku.edu.cn (Y.W.); haoyt@bjmu.edu.cn (Y.H.); fengguoshuang@mail.ccmu.edu.cn (G.F.)

Received: February 7, 2026; Accepted: June 10, 2026; Published Online: June 11, 2026; https://doi.org/10.59717/j.xinn-med.2026.100229

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Care delivery needs the Development, Implementation, And MONitoring for Dependable AI model framework.
- Development requires clear use cases, reliable data, standardized predictors, and rigorous validation.
- Implementation requires workflow integration, transparent governance, fairness, and accountability.
- Monitoring requires ongoing evaluation, model updates, and cost-effectiveness assessment.
- The proposed DIAMOND checklist offers a practical path from promising algorithms to dependable care.

REPORT Digital Medicine

www.the-innovation.org/the-innovation-medicine

From code to continuous care: Key considerations for developing and implementing dependable AI prediction models in healthcare delivery

Qiaochu Wei,[1,2] Miao Cui,[1,2] Yujuan Qi,[3] Qiang Zhang,[3] Zhonghua Liu,[4] Canqing Yu,[1,2,5] F.D. Richard Hobbs,[6,7] David C. Christiani,[8] Yi Li,[9] Ruyang Zhang,[10] Feng Chen,[10] Guoshuang Feng,[11,*] Yuantao Hao,[1,2,5,*] and Yongyue Wei[1,2,5,*]

[1]Center for Public Health and Epidemic Preparedness & Response, Peking University, Beijing 100191, China
[2]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China
[3]Qinghai Provincial people's Hospital, Qinghai Province Research Center for High Altitude Medicine, Xining, Qinghai 810001, China
[4]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York NY 10032, USA
[5]Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China
[6]Nuffield Department of Primary Health Care Sciences, University of Oxford, Oxford OX2 6GG, UK
[7]Oxford Institute of Digital Health, University of Oxford, Oxford OX2 6GG, UK
[8]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston MA 02115, USA
[9]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA
[10]Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China
[11]Big Data Center, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health,Beijing 100045, China
*Correspondence: ywei@pku.edu.cn (Y.W.); haoyt@bjmu.edu.cn (Y.H.); fengguoshuang@mail.ccmu.edu.cn (G.F.)

Received: February 7, 2026; Accepted: June 10, 2026; Published Online: June 11, 2026; https://doi.org/10.59717/j.xinn-med.2026.100229

Citation: Wei Q., Cui M., Qi Y., et al. (2026). From code to continuous care: Key considerations for developing and implementing dependable AI prediction models in healthcare delivery. *The Innovation Medicine* **4**:100229.

Predictive models in healthcare are widely published, yet few achieve routine clinical use due to gaps in methodological rigor, workflow integration, and governance. Existing guidelines primarily focus on clinical settings, with few addressing broader healthcare delivery contexts. We propose a practical framework for translating code to continuous care: Development, Implementation, And MONitoring for Dependable AI prediction model (DIAMOND). Models must be built for explicit clinical use cases, supported by interoperable data, standardized predictors, and rigorous validation with prospective designs. Translation into practice requires workflow integration, proportionate regulatory oversight of intended use, transparency, uncertainty, bias, and accountability, and continued post-deployment evaluation—testing transportability across settings, monitoring and updating for data shift and performance degradation, and assessing health-economic impact to inform iterative refinement. By systematically linking these stages, the DIAMOND framework provides a structured pathway for advancing AI predictive models from promising algorithms to dependable clinical tools.

INTRODUCTION

The medical field is witnessing a surge in the development of predictive models.[1] Several are widely used in clinical practice, both in traditional analogue form and as integrated digital tools.[2] However, only a limited number of models are incorporated into electronic health record (EHR), and a substantial proportion of models remain underutilized.[3] Under such a condition, guidance on prediction model development for better clinical practice was established. DECIDE-AI provides a consensus reporting guideline for early-stage, live clinical evaluation of AI-driven decision-support systems, whereas TRIPOD+AI extends TRIPOD to harmonized reporting of prediction model development and validation using regression or machine-learning methods.[4–5] However, the central bottleneck is often safe and sustainable translation into prediction-action-intervention workflows rather than model performance metrics alone. Moreover, the FDA's SaMD (including AI/ML-enabled) guidance adopts a total product lifecycle perspective. The 2025 FUTURE-AI framework encompasses the entire AI lifecycle, including six guiding principles: fairness, universality, traceability, usability, robustness, and explainability. However, these guidelines primarily focus on clinical models and lack a broader perspective on the healthcare system. Furthermore, they adopt a single perspective, neglecting to assess clinical needs from the combined viewpoints of doctors, the healthcare system, and agencies.[6–9]

To address these existing gaps, we introduce the DIAMOND checklist — a structured framework designed to guide the development, implementation, and ongoing monitoring of reliable AI prediction models. The checklist outlines three essential steps in design and translation, including aligning models with real-world needs, rigorously validating performance, proper feedback mechanisms and ensuring continuous monitoring and updates to maintain effectiveness and ethical compliance. By considering the lifecycle of AI models from a systematic and comprehensive perspective, we can facilitate smoother integration of these models into healthcare systems, enhancing their practical utility and impact.

MATERIALS AND METHODS

To develop the DIAMOND framework, we conducted a structured literature review of evidence on the development, implementation, and post-deployment governance of artificial intelligence-based clinical prediction models. Two reviewers independently searched PubMed, Embase, and Web of Science for publications published between 1 January 2019 to 31 May 2026 using terms related to artificial intelligence, machine learning, deep learning, clinical prediction, healthcare, development, implementation, post-deployment monitoring and workflow integration. The search was supplemented by grey-literature sources including medRxiv, Google Scholar, guideline repositories, and regulatory documents, as well as manual screening of reference lists.

Eligible sources included peer-reviewed studies, consensus statements, regulatory guidance, and high-quality preprints addressing diagnostic, prognostic, triage, decision-support, or algorithm governance models across any healthcare specialty. Titles, abstracts, and full texts were screened independently by two reviewers; disagreements were resolved through discussion with a third reviewer.

To illustrate regional heterogeneity relevant to model implementation, we analysed four population-based datasets: UK Biobank (UKB; >500,000 participants aged 40–69 years, recruited across the UK in 2006–2010), China Kadoorie Biobank (CKB; >512,000 adults from ten Chinese regions, recruited in 2004–2008), Yinzhou electronic health records (EHR; residents aged ≥18 years since 2009), and Shenzhen health examination records (adults aged ≥ 65 years). For each cross-cohort comparison, participants were first restricted to overlapping age ranges and then harmonized using 1:1 nearest-neighbour propensity score matching without replacement, with a caliper of 0.2 standard deviations of the logit propensity score. Propensity scores were estimated from established lung cancer risk models **(details in the Supplemental materials and methods)**, which incorporate demographic, smoking-related, respiratory, occupational, metabolic, and cancer-history variables. Covariate balance was evaluated using standardized mean differences. Regional risk patterns were compared using incidence rates, Kaplan–Meier curves, and Cox proportional hazards models.

RESULTS

Factors to consider after implementation: long-term surveillance and lifecycle governance of deployed models

In clinical prediction modelling, development involves defining the study aim, selecting predictors, preparing data, and estimating parameters, while crucially aligning these steps with real-world clinical needs.[10] If developers neglect real-world clinical needs, or if the healthcare system is not prepared to embed the AI model into routine care, the model may remain a paper prototype. Accordingly, this stage should address five key aspects: application scenarios, predictor accessibility, model validation, workflow integration, and system infrastructure.[11]

Clarify the deployment scenarios of predictive models. It is essential for developers to clearly define a model's objective and intended deployment scenario to ensure alignment with real-world needs. In the case of hospital-based diagnostic support, models' clinical impact must be measured. Meanwhile, for personal health management, real-time data collection is essential. Besides, for primary, secondary, and tertiary care settings, screening efficiency, incomplete-data handling, and diagnostic precision, are emphasized respectively.[12] Primary care AI applications emphasize simplicity, transparency, and clinical relevance at the frontlines,[13] while more complex tertiary applications often require deeper data integration. Achieving such integration depends on standardized and interoperable clinical data across the end-to-end workflow.

Assess whether data and system infrastructure align with model requirements. The integrity of healthcare AI fundamentally depends on high-quality data, especially electronic health records (EHRs), while they were designed to document individual episodes of care, and sometimes administrative tasks such as facilitate the payments for that care, rather than supporting clinical research.[14] This underscores the necessity of systematic data provenance review. Data collection should follow standardized protocols to facilitate integration and exchange.[15] For example, interoperability and data modeling standards (e.g., health level 7 fast healthcare interoperability resources, HL7 FHIR) ensure consistent capture and exchange of clinical data, while common data models (e.g., observational medical outcomes partnership common data model, OMOP CDM) harmonize multi-site observational data. Annotation should involve expert(a domain-qualified clinician, with their role, specialty, years of experience, or certification reported) validation to ensure accuracy. Data storage must prioritize security and accessibility, ensuring data remains intact for analysis, and data management should include classification, organization, coding, and retrieval to preserve data quality.

Moreover, the integration of multimodal data sources (e.g., images, video records, diagnostic notes) further complicates data infrastructure requirements. Unfortunately, many healthcare facilities lack adequate IT infrastructure, including reliable internet connectivity, computing resources, and standardized data formats. Although we have developed federated learning infrastructures to integrate heterogeneous datasets across multiple hospitals, the underlying data still reflected variability in completeness and quality.[16] Therefore, before deploying prediction models, it is important to assess the healthcare system's digital maturity, such as the quality of the data captured. If the system is ready to embed an AI model into routine workflows, the next step is to determine whether the required inputs have been consistently defined, standardized, and validated.

Assess whether predictive factors' testing systems are universally applicable. Identifying high-impact predictive factors while neglecting their clinical accessibility, measurement standardization, and generalizability often widens the gap between research and practice,[17] especially in low-middle-income settings.[18] Candidate factors should be both clinically accessible and demonstrably associated with the outcomes of interest.[19] They should also be sufficiently prevalent in the target population and be recorded well before the anticipated outcome to allow for timely analysis.[20]

Measurement methods must be widely recognized, accessible, and standardized to ensure reliability and validity across settings and populations.[19] For example, the measurement equipment for hypersensitive C-reactive protein (hs-CRP) differ across institutions such as use of different test reagents with differing reference ranges.[21] Therefore multicenter data typically require normalization according to reference interval of each value.[22] Similarly, non-targeted, non-absolute-quantitative biomarkers generated from mass-spectrometry–based "omics" assays, should not be used in risk prediction models, as they lack measurable and reproducible clinical readouts. These challenges highlight the need to advance clinical laboratory standardization by adopting internationally recognized testing methods and calibration standards. With key factors standardized, prediction model development is facilitated using step-by-step methodological guidance,[10] but clinical implementation ultimately depends on rigorous validation.

Consider multidimensional and prospective validation in research design. A comprehensive and robust evaluation framework encompasses several key aspects: discrimination, calibration, clinical utility, and, if applicable, economic evaluation.[23-25] Notably, while "accuracy" is frequently prioritized, it represents a threshold-dependent classification metric that often obscures the model's global performance.[7] Internal validations are instrumental in model selection and overfitting prevention, but limitation is inherent due to potential optimism bias. Therefore, external validation is indispensable for verification (Figure 1). We focus on discrimination (AUC) as a global measure, while classification metrics (sensitivity/specificity) are reported contingent on specific decision thresholds relevant to the clinical setting. Despite the abundance, few AI prediction models undergo rigorous external validation or prospective trial examination. When assessed in real-world settings,[26-28] validation should go beyond accuracy and precision, incorporating robustness, adversarial testing, and fairness evaluations. It is also essential to balance ideal metrics with data, cost, and regulatory constraints.

Randomized controlled trials (RCTs) are the gold standard for validating the efficacy of AI models. They are used to evaluate the effects of interventions, with appropriate control procedures such as sham-models.[29] The sham model acts as a placebo control for the AI system, providing the scientific foundation upon validation. It needs to be developed specifically and evaluated for its rationality, with typical training objectives to achieve the same specificity as the real model while having a sensitivity of zero.[30] Zhang and his colleagues designed a study demonstrating that, compared to where radiologists worked independently, the sham-AI model did not impact their diagnostic performance negatively.[31] Prospective evaluation, ideally in pragmatic trials or target-trial emulation, remains the most reliable way to establish clinical impact.[32] Nevertheless, RCT-level evidence does not make an AI model "plug-and-play"; additional software engineering is usually needed for workflow integration.

Consider whether models can be integrated into healthcare processes. The key to integrating prediction models into clinical workflows is the changes of doctors' work mode. For example, automated pop-up alert systems provide real-time and actionable information, improving situation awareness and clinical management in pediatric sepsis.[33] However, the frequency and threshold for alert generation needs to be calibrated to avoid excessive false positives.[34] Although risks of automation bias, alert fatigue, and ongoing human oversight to interpret alerts exist, clinicians generally appreciate the support provided by automated alerts, concerning about increased workload and potential overreliance necessitating balanced system design.[34] Therefore, the content of alerts should be concise, contextually relevant, and actionable, ideally providing not only risk scores but also explanatory information or suggested interventions. For informed decision-making and confidence in using predictive models, four key informational elements are essential: (1) information about model developers and users, (2) methodology, (3) peer review and model updates, and (4) population validation.[35] Furthermore, expert interviews revealed equity-related concerns, particularly regarding how the model's application might intersect with structural inequities. For example, factors like the capacity of advanced care planning (ACP) workflows and discharge timing can significantly impact the net benefits of model-triggered interventions for advanced care planning.

Factors to consider during implementation: establishing a robust framework for model governance

Real-world metrics may disrupt the feasibility of model implementation. Without regulatory oversight across the total product life cycle, deployment is unlikely to be sustainable (Figure 2). From the perspective of regulatory agencies, the assessment of medical AI products involves pre-market evaluation, intended use and predictive uncertainty, model bias and unfairness and

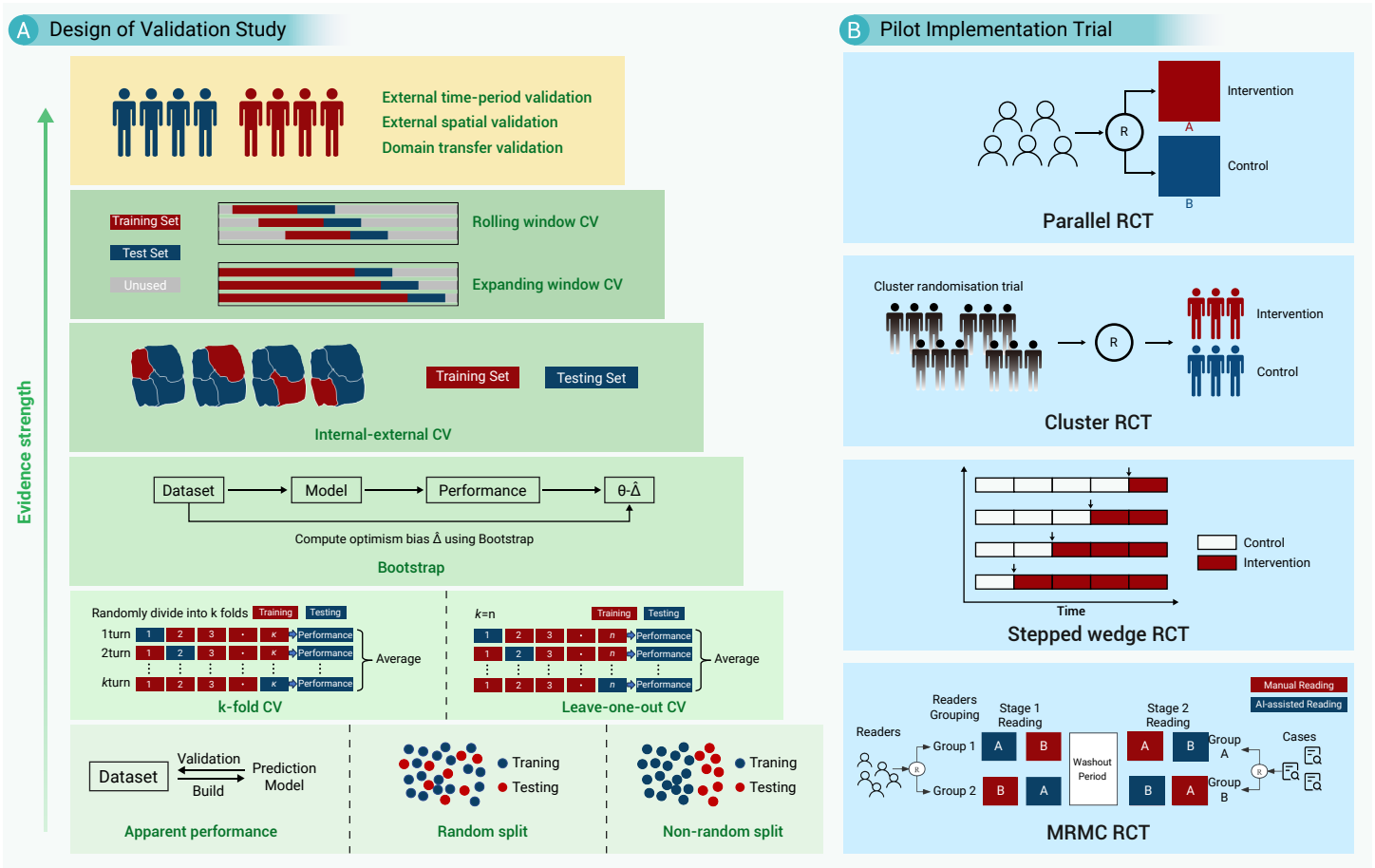


Figure 1. Hierarchical evidence strength pyramid for clinical prediction model validation.

accountability arrangements.

Assessing models' transparency, explainability and reproducibility Transparency in AI decision-making is essential for clinician's trust and patient's sense of safety.[36] Agencies should ensure transparency and reproducibility by several measures: Assessing study design protocols, details of data curation, feature engineering, model training and validation processes; assessing de-identified feature data or data simulation scripts in compliance with ethical requirements; assessing trained model weights, model algorithms or online prediction tools provided by developers{Yang, 2024 #38}. In line with best practices for reproducible research, model algorithm should be hosted on code repositories (e.g., GitHub) to ensure robust version control. Beyond model's transparency, demographic representation within training datasets is critical to achieving efficacious and equitable outcomes.[37] AI models must be rigorously validated across diverse patient populations to ensure generalizability and minimize biases. Regulatory frameworks must be flexible enough to adapt to the rapid advancements in AI technology, ensuring that certification processes do not stifle innovation while still safeguarding patient health.[38] Also, data privacy, security, and informed consent must be integrated into the regulatory frameworks to protect patient rights and sustain public trust in AI technologies.[39] Furthermore, these frameworks must address the needs of frontline clinicians by distinguishing between "local" explanations tailored to individual patients and "global" logic governing the model as a whole. Clinical prediction models, which rely on longitudinal data and may evolve over time. Broad consent with dynamic consent, facilitates patient authorization for the use of health data and biological samples in future research, while allowing for ongoing updates and withdrawals of consent.[40] In 2022, national institute for health and care excellence(NICE) updated its Digital Health Evidence Standards Framework (ESF) to include AI with adaptive algorithms, aligning regulatory requirements to address the need for "sustainable evidence and continuous monitoring." Similarly, the European Medicines Agency (EMA) reflection paper on AI in the medicinal product lifecycle emphasizes the need for post-authorization surveillance and adaptability, proposing frameworks for algorithmic updates that do not require full re-clearance, ensuring ongoing monitoring under established legal frameworks like data protection and cybersecurity. This advance enhances dependability in AI prediction models and promoting long-term participation.[41]

Specifying intended use and operationalizing predictive uncertainty. The intended use of a clinical AI model extends beyond simple inclusion/exclusion criteria; it must clearly define the intended users, patient groups, and clinical context of use. For instance, applying a model developed on patients aged 40–70 to all-age adult population is not a minor adjustment, but raises questions of transportability. This requires evidence that the deployment environment matches the development setting in population characteristics, predictor availability, and outcome definition—a point also emphasized in Good Machine Learning Practice principles, which align model design and evaluation with intended use.[42]

Clinical decisions seldom depend solely on point predictions; thus, uncertainty should be quantified and communicated transparently. This includes uncertainties arising from data limitations, model assumptions, and dataset shifts. Presenting confidence intervals or band of prediction values helps users distinguish between high-risk cases supported by strong evidence and those with fragile evidence. Explicit uncertainty quantification is increasingly regarded as a key factor in safety and decision quality, bridging statistical output with human factors in clinical practice.[43] For example, confidence estimates for machine learning models enhance reliability, analyzing uncertainties improves clinical decision accuracy and robustness.[43–44] Regarding hallucination detection in unsupervised LLM dialogue, a knowledge uncertainty-based measure has been proposed, with uncertainty quantification identified as key for detecting unreliable outputs when external validation is not possible. Effective communication of uncertainty not only supports safer decisions but also relates to broader concerns of model unfairness and bias.

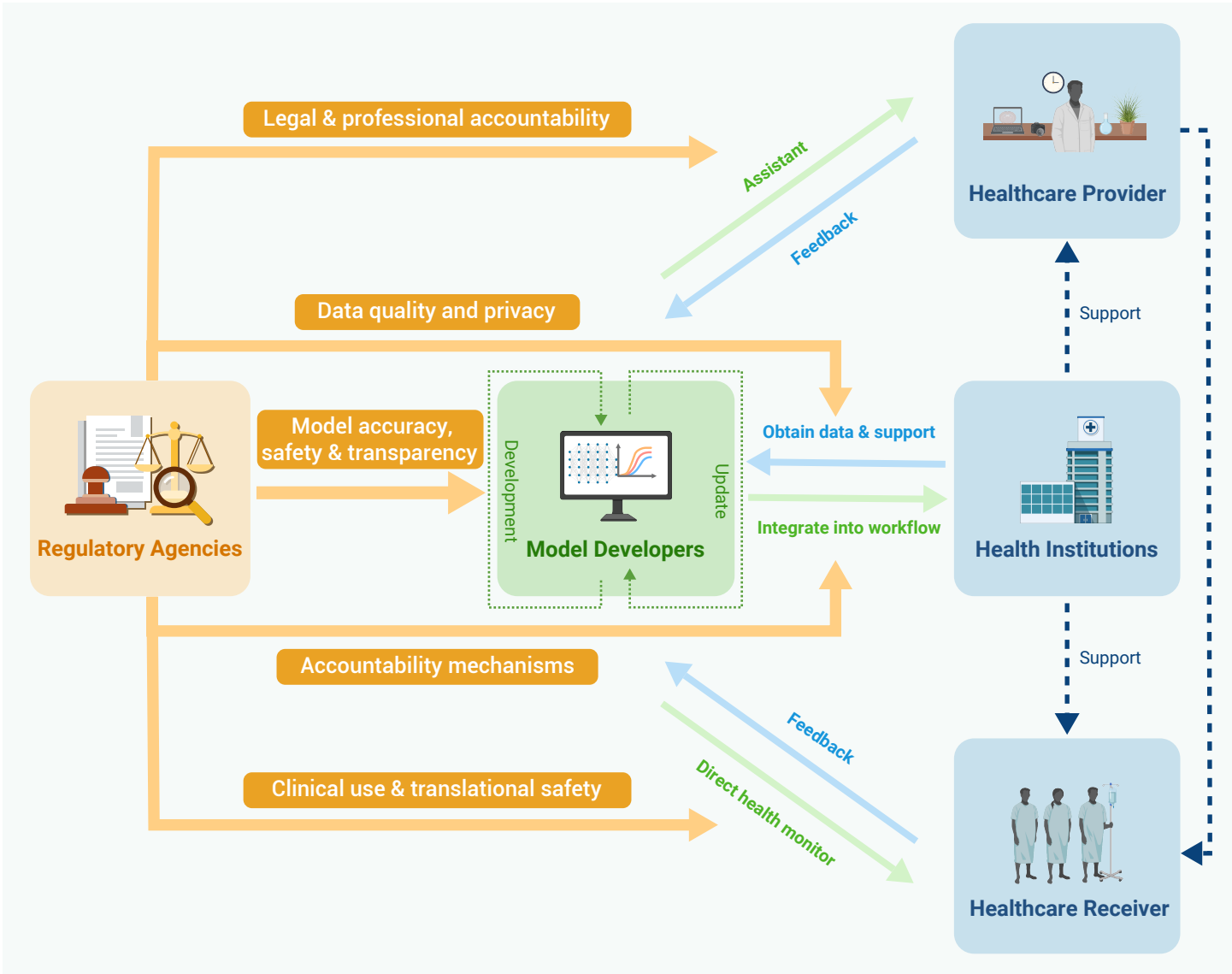


Figure 2. Governance structure and responsibilities of key stakeholders in clinical prediction model This figure illustrates the interactions among model developers, healthcare providers, health institutions, healthcare receivers, and regulatory agencies across the AI model development-implementation-regulatory lifecycle. Feedback loops support model refinement, while regulatory agencies oversee all stakeholders by defining standards for data governance, model certification, transparency, and post-market monitoring, ensuring safe and accountable clinical use.

Addressing model bias and unfairness. Sample selection bias arises when real-world patients differ from the training population, leading to uneven model performance across demographic or clinical subgroups. This bias can arise from various factors, such as differences between the real-world population and the training samples, as well as clinicians' understanding of the AI model. In contrast, end-user bias stems from physician mistrust in AI outputs, complex user interfaces, and documentation burdens, which lead to inconsistent use of AI-generated recommendations. In this context, adhering to established guidelines and frameworks is essential to ensure sustained accuracy and equity of AI algorithms.[45] Bias in clinical AI should be addressed in two stages: first, by reducing sample selection bias through the use of representative data, alignment with the target population, local validation, drift monitoring, recalibration, and site-specific adaptation; and second, by mitigating end-user bias through interpretable design, workflow integration, usability optimization, training, and human-factor safeguards that promote appropriate trust and consistent use.

Unfairness is often driven by bias. Detecting unfairness requires translating fairness into auditable quantitative evidence by systematically comparing performance, error types, calibration, and formal fairness metrics across subgroups defined by sensitive attributes and their intersections. At the same time, it is necessary to examine whether the model objectives and proxy variables themselves introduce structural disparities, thereby identifying where unfairness occurs and assessing its severity. Advanced techniques like the Double Deep Q-Network (DDQN) address model unfairness by reformulating clinical classification tasks as reinforcement learning problems, with a custom reward function optimizing both "clinically accurate classification" and fairness regarding sensitive attributes (e.g., race, hospital site).[46] Notably, The FDA's Proposed Regulatory Framework for AI/ML-based Software as a Medical Device (SaMD) emphasizes the need for real-world performance monitoring, including tracking model performance to identify existing or emerging bias.[47] While this framework places meaningful responsibility on AI model developers, healthcare institutions must bear the responsibility to maintain awareness of their local populations and clinical practice shifts that are commonly opaque to commercial platforms. Beyond assessing model bias and unfairness, regulatory agencies should establish clear accountability arrangements.

Navigating moral accountability in AI-driven clinical decision-making. Determining moral accountability in complex socio-technical systems is both significant and inherently challenging. A key debate in patient safety concerns how responsibility should be balanced between individual clinicians and their organizations.[48] Clinical authority is not delegated to the AI system. The final decision remains with the clinician, as human clinicians tend to maintain

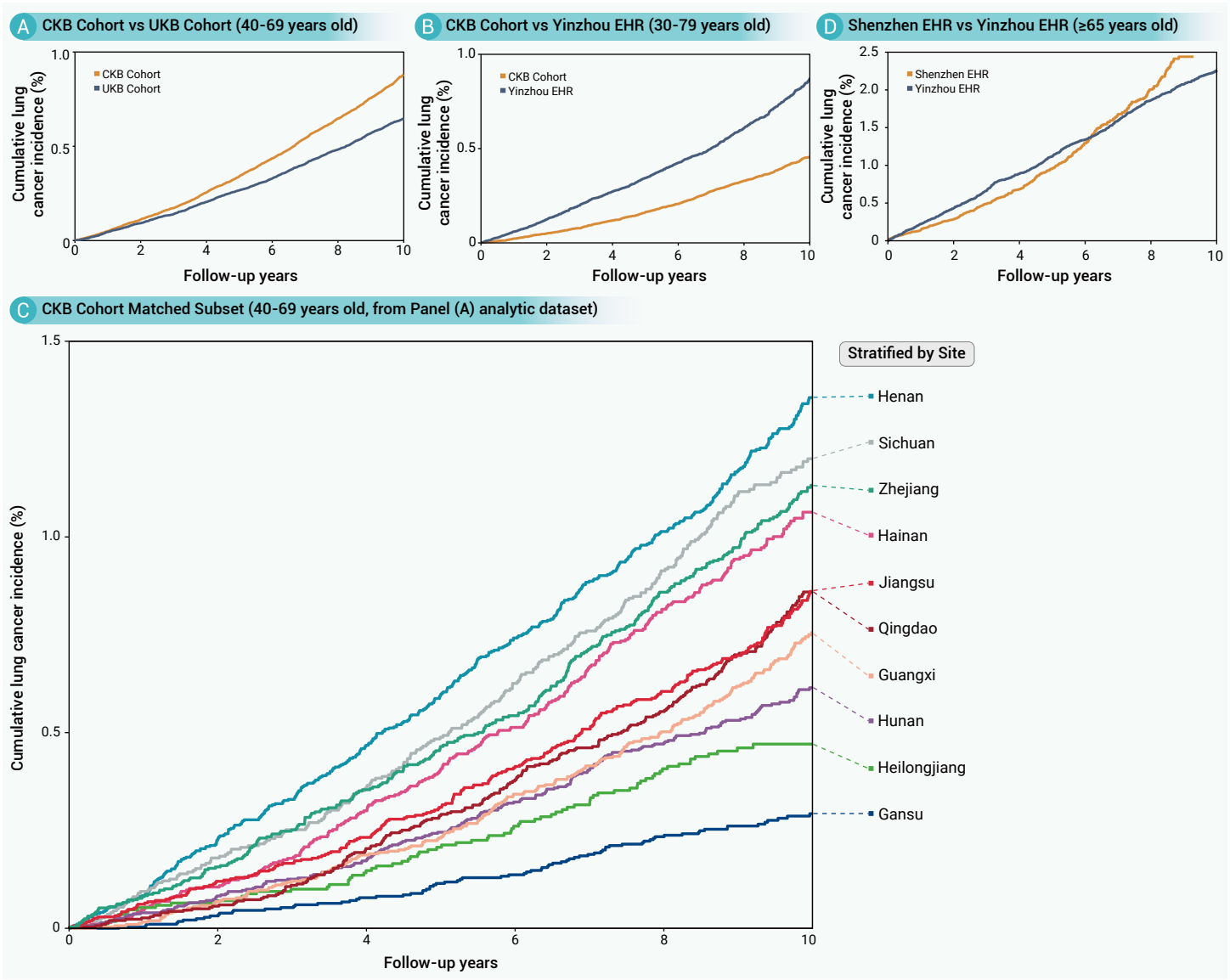


Figure 3. Cross-regional divergence in disease incidence trend — lung cancer as example This figure shows Kaplan–Meier curves of cumulative lung cancer incidence (y-axis) over time in years of follow-up (x-axis) for comparable participants from different cohorts that are matched with traditional lung cancer risk factors. (A) The figure compares subpopulations from UKB versus CKB to illustrate cross-country differences in lung cancer incidence trend among individuals with similar characteristics of lung cancer risk factors. (B) The figure contrasts subpopulations of CKB with Yinzhou EHR to highlight differences between a historic, large-scale cohort covering 10 provinces in China and contemporary EHR of eastern coastal city. (C) The figure depicts the incidence trends, stratified by the 10 provinces within the analytical subset of CKB cohort shown in Fig. B, to illustrate domestic China cross-regional heterogeneity present. (D) The figure compares Yinzhou EHR and Shenzhen EHR to reflect regional variations in lung cancer incidence trend within China.

treatment despite digital systems potentially recommending abrupt therapeutic changes.[49–50] However, when clinicians alter their decisions based on automated recommendations, the boundaries of responsibility become blurred.[51] This underscores the need for shared accountability in the "automation trade-off." Clinical institutions must ensure that AI systems are safe and reliable by requiring manual review of AI recommendations, maintaining decision logs, and enabling timely interventions in the event of anomalies. Additionally, developers and healthcare institutions must implement transparent, auditable designs and comprehensive training programs to mitigate the risk of automation bias.[52]

Numerous scholars have debated what constitutes sufficient control or understanding for moral responsibility.[53] However, when AI systems are involved in clinical decision-making, the extent to which clinicians can be held accountable for patient harm remains unclear. They lack direct control over system outputs, and many AI models are inherently opaque.[54]

Safety assurance aims to establish confidence in a system's safety, typically communicated through a safety case—a structured, evidence-based argument prepared by developers.[55] However, static safety cases are poorly suited to the dynamic nature of clinical settings and machine learning. The adaptive behavior of AI systems can alter their operating context, undermining the assumptions on which the original safety case was built.[56] Therefore, AI developers and safety engineers must be included in assessments of moral accountability for patient harm. A shift from static to dynamic assurance is essential.[57] This requires consensus on acceptable levels of control reduction and epistemic uncertainty prior to implementation and within specific clinical contexts.

Factors to consider after implementation: long-term surveillance and lifecycle governance of deployed models

After implementation, there are always changes in patient demographics, variations in clinical techniques, and confounding medical interventions. Among these factors, effective interventions modify outcomes, thereby introducing bias into performance assessment.[6] Thus, post-deployment assessment of predictive AI models is necessary for developers. Specifically, it includes three core components: geographic heterogeneity and model recalibration, continuous monitoring and updating, and health economic evalua-

Table 1. Illustration of the DIAMOND checklist using the QRISK programme as an example.

Consideration	Example
Consideration 1	The QRISK family (QRISK1–3) and the QR4 update were developed to estimate 10-year risk of incident cardiovascular disease (CVD) in UK adults and to support primary-prevention decisions (e.g., statin initiation and lifestyle counselling) in routine primary care.
Consideration 2	The models were derived using routinely collected general-practice EHR data from the nationally representative QResearch database (EMIS), with linkage to outcome sources such as ONS mortality (and, in newer work, linked hospital data), enabling population-scale development and evaluation.
Consideration 3	Development prioritised predictors that are clinically meaningful and routinely captured in EHR workflows, balancing incremental predictive value against deployability at scale.[70]
Consideration 4	To demonstrate generalisability, external validation in independent UK datasets—initially THIN and later CPRD—has been central to the QRISK lifecycle, reducing reliance on "home" data and strengthening confidence in transportability across settings. Model performance was assessed using discrimination (C statistic/AUC) and calibration; in QR4, shrinkage diagnostics were close to 1, indicating minimal overfitting, and external validation showed modest yet consistent discrimination gains over QRISK3.[71–72]
Consideration 5	In routine UK primary care, NICE guidance recommends QRISK3 for adults aged 25–84 years without established CVD to support shared decision-making. It operationalises prevention by recommending atorvastatin 20 mg for primary prevention when 10-year QRISK3 ≥10%, while recognising that embedded clinical systems may lag in version updates.[72]
Consideration 6	To reinforce trust and uptake, transparent reporting and critical appraisal should align with established methodological standards—TRIPOD for complete reporting and PROBAST for risk-of-bias/applicability assessment—to support transparency, explainability, and reproducibility.[73]
Consideration 7	In practice, implementation commonly relies on EHR-integrated decision-support engines rather than public demonstrators; accordingly, the QRISK website explicitly distinguishes the non-clinical demonstrator from the EP QRISK3 Engine registered as a Class I medical device for use within clinical systems.
Consideration 8	Methodologically, QRISK3 and QR4 were developed using sex-specific Cox-based time-to-event models to estimate 10-year risk, with clear separation between derivation and validation populations (practice-level split and/or independent databases) to mitigate optimism and quantify generalisability.[2]
Consideration 9	Real-world experience shows that software-layer failures can be clinically material: an MHRA investigation reported code-mapping errors in one integrated QRISK2 calculator, requiring temporary disablement and remediation—underscoring governance, surveillance, and corrective action as integral to safe deployment.
Consideration 10	Consistent with this emphasis on robustness, ongoing external validation in independent UK datasets—again spanning THIN and CPRD—remains a core feature of the QRISK lifecycle, sustaining confidence in transportability as models evolve.[71]
Consideration 11	Because baseline CVD incidence, coding completeness, treatment patterns, and population risk-factor distributions evolve over time, QRISK is framed as an updating programme, with periodic recalibration/versioning used to preserve calibration and clinical relevance.
Consideration 12	Health-economic evidence is embedded in UK guideline processes: the BMJ summary of updated NICE guidance notes that high-intensity statins are cost-effective across plausible age–risk combinations, while NICE retained the 10% 10-year QRISK threshold to prioritise uptake among those most likely to benefit.[74]

THIN: the health improvement network; CPRD: clinical practice research datalink; AUC: area under the curve; NICE: national institute for health and care excellence; TRIPOD: transparent reporting of a multivariable prediction model for individual prognosis or diagnosis; PROBAST: prediction model risk of bias assessment tool; EP: endeavour predict; MHRA: Medicines and Healthcare Products Regulatory Agency; BMJ: British Medical Journal.

tion.[58]

Geographic heterogeneity drives risk miscalibration: evidence across cohorts and regions Cross regional heterogeneity in disease risk is often underestimated or even overlooked. Using UKB, CKB, and two Chinese regional EHR datasets (Yinzhou and Shenzhen), we propensity-score–matched individuals on key lung cancer risk factors and still observed marked between-setting differences in absolute risk (Figures 3A-D). This indicates that there are still non-negligible driving factors for lung cancer occurrence that have not been identified, such as genetic factors, molecular biomarkers, as well as gene-environment and molecular-environment interactions. Additionally, different exposure patterns across various regions that are still unexplored also play an important role. Consequently, a model initially developed within a specific set often exhibits miscalibration in a new environment, underscoring the need for model recalibration. Such miscalibration is likely due to genetic diversity, environmental differences, and variations in screening practices. Models that are well-developed and validated in high-quality cohorts may exhibit different performances when applied to real-world big data, owing to differences in data quality.[59] Thus, there is an urgent need to advance methodological innovation in disease risk prediction modelling, in addition to large-scale electronic healthcare data from geographically or socioeconomically comparable populations.[7] The generality of the model techniques including local recalibration/model updating when miscalibration is detected (e.g., updating intercept/slope or other updating strategies).

Continuous monitoring and updating mechanisms. AI model deployment is not a one-stop procedure but rather a continuous, iterative process. A salient example is the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial model (PLCOm2012), a widely used lung cancer risk prediction tool developed to support risk-based selection for low-dose CT screening. Evidence has shown that risk-model–based eligibility (e.g., PLCOm2012) can outperform categorical criteria in identifying individuals who may benefit from screening.[60] After deployment, the model's behavior interacts with clinicians' practice, impacting its performance and requiring further tuning or retraining. However, such adjustments carry inherent risks: they may further deteriorate performance or lead to unintended consequences. Therefore, it is critical to log all details of the entire deployment process, including when the model was deployed, how it interacted with clinicians, and how its performance changed over time. To address this challenge and unravel AI model failure mechanisms, Liu et al proposed a medical algorithmic audit framework that fosters structured feedback among the end-users, model developers, and ITS team.[61]

Furthermore, AI models require ongoing validation using real-world data. This is a proactive, label-agnostic monitoring pipeline incorporating transfer and continual learning detects and mitigates harmful data shifts, preserving

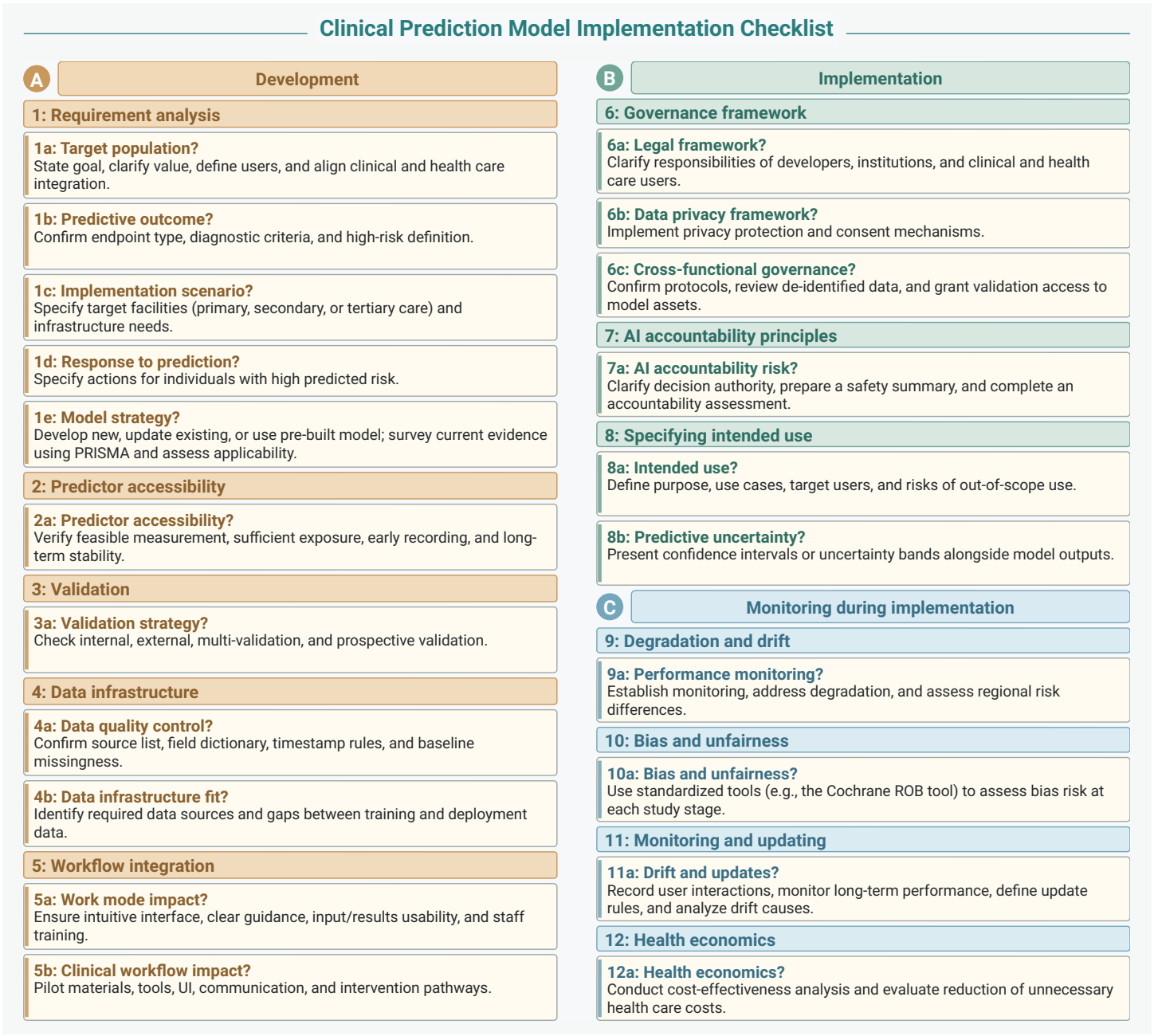


Figure 4. DIAMOND checklist: key considerations of prediction model development, implementation, and monitoring after implementation.

model performance over time.

The sustained effectiveness of a model depends on both technical adjustments and the influence of clinical practices and patient behavior. Healthcare professionals' and patients' perceptions of machine-learning-based risk prediction models can impact their use and effectiveness.[62] Considering the above, post-implementation impact should also be evaluated to enable regulatory agencies to determine the AI model's cost-effectiveness.

Health economic evaluation Although the evaluation of machine learning models typically focuses on measures of performance, they do not reflect the consequences of taking action based on the model's output.[63] If the projected outcomes are unfavorable (e.g., high costs per quality-adjusted life year), the rationale for conducting large-scale empirical impact studies may be substantially undermined. Accordingly, after model deployment, health economic evaluation (e.g., Markov models, decision trees or simulations assessing cost-effectiveness) is a core component.[64] For example, in lung cancer screening, predictive models such as OWL have shown superior screening efficiency compared to traditional guideline-based approaches in European populations.[65-67] Besides, the PLC_{m2012} model was shown to be cost-effective in the United States, while the NCC-LC_{m2021} model in China was well below the local willingness-to-pay threshold.[68–69] Developing health economic models' post-deployment is recommended, as these models can provide actionable feedback to the developing stage. This is critical to build a multi-stakeholder feedback loop for real-world monitoring and continuous model refinement.

DISCUSSION

The QRISK programme—most notably implemented as QRISK3 (introduced in 2017) and more recently extended through the QR4 algorithm—was developed to estimate an individual's absolute 10-year risk of incident cardiovascular disease (CVD) among UK adults, thereby supporting primary prevention decisions, including lifestyle optimisation and lipid-lowering therapy where appropriate. Derived from routinely collected, anonymised primary-care electronic health record (EHR) data from the QResearch database, which has now been reported to include health records from more than 40 million patients, the QRISK models provide an excellent example of the application of our DIAMOND checklist (Table 1).

CONCLUSION

In summary, closing the translational gap from algorithm development to clinical implementation necessitates careful attention throughout the entire model lifecycle. This spans from alignment with real-world clinical needs to sustained validation of efficacy and ethical compliance post-deployment. By prioritizing rigorous testing protocols, transparent reporting standards, iterative feedback loops, and robust governance framework, as detailed in the DIAMOND Checklist(as detailed in Figure 4), we can not only enhance the model's practical utility, but also ensure its enduring effectiveness in optimizing patient outcomes and healthcare system efficiency.

REFERENCES

1. Wei Q., Cui M., Liu Z., et al. (2025). Integrating statistical design and inference: A roadmap for robust and trustworthy medical AI. *Innov. Med.* **3**:100145. DOI:10.59717/j.xinn-med.2025.100145
2. Hippisley-Cox J., Coupland C., Vinogradova Y., et al. (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *BMJ* **335**:136. DOI:10.1136/bmj.39261.471806.55
3. Hobbs F.D. (2015). Prevention of cardiovascular diseases. *BMC Med.* **13**:261. DOI:10.1186/s12916-015-0507-0
4. Vasey B., Clifton D.A., Collins G.S., et al. (2021). DECIDE-AI: New reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat. Med.* **27**:186–187. DOI:10.1038/s41591-021-01229-5
5. Collins G.S., Moons K.G.M., Dhiman P., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**:e078378. DOI:10.1136/bmj-2023-078378
6. Ansari S., Baur B., Singh K., et al. (2025). Challenges in the postmarket surveillance of clinical prediction models. *NEJM AI* **2**:AIp2401116. DOI:10.1056/AIp2401116
7. Van Calster B., Collins G.S., Vickers A.J., et al. (2025). Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: Overview and guidance. *Lancet Digit. Health* **7**:100916. DOI:10.1016/j.landig.2025.100916
8. Lekadir K., Frangi A.F., Porras A.R., et al. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **388**:e081554. DOI:10.1136/bmj-2024-081554
9. Saelmans A., Seinen T., Pera V., et al. (2025). Implementation and updating of clinical prediction models: A systematic review. *Mayo Clin. Proc. Digit. Health* **3**:100228. DOI:10.1016/j.mcpdig.2025.100228
10. Efthimiou O., Seo M., Chalkou K., et al. (2024). Developing clinical prediction models: A step-by-step guide. *BMJ* **386**:e078276. DOI:10.1136/bmj-2023-078276
11. Liu X., Peng Y., Li N., et al. (2025). The necessity and feasibility assessment tool of the clinical prediction model for individual prognosis before its startup: A multi-sectoral Delphi consensus study. *J. Evid. Based Med.* e70106. DOI:10.1111/jebm.70106.
12. Zhou Y., Nie H., Gong X., et al. (2025). A three-tier AI solution for equitable glaucoma diagnosis across China's hierarchical healthcare system. *NPJ Digit. Med.* **8**:400. DOI:10.1038/s41746-025-01835-4
13. El-Sherbini A.H., Hassan Virk H.U., Wang Z., et al. (2023). Machine-learning-based prediction modelling in primary care: State-of-the-art review. *AI* **4**:437–460. DOI:10.3390/ai4020024
14. Abbasi A.B., Curtis L.H. and Califf R.M. (2025). The promise of real-world data for research: What are we missing. *N. Engl. J. Med.* **393**:318–321. DOI:10.1056/NEJMp2416479
15. Liu H. and Ding G. (2024). Scientific wellness in China: Innovations and implementation of data- and AI-driven health. *Innov. Med.* **2**:100103. DOI:10.59717/j.xinn-med.2024.100103
16. Field M., Thwaites D.I., Carolan M., et al. (2022). Infrastructure platform for privacy-preserving distributed machine learning development of computer-assisted theragnostics in cancer. *J. Biomed. Inform.* **134**:104181. DOI:10.1016/j.jbi.2022.104181
17. Chekroud A.M., Hawrilenko M., Loho H., et al. (2024). Illusory generalizability of clinical prediction models. *Science* **383**:164–167. DOI:10.1126/science.adg8538
18. Yang J., Dung N.T., Thach P.N., et al. (2024). Generalizability assessment of AI models across hospitals in a low-middle and high-income country. *Nat. Commun.* **15**:8270. DOI:10.1038/s41467-024-52618-6
19. de Hond A.A.H., Leeuwenberg A.M., Hooft L., et al. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: A scoping review. *NPJ Digit. Med.* **5**:2. DOI:10.1038/s41746-021-00549-7
20. Feng J., Phillips R.V., Malenica I., et al. (2022). Clinical artificial intelligence quality improvement: Towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit. Med.* **5**:66. DOI:10.1038/s41746-022-00611-y
21. Xu H., Feng G., Ma C., et al. (2023). AMHconverter: An online tool for converting results between the different anti-Müllerian hormone assays of Roche Elecsys®, Beckman Access, and Kangrun. *PeerJ* **11**:e15301. DOI:10.7717/peerj.15301
22. Pasqualetti S., Mussap M., Monteverde E., et al. (2024). C-reactive protein and brain natriuretic peptides harmonization. *Clin. Chim. Acta* **562**:119848. DOI:10.1016/j.cca.2024.119848
23. Collins G.S., Dhiman P., Ma J., et al. (2024). Evaluation of clinical prediction models (part 1): From development to external validation. *BMJ* **384**:e074819. DOI:10.1136/bmj-2023-074819
24. Riley R.D., Archer L., Snell K.I.E., et al. (2024). Evaluation of clinical prediction models (part 2): How to undertake an external validation study. *BMJ* **384**:e074820. DOI:10.1136/bmj-2023-074820
25. Riley R.D., Snell K.I.E., Archer L., et al. (2024). Evaluation of clinical prediction models (part 3): Calculating the sample size required for an external validation study. *BMJ* **384**:e074821. DOI:10.1136/bmj-2023-074821
26. Wong A., Otles E., Donnelly J.P., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.* **181**:1065–1070. DOI:10.1001/jamainternmed.2021.2626
27. Han R., Acosta J.N., Shakeri Z., et al. (2024). Randomised controlled trials evaluating artificial intelligence in clinical practice: A scoping review. *Lancet Digit. Health* **6**:e367–e373. DOI:10.1016/S2589-7500(24)00047-5
28. Rajkomar A., Dean J. and Kohane I. (2019). Machine learning in medicine. *N. Engl. J. Med.* **380**:1347–1358. DOI:10.1056/NEJMra1814259
29. Liu X., Faes L., Kale A.U., et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digit. Health* **1**:e271–e297. DOI:10.1016/S2589-7500(19)30123-2
30. Mayfield J.D. and Romero J. (2025). Establishing a chain of evidence for AI in radiology: Sham AI and randomized controlled trials. *Radiol. Artif. Intell.* **7**:e250334. DOI:10.1148/ryai.250334
31. Shi Z., Hu B., Lu M., et al. (2025). Development and validation of a sham-AI model for intracranial aneurysm detection at CT angiography. *Radiol. Artif. Intell.* **7**:e240140. DOI:10.1148/ryai.240140
32. Hubbard R.A., Gatsonis C.A., Hogan J.W., et al. (2024). Target trial emulation for observational studies: Potential and pitfalls. *N. Engl. J. Med.* **391**:1975–1977. DOI:10.1056/NEJMp2407586
33. Kandaswamy S., Muthu N., Braykov N., et al. (2025). Human performance evaluation of a pediatric artificial intelligence sepsis model. *J. Am. Med. Inform. Assoc.* **32**:1552–1561. DOI:10.1093/jamia/ocaf106
34. Saha S., Ross H., Velmovitsky P.E., et al. (2025). Machine learning-enhanced expert system for detecting heart failure decompensation using patient-reported vitals and electronic health records. *Sci. Rep.* **15**:30979. DOI:10.1038/s41598-025-16376-9
35. Nong P., Raj M. and Platt J. (2022). Integrating predictive models into care: Facilitating informed decision-making and communicating equity issues. *Am. J. Manag. Care* **28**:18–24. DOI:10.37765/ajmc.2022.88812
36. Amann J., Blasimme A., Vayena E., et al. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **20**:310. DOI:10.1186/s12911-020-01332-6
37. Muralidharan V., Adewale B.A., Huang C.J., et al. (2024). A scoping review of reporting gaps in FDA-approved AI medical devices. *NPJ Digit. Med.* **7**:273. DOI:10.1038/s41746-024-01270-x
38. Singh R., Paxton M. and Auclair J. (2025). Regulating the AI-enabled ecosystem for human therapeutics. *Commun. Med.* **5**:181. DOI:10.1038/s43856-025-00910-x
39. Williams M., Karim W., Gelman J., et al. (2024). Ethical data acquisition for LLMs and AI algorithms in healthcare. *NPJ Digit. Med.* **7**:377. DOI:10.1038/s41746-024-01399-9
40. Bruns A. and Winkler E.C. (2024). Dynamic consent: A royal road to research consent? *J. Med. Ethics* jme-2024-110153. DOI:10.1136/jme-2024-110153.
41. Mascalzoni D., Melotti R., Pattaro C., et al. (2022). Ten years of dynamic consent in the CHRIS study: Informed consent as a dynamic process. *Eur. J. Hum. Genet.* **30**:1391–1397. DOI:10.1038/s41431-022-01160-4
42. Kang D.Y., DeYoung P.N., Tantiongloc J., et al. (2021). Statistical uncertainty quantification to augment clinical decision support: A first implementation in sleep medicine. *NPJ Digit. Med.* **4**:142. DOI:10.1038/s41746-021-00515-3
43. Kompa B., Snoek J. and Beam A.L. (2021). Second opinion needed: Communicating uncertainty in medical machine learning. *NPJ Digit. Med.* **4**:4. DOI:10.1038/s41746-020-00367-3
44. Tsaneva-Atanasova K., Pederzanil G. and Laviola M. (2025). Decoding uncertainty for clinical decision-making. *Philos. Trans. A Math. Phys. Eng. Sci.* **383**:20240207. DOI:10.1098/rsta.2024.0207
45. Widner K., Virmani S., Krause J., et al. (2023). Lessons learned from translating AI from development to deployment in healthcare. *Nat. Med.* **29**:1304–1306. DOI:10.1038/s41591-023-02293-9
46. Yang J., Soltan A.A.S., Eyre D.W., et al. (2023). Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nat. Mach. Intell.* **5**:884–894. DOI:10.1038/s42256-023-00697-3
47. Ebad S.A., Alhashmi A., Amara M., et al. (2025). Artificial intelligence-based software as a medical device (AI-SaMD): A systematic review. *Healthcare (Basel)* **13**. DOI:10.3390/healthcare13070817.
48. Wachter R.M. (2013). Personal accountability in healthcare: Searching for the right balance. *BMJ Qual. Saf.* **22**:176–180. DOI:10.1136/bmjqs-2012-001227
49. Yapps B., Shin S., Bighamian R., et al. (2017). Hypotension in ICU patients receiving vasopressor therapy. *Sci. Rep.* **7**:8551. DOI:10.1038/s41598-017-08137-0
50. Lee B., Patel S., Favorito C., et al. (2025). Development and commercialization

pathways of AI medical devices in the United States: Implications for safety and regulatory oversight. *NEJM AI* **2**:AIra2500061. DOI:10.1056/AIra2500061
51. Nouis S.C., Uren V. and Jariwala S. (2025). Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making: A qualitative study of healthcare professionals' perspectives in the UK. *BMC Med. Ethics* **26**:89. DOI:10.1186/s12910-025-01243-z
52. Habli I., Lawton T. and Porter Z. (2020). Artificial intelligence in health care: Accountability and safety. *Bull. World Health Organ.* **98**:251–256. DOI:10.2471/blt.19.237487
53. Talbert M. (2016). *Moral responsibility: An introduction* (John Wiley & Sons).
54. LeCun Y., Bengio Y. and Hinton G. (2015). Deep learning. *Nature* **521**:436–444. DOI:10.1038/nature1459
55. Cleland G.M., Sujan M.A., Habli I., et al. (2012). Evidence: Using safety cases in industry and healthcare (The Health Foundation).
56. Denney E., Pai G. and Habli I. (2015). Dynamic safety cases for through-life safety assurance. *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*.
57. Hwang T.J., Kesselheim A.S. and Vokinger K.N. (2019). Lifecycle regulation of artificial intelligence- and machine learning-based software devices in medicine. *JAMA* **322**:2285–2286. DOI:10.1001/jama.2019.16842
58. Vasey B., Nagendran M., Campbell B., et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **28**:924–933. DOI:10.1038/s41591-022-01772-9
59. Zhang L., Wang X., Chen Q., et al. (2025). Lung cancer risk assessment by prediction model: A global perspective. *Thorax* **80**:890–899. DOI:10.1136/thorax-2023-221253
60. Tammemägi M.C., Katki H.A., Hocking W.G., et al. (2013). Selection criteria for lung-cancer screening. *N. Engl. J. Med.* **368**:728–736. DOI:10.1056/NEJMoa1211776
61. Liu X., Glocker B., McCradden M.M., et al. (2022). The medical algorithmic audit. *Lancet Digit. Health* **4**:e384–e397. DOI:10.1016/s2589-7500(22)00003-6
62. Giddings R., Joseph A., Callender T., et al. (2024). Factors influencing clinician and patient interaction with machine learning-based risk prediction models: A systematic review. *Lancet Digit. Health* **6**:e131–e144. DOI:10.1016/s2589-7500(23)00241-8
63. Kattan M.W. and Gerds T.A. (2020). A framework for the evaluation of statistical prediction models. *Chest* **158**:S29–S38. DOI:10.1016/j.chest.2020.03.005
64. Alba A.C., Agoritsas T., Walsh M., et al. (2017). Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *JAMA* **318**:1377–1384. DOI:10.1001/jama.2017.12126
65. Pan Z., Zhang R., Shen S., et al. (2023). OWL: An optimized and independently validated machine learning prediction model for lung cancer screening based on the UK Biobank, PLCO, and NLST populations. *eBioMedicine* **88**:104443. DOI:10.1016/j.ebiom.2023.104443
66. Ye Z., Sun Y., Yin Y., et al. (2025). Assessment and recalibration of seventeen lung cancer risk prediction models in approximately one million Chinese population utilising healthcare big data: A retrospective cohort analysis. *Lancet Reg. Health West. Pac.* **58**:101575. DOI:10.1016/j.lanwpc.2025.101575
67. Feng X., Goodley P., Alcala K., et al. (2024). Evaluation of risk prediction models to select lung cancer screening participants in Europe: A prospective cohort consortium analysis. *Lancet Digit. Health* **6**:e614–e624. DOI:10.1016/S2589-7500(24)00123-7
68. Zhang T., Wang Y., Chen X., et al. (2025). Cost-effectiveness of risk model-based lung cancer screening in smokers and nonsmokers in China. *BMC Med.* **23**:315. DOI:10.1186/s12916-025-04065-3
69. Toumazis I., Cao P., de Nijs K., et al. (2023). Risk model-based lung cancer screening: A cost-effectiveness analysis. *Ann. Intern. Med.* **176**:320–332. DOI:10.7326/m22-2216
70. Hippisley-Cox J., Coupland C. and Brindle P. (2017). Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ* **357**:j2099. DOI:10.1136/bmj.j2099
71. Hippisley-Cox J., Coupland C., Vinogradova Y., et al. (2008). Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: A validation study. *Heart* **94**:34–39. DOI:10.1136/hrt.2007.134890
72. Hippisley-Cox J., Coupland C.A., Bafadhel M., et al. (2024). Development and validation of a new algorithm for improved cardiovascular risk prediction. *Nat. Med.* **30**:1440–1447. DOI:10.1038/s41591-024-02905-y
73. Collins G.S., Reitsma J.B., Altman D.G., et al. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Br. J. Surg.* **102**:148–158. DOI:10.1002/bjs.9736
74. Samarasekera E.J., Clark C.E., Kaur S., et al. (2023). Cardiovascular disease risk assessment and reduction: Summary of updated NICE guidance. *BMJ* **381**:p1028. DOI:10.1136/bmj.p1028

FUNDING AND ACKNOWLEDGMENTS

This study was supported by Noncommunicable Chronic Diseases-National Science and Technology Major Project (No. 2024ZD0521703 to Y.W., 2025ZD0551802 to Q.Z.) and the National Natural Science Foundation of China (No. 82473728 to Y.W., 82220108002 to F.C.). This study was supported by High-performance Computing Platform of Peking University. The authors sincerely thank Ziwei Xi and Ziqing Ye for aiding with analysis and graphics. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Y.W., Y.H., and G.F. conceived the concept and proposed the original idea of this review. Q.W. and M.C. performed analysis and drafted the original manuscript. Y.Q., Q.Z., Z.L. Y.C., D.C.C., F.D.R.H. and Y.L. provided constructive comments and further optimized the manuscript. Y.W. developed the framework of the article and provided funding support. All authors critically reviewed the manuscript, and approved the final manuscript.

DECLARATION OF INTERESTS

Guoshang Feng is an Editorial Board member of The Innovation Medicine and was blinded from reviewing or making final decisions on the manuscript. Peer review was handled independently of this member and their research group. The other authors declare no conflicts of interest.

DATA AND CODE AVAILABILITY

The reviewed evidence was obtained from cited publications, guidelines, and regulatory documents. UK Biobank data are available upon approved application. Local electronic health record data require approval from the relevant data custodians and ethics committees. Analytical code is available from the corresponding author upon reasonable request.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.59717/j.xinn-med.2026.100229.