

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/189812/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Chi, Hao, Liu, Xin, Yan, Yalan, Wang, Lexin, Luo, Peng and Xu, Jiayu 2026. Trustworthy AI for disease pathogenesis and precision medicine: A disease-agnostic framework from data to deployment. *Genes & diseases* , 102480. 10.1016/j.gendis.2026.102480

Publishers page: <https://doi.org/10.1016/j.gendis.2026.102480>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Journal Pre-proof

Trustworthy AI for disease pathogenesis and precision medicine: A disease-agnostic framework from data to deployment

Hao Chi, Xin Liu, Yalan Yan, Lexin Wang, Peng Luo, Lanqian Su, Lai Jiang, Xuancheng Zhou, Gang Huang, Yuxuan Jiang, Shiyang Yin, Ke Su, Haiqing Chen, Ge Zhang, Pengpeng Zhang, Hao Zhang, Bicheng Ye, Qiang Cao, Jie Liu, Vedbar Khadka, Hua Yang, Guanhu Yang, Jiayu Xu, Peng Wang, Xuelian Wang, Zhijia Xia, Xiaosong Li, Youping Deng, Zhenjiang Li, Ke Xu

PII: S2352-3042(26)00452-6

DOI: <https://doi.org/10.1016/j.gendis.2026.102480>

Reference: GENDIS 102480

To appear in: *Genes & Diseases*

Received Date: 7 March 2026

Revised Date: 13 May 2026

Accepted Date: 1 June 2026

Please cite this article as: Chi H, Liu X, Yan Y, Wang L, Luo P, Su L, Jiang L, Zhou X, Huang G, Jiang Y, Yin S, Su K, Chen H, Zhang G, Zhang P, Zhang H, Ye B, Cao Q, Liu J, Khadka V, Yang H, Yang G, Xu J, Wang P, Wang X, Xia Z, Li X, Deng Y, Li Z, Xu K, Trustworthy AI for disease pathogenesis and precision medicine: A disease-agnostic framework from data to deployment, *Genes & Diseases*, <https://doi.org/10.1016/j.gendis.2026.102480>.

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: <https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article>. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



1 Trustworthy AI for disease pathogenesis and precision medicine: A

2 disease-agnostic framework from data to deployment

3
4 Hao Chi^{a,b,w}, Xin Liu^{c,d,w}, Yalan Yan^{e,w}, Lexin Wang^{a,f,w}, Peng Luo^{g,w}, Lanqian Su^e,
5 Lai Jiang^e, Xuancheng Zhou^e, Gang Huang^e, Yuxuan Jiang^e, Shiyang Yin^e, Ke Su^h,
6 Haiqing Chen^a, Ge Zhangⁱ, Pengpeng Zhang^j, Hao Zhang^k, Bicheng Ye^l, Qiang Cao^m,
7 Jie Liuⁿ, Vedbar Khadka^b, Hua Yang^b, Guanhu Yang^{o,p,q}, Jiayu Xu^r, Peng Wang^s,
8 Xuelian Wang^t, Zhijia Xia^u, Xiaosong Li^{a,f,*}, Youping Deng^{b,*}, Zhenjiang Li^{c,*}, Ke
9 Xu^{v,*}

10
11 ^a Western Institute of Digital-Intelligent Medicine, Chongqing 401329, China

12 ^b Department of Quantitative Health Sciences, John A. Burns School of Medicine,
13 University of Hawai‘i at Mānoa, Honolulu, HI 96813, USA.

14 ^c Department of Radiation Physics, Shandong Cancer Hospital and Institute,
15 Shandong First Medical University and Shandong Academy of Medical Sciences,
16 Jinan, Shandong 250117, China

17 ^d Department of Gynecological Radiotherapy, Harbin Medical University Cancer
18 Hospital, Harbin, Heilongjiang 150081, China

19 ^e Clinical Medical College, Southwest Medical University, Luzhou, Sichuan 646000,
20 China

21 ^f Clinical Molecular Medicine Testing Center, Key Laboratory of Clinical Laboratory
22 Diagnostics (Chinese Ministry of Education), The First Affiliated Hospital of

- 23 *Chongqing Medical University, Chongqing 400016, China*
- 24 ^g *Department of Oncology, Zhujiang Hospital, Southern Medical University,*
 25 *Guangzhou, Guangdong 510280, China*
- 26 ^h *Department of Oncology, The Affiliated Traditional Chinese Medicine Hospital of*
 27 *Southwest Medical University, Luzhou, Sichuan 646000, China*
- 28 ⁱ *Department of Cardiology, The First Affiliated Hospital of Zhengzhou University,*
 29 *Zhengzhou, Henan 450052, China*
- 30 ^j *Department of Lung Cancer, Tianjin Lung Cancer Center, National Clinical*
 31 *Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy,*
 32 *Tianjin's Clinical Research Center for Cancer, Tianjin Medical University Cancer*
 33 *Institute and Hospital, Tianjin 300060, China*
- 34 ^k *Department of Neurosurgery, The Second Affiliated Hospital of Chongqing Medical*
 35 *University, Chongqing 400016, China*
- 36 ^l *Liver Disease Center of Integrated Traditional Chinese and Western Medicine,*
 37 *Department of Radiology, Zhongda Hospital, Medical School of Southeast University,*
 38 *Nurturing Center of Jiangsu Province for State Laboratory of AI Imaging &*
 39 *Interventional Radiology (Southeast University), Nanjing, Jiangsu 210009, China*
- 40 ^m *Faculty of Applied Sciences, Macao Polytechnic University, R. de Luís Gonzaga*
 41 *Gomes, Macao 999078, China*
- 42 ⁿ *Department of General Surgery, Dazhou Central Hospital, Dazhou, Sichuan 635000,*
 43 *China*
- 44 ^o *Faculty of Chinese Medicine, State Key Laboratory of Quality Research in Chinese*

*Medicine, University Hospital, Macau University of Science and Technology, Macau
999078, China*

^p *Department of Specialty Medicine, Ohio University, Athens, OH 45701, USA*

^q *Research Department, Swiss University of Traditional Chinese Medicine, Bad
Zurzach 5330, Switzerland*

^r *School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG,
United Kingdom*

^s *Department of Oncology, Southwest Hospital, Third Military Medical University
(Army Medical University), Chongqing 400038, China*

^t *Department of Oncology and Hematology, Zhongxian People's Hospital, Chongqing
404300, China*

^u *Department of General, Visceral, and Transplant Surgery, Ludwig-Maximilians-
University Munich, Munich 80539, Germany*

^v *Department of Oncology, Chongqing Academy of Medical Sciences, Chongqing
General Hospital, Chongqing University, Chongqing 401147, China*

* Corresponding authors.

*E-mail addresses: lixiaosong@cqmu.edu.cn (X. Li), dengy@hawaii.edu (Y. Deng),
zjli@email.sdfmu.edu.cn (Z. Li), cqghxuke@cqu.edu.cn (K. Xu).*

^w These authors have contributed equally to this work and share first authorship.

Abstract

Artificial intelligence (AI) is rapidly reshaping biomedical research and clinical practice, yet many models that perform well on benchmarks fail to deliver mechanistic insight, robust generalization, or measurable clinical value. This review presents a disease-agnostic roadmap for trustworthy AI in disease pathogenesis and precision medicine across multi-omics, medical imaging, digital pathology, electronic health records, and longitudinal real-world data. We organize the evolving model landscape from classical machine learning to deep learning, self-supervised and foundation models, graph and knowledge-guided learning, multimodal fusion, and generative approaches, while emphasizing how methodological choices interact with data provenance, endpoint validity, label design, and clinically meaningful evaluation. To bridge prediction and biology, we synthesize interpretation strategies and propose an evidence ladder that distinguishes association from robustness, biological coherence, causal support, perturbation evidence, and clinical utility. We further distill practical principles for reliable evaluation and deployment, including leakage control, external and temporal validation, calibration, decision-curve-informed thresholding, subgroup reporting, uncertainty communication, and post-deployment monitoring. By linking technical advances to evidence standards and lifecycle governance, this review provides a reusable framework for converting AI-derived signals into testable mechanisms and clinically actionable precision medicine.

Keywords: Artificial intelligence; Causal inference; Generalization; Interpretability; Precision medicine

89

90 **Introduction**

91 Artificial intelligence (AI) is now used across much of biomedicine, from analyzing
92 molecular profiles to interpreting medical images and extracting signals from
93 electronic health records (1). These tools promise to support two closely related goals:
94 improving patient-level prediction (diagnosis, prognosis, treatment response) and
95 advancing biological understanding of why disease develops and progresses (2). In
96 practice, however, strong performance in a single dataset often does not translate into
97 robust clinical utility or credible mechanistic insight (3). This gap matters because
98 precision medicine is inherently comparative and context-dependent—models must
99 work across institutions, populations, and time, and the claims they motivate should
100 be testable and meaningful beyond the dataset in which they were trained (4).

101 A central reason is that biomedical data are heterogeneous and shaped by how they
102 are collected (5). Sequencing platforms, imaging scanners, pathology staining, clinical
103 documentation, and treatment pathways all imprint structure on the data (6). If a
104 model captures those patterns rather than disease biology, it can appear accurate while
105 failing when the setting changes (7). Even when discrimination is high, probabilities
106 may be poorly calibrated, thresholds for action may be unclear, and subgroup
107 performance can differ in ways that are clinically important (8). For work that aims to
108 inform pathogenesis, the bar is higher still: associations highlighted by a model need
109 to be stable, biologically coherent, and supported by evidence that is less sensitive to
110 confounding, measurement artifacts, and correlated proxies (9).

For these reasons, “trustworthy AI” should be understood as an end-to-end property rather than a single algorithmic feature (10). It depends on data provenance and endpoint definition, careful evaluation that avoids leakage, explicit testing under external and temporal validation, and an honest accounting of uncertainty and limitations (11). It also requires interpretability that matches the scientific or clinical question and, when claims move toward mechanism or treatment choice, methods that connect prediction to causal reasoning and to verification through independent data or perturbation evidence (12). Without this scaffolding, it is easy to produce models that are impressive on paper but difficult to reproduce, hard to validate, and risky to deploy (13).

This review takes a disease-agnostic approach because the same failure modes—leakage, confounding, shortcut learning, poor calibration, and drift—recur across conditions and data modalities. We organize the methodological landscape from classical machine learning to deep learning, self-supervised and foundation models, graph and knowledge-guided approaches, multimodal fusion, and generative methods but keep the emphasis on the evidence chain that makes results reliable and reusable. Our main contribution is an integrated framework that links data provenance, endpoint definition, interpretation, causal reasoning, validation, and deployment governance. Within this framework, the evidence ladder and the proposed model-and-data-card concept help distinguish association, mechanistic plausibility, and clinical utility. To provide a structured overview, we outline an end-to-end roadmap spanning data integration, modeling paradigms, causal interpretation,

validation, and deployment (Fig. 1).

Data foundations across diseases and modalities

Multi-omics, imaging/pathology, electronic health record (EHR)/text,

longitudinal and real-world data

A disease-agnostic AI framework starts with the recognition that biomedical “data” are not a single object but a collection of measurement systems, each with its own noise structure, missingness patterns, and biases. In multi-omics studies, inputs may include germline variation, somatic alterations, transcriptomes, epigenomes, proteomes, metabolomes, microbiomes, or derived pathway and cell-state scores (14). These layers differ in dynamic range, technical artifacts, and biological interpretability; for example, batch effects can dominate gene expression-based signals, while sparse, high-dimensional mutation features can drive instability if not handled carefully (15). A practical implication is that preprocessing choices—normalization, batch correction, filtering, feature aggregation, and data splitting—can change conclusions as much as model choice (16). For disease-agnostic reviews, it is often more informative to emphasize how signals are constructed (raw features versus biologically structured summaries) than to list specific assays (17). Medical imaging and digital pathology introduce additional complexities because models can exploit visual cues unrelated to disease biology (18). Scanner type, reconstruction kernels, resolution, staining protocols, slide preparation, compression artifacts, and even site-specific layout patterns can act as shortcuts (19). At the same time, imaging and pathology offer advantages: spatial context, morphological

phenotypes, and the possibility of weak supervision (*e.g.*, slide-level labels) that can scale (20). In pathology, multiple instance learning and tile-based pipelines are common, which makes choices around tiling, color normalization, tissue masking, and slide-level aggregation central to model validity (21). In radiology, harmonization and sensitivity to acquisition protocols should be treated as first-order concerns rather than afterthoughts (22).

EHR and clinical text data are attractive because they reflect real-world phenotypes and care pathways, but they are also highly confounded by the healthcare system (9). Codes, medications, labs, vitals, notes, and procedures are influenced by clinician behavior, institutional practice patterns, reimbursement rules, and the availability of tests (23). For many tasks, the strongest predictors can be proxies for clinical suspicion or intensity of care rather than underlying disease status (9). This does not make EHR-based prediction unusable, but it does change what a model's features mean and how generalization should be tested. Text adds further challenges: changes in documentation templates over time, differences in language across sites, and privacy constraints that limit sharing and replication.

Longitudinal and real-world data are increasingly important for precision medicine because many clinically meaningful questions are time-dependent: disease trajectories, treatment sequences, competing risks, and evolving definitions of response (24).

Longitudinal designs can reduce ambiguity about temporal ordering, but they also make leakage and bias easier to introduce. A time stamp attached to every variable is not sufficient if the underlying process generating the variable is itself influenced by

earlier clinical decisions. Consequently, time-aware splitting, explicit modeling of censoring, and careful definition of index times, prediction windows, and follow-up windows are essential when claims extend beyond static prediction. Different biomedical data modalities introduce distinct validity threats, ranging from batch effects and leakage in omics to site signatures in imaging and temporal biases in real-world data (Fig. 2).

Endpoint and label design (diagnosis, prognosis, response, and mechanistic phenotypes)

In biomedical AI, endpoints and labels are often the limiting factors. Many studies inherit labels from registries, pathology reports, billing codes, or clinician-assigned outcomes because they are available at scale. However, label availability does not guarantee label validity. Diagnosis labels can be noisy, heterogeneous across coding systems, or reflect uncertainty in borderline cases (25). Prognostic endpoints such as overall survival are objective but entangled with treatment access, comorbidities, and follow-up patterns (26). Treatment response labels can be especially fragile: “response” may depend on imaging schedules, measurement variability, line of therapy, and clinician decisions to stop or switch treatment (27). When the label reflects a care process, the model may learn the process rather than the biology (7).

A useful disease-agnostic distinction is between (i) labels intended for operational prediction (*e.g.*, short-term risk of an event within a defined window) and (ii) labels intended to support mechanistic interpretation (*e.g.*, pathway activity, cell-state transitions, or resistance phenotypes) (28). For mechanistic goals, it is often preferable

to define intermediate phenotypes that are closer to biology and less confounded by care pathways—while acknowledging that such phenotypes can be harder to measure and validate (29). For example, in multi-omics, mechanistic phenotypes might include inferred pathway activation, immune states, or metabolic programs; in pathology, they might include quantified morphological patterns that can be cross-checked by expert review (30).

Label design should also align with the intended use. A model meant to support clinical decisions must map outputs to an action, which requires a clear prediction target, a time horizon, and a decision threshold (13). In contrast, a model meant to nominate mechanisms or targets should be evaluated for robustness of the discovered signals across cohorts and for biological coherence, not solely for discrimination (31). Mixing these goals without stating which is primary leads to overstated claims. For disease-agnostic work, being explicit about whether the endpoint is a clinical decision target, a surrogate, or a mechanistic construct is one of the most effective ways to improve clarity and credibility.

Data quality, bias, confounding, missingness, and harmonization

The most common threats to validity recur across modalities. Confounding arises when variables associated with both the input features and the outcome are not properly accounted for; in biomedicine, confounders can include age, sex, ancestry, comorbidity, socioeconomic factors, tumor stage, and treatment intensity (32). In omics, batch and center effects can masquerade as biological differences (33). In imaging and pathology, acquisition and processing pipelines can embed site

signatures. In EHR, documentation and ordering practices create systematic differences across clinicians and institutions (34).

Missingness is not merely a technical inconvenience; it often carries meaning (35).

Tests may be absent because a clinician judged them unnecessary, because of resource limitations, or because a patient's condition made measurement difficult. Such informative missingness can be predictive, but it can also reduce transportability (36).

Strategies such as imputation, missingness indicators, and model architectures that handle missing modalities can help, but they should be paired with sensitivity analyses: performance should be checked in scenarios where missingness patterns shift, and mechanistic conclusions should not hinge on variables that are missing for reasons tied to care processes (37). In longitudinal settings, missing-data handling must also be time-aware. Global-mean imputation or bidirectional imputation across an entire time series can leak future information into earlier prediction states. Safer approaches include forward filling only from prior observations, anchoring missingness indicators to the prediction window, and fitting imputation models strictly within the training-set time window before applying them unchanged to validation or test data.

Harmonization can occur at multiple levels. At the feature level, one can standardize units, normalize intensities, adjust batch effects, or apply domain adaptation (5). At a higher level, one can move from raw features to biologically structured summaries—pathway scores, cell-type proportions, radiomic signatures designed to be stable, or representation learning via self-supervision. Disease-agnostic guidance

should stress that harmonization is not automatically beneficial: aggressive batch correction can remove real biology, while insufficient correction can inflate spurious associations. The safest approach is to treat harmonization as a hypothesis-driven step and to evaluate its impact using external or temporal validation, not just internal cross-validation.

Privacy-preserving learning and data governance

Large, diverse datasets are central to generalizable biomedical AI, yet privacy and governance constraints often limit data sharing. This tension has motivated privacy-preserving strategies such as federated learning, secure aggregation, and related approaches that allow model training across sites without centralizing raw data (38). These methods can broaden demographic and institutional diversity, but they do not remove the need for rigorous validation; federated training can still suffer from site-specific shortcuts and heterogeneity in data quality, labels, and workflows (39). Moreover, governance decisions—who can access data, how consent is managed, what can be shared, and how models are audited—shape which studies can be reproduced and how claims can be verified.

For a review aimed at trustworthiness, the practical point is that governance should be viewed as part of scientific rigor. Transparent reporting of data provenance, inclusion and exclusion criteria, de-identification procedures, and permissible reuse is not bureaucratic detail; it determines whether a model's behavior can be independently checked. When data cannot be shared, alternative mechanisms become more important: sharing code, documenting preprocessing and splitting, releasing trained

weights when appropriate, and providing detailed evaluation protocols that others can apply to independent cohorts. In precision medicine contexts, governance also intersects with accountability: models that influence decisions should have clear versioning, audit trails, and monitoring plans, particularly when clinical practice evolves.

Model taxonomy and emerging paradigms

Classical ML and survival modeling in biomedical prediction

In many biomedical studies, the most durable models are still classical ones. Regularized regression and tree-based ensembles often remain competitive when sample sizes are limited, when features are noisy, or when investigators intentionally compress biology into stable summaries such as pathway scores, cell-type fractions, or curated signatures (40). Their practical appeal is not only performance but also the ability to audit inputs, understand failure modes, and establish credible baselines. In precision medicine, time-to-event outcomes are common, so survival modeling is frequently the appropriate starting point (41). Here, the key question is rarely which algorithm is most fashionable; it is whether the modeling choices match the endpoint, censoring structure, prediction horizon, and intended use of the prediction (42). Survival prediction should not rely on a single static C-index alone. Time-dependent AUC, Brier score, and calibration curves should be reported at clinically meaningful horizons—for example, 1-year, 3-year, or disease-specific postoperative follow-up points—especially when predictions may guide surveillance intensity, adjuvant therapy, or patient stratification.

From a review perspective, the recurring lesson is that classical methods tend to succeed when the “data–endpoint–evaluation” triangle is well aligned. When reported gains over baselines are modest, they are often explained by careful splitting, leakage control, and external validation rather than model novelty (43). Conversely, many headline performance claims evaporate once evaluation is tightened or once transportability is tested (44). For mechanistic interpretation, classical models can be preferable when the feature space is biologically structured because stability and interpretability become more meaningful than raw discrimination.

Deep learning for images, pathology, and sequences

Deep learning has reshaped biomedical AI most visibly in imaging and pathology, where the raw signal is high-dimensional and inherently spatial. The dominant story here is not that deep networks “replace” classical models but that they enable representation learning when handcrafted features are inadequate. At the same time, imaging and pathology amplify the risk of shortcut learning: models can rely on acquisition signatures, staining patterns, compression artifacts, or site-specific workflows that correlate with labels in a given dataset but do not reflect disease biology (7).

In pathology, whole-slide analysis has encouraged tile-based pipelines and multiple instance learning because slide-level labels are common, while region-level annotations are expensive (45). The conceptual contribution of this literature is the shift toward weak supervision and scalable training, but the translational bottleneck often lies in reproducibility across labs: tiling strategy, color normalization, and slide

quality filters can materially change what the model learns (45). In radiology, similar issues arise with scanner heterogeneity and protocol drift. For sequence-based tasks, deep learning provides powerful embeddings and pattern discovery, but the same caution applies: if labels encode measurement or ascertainment processes, deep models may learn those processes rather than the biology of interest (46).

Self-supervised learning and foundation models

Self-supervised learning addresses a structural constraint in biomedicine: labeled data are scarce, while unlabeled archives are large. The practical promise is improved feature representations that transfer to many downstream tasks with fewer annotations (46). Foundation models extend this idea by scaling pretraining, often with the goal of producing modality-specific encoders that can be fine-tuned across diseases and tasks; large-scale pretrained models have been reported for ophthalmic imaging and for pathology-scale WSIs (47, 48). In a disease-agnostic review, these developments matter less as a list of architectures and more as a change in workflow: representation learning becomes a shared substrate, and task-specific modeling becomes comparatively lighter.

However, the translational question shifts rather than disappears. Pretraining can improve average performance and reduce the need for large labeled cohorts, but it can also propagate biases from the pretraining corpus, obscure interpretability, and encourage overconfidence in generalization (49). A common pattern is strong within-domain transfer but weaker cross-institutional robustness than expected. For trustworthy use, foundation models should therefore be evaluated as one would

evaluate any candidate clinical model—under external and temporal validation, with calibration checks and subgroup reporting—rather than treated as inherently generalizable.

Graph learning and knowledge-guided AI

Graph-based learning and knowledge-guided modeling represent a recurrent ambition in biomedical AI: to encode biological structure so that models are less likely to chase spurious associations and more likely to produce interpretable, testable hypotheses.

Networks—gene regulation, protein interactions, pathways, and cell–cell communication—provide an appealing scaffold, and graph neural networks offer a principled mechanism for exploiting it. Even when graphs do not improve headline metrics, they can improve scientific usability by organizing features into coherent biological modules and by supporting interpretation at the pathway or network level rather than at the level of individual variables (50).

The limitation is that biological knowledge is incomplete, biased toward well-studied components, and context-dependent. Highly studied genes such as *TP53*, *EGFR*, or *MYC* often occupy densely connected positions in protein-interaction and pathway databases, whereas rare-disease genes, non-hotspot targets, and context-specific regulators may be represented by sparse or uncertain edges. Graph construction choices can therefore dominate results, and claims can become circular if known biology is used both to build the model and to validate its outputs. More persuasive applications use knowledge as an inductive bias while still requiring independent validation: degree normalization, edge-confidence weighting, sensitivity analyses

with alternative network sources, masking or down-weighting hub nodes during robustness testing, and checking whether conclusions remain stable when high-degree canonical nodes are removed.

Multimodal fusion and missing-modality robustness

Many clinically meaningful questions cannot be answered well by a single modality. Molecular assays capture pathway activity, pathology captures tissue architecture, imaging captures spatial phenotypes at the organ scale, and EHR captures comorbidity and treatment trajectories. Multimodal learning aims to integrate these perspectives, either to boost prediction or to triangulate biologically meaningful states. The key theme across this literature is that fusion is not purely a technical choice; it is a statement about what information is complementary, how modality-specific noise is handled, and how the model should behave when inputs are missing or discordant. Fusion strategies differ in their interpretive implications. Early fusion is simple and efficient but vulnerable to scale imbalance, missing modalities, and dominance by high-dimensional inputs. Late fusion is practical when some modalities are unavailable, although it may miss nonlinear cross-modal interactions. Intermediate fusion, including attention-based or contrastive architectures, can learn shared representations and cross-modal dependencies and may therefore be better suited for exploring complex pathogenic mechanisms (51, 52). However, attention weights or shared embeddings are not mechanistic evidence by themselves; they require cohort-level stability, biological coherence, and orthogonal validation.

Generative AI: Opportunities and pitfalls for biomedicine

Generative approaches are being adopted for augmentation, imputation, denoising, representation learning, and clinical-text workflows. For a review centered on trustworthiness, their value is best framed in terms of where they can legitimately support the pipeline rather than where they can replace evidence. Synthetic augmentation can help when classes are rare, imputation can reduce the missing-data burden, and generative pretraining can produce useful embeddings. However, each of these uses carries risks: synthetic data can amplify bias, imputation can create illusory certainty, and generated outputs can be mistaken for validated findings (53, 54). Large language models add further concerns in biomedical and clinical-text settings, including hallucination, lack of traceability, prompt sensitivity, version drift, and summaries that omit uncertainty or overstate conclusions. Reviews of synthetic data generation in healthcare also stress the need to evaluate utility, bias, and privacy rather than assuming that synthetic data are automatically safe or valid (55).

The most consistent message emerging from this space is that generative AI should be treated as hypothesis- and workflow-supporting technology, not as a substitute for validation. When generative models affect scientific or clinical conclusions, they must be evaluated under the same standards as predictive models: external validity, calibration where relevant, sensitivity to shift, and transparency about how generated content was used. In precision medicine, where decisions have consequences, the burden is to show that generative components improve reliability rather than merely improving internal metrics (Fig. 3).

From prediction to mechanism: Interpretability and causal reasoning

Interpretability levels and how to judge explanations

Interpretability is not a cosmetic add-on in biomedical AI. It plays different roles depending on the aim. For operational prediction, explanations mainly support trust, debugging, and safe use. For mechanistic work, explanations are part of the evidence because they guide biological claims and downstream validation. These two settings place different demands on what counts as a “good” explanation, and treating them as the same is a common source of overstatement (56).

A practical way to evaluate explanations is to focus on behavior rather than on the specific method used. Reliable explanations tend to be stable under reasonable changes in data splits and preprocessing, and they remain recognizable when tested in independent cohorts (57). They also tend to be specific in the sense that highlighted signals map to plausible biology rather than to known technical sources of variation such as batch, site, or processing artifacts (7). Finally, they are useful when they produce hypotheses that can be checked with independent data or perturbation evidence rather than remaining purely descriptive (58). In many studies, these checks are more informative than choosing one attribution algorithm over another (59).

Global and local explanations should also be kept conceptually separate. Global explanations summarize model reliance at the population level, while local explanations attempt to justify a single prediction. Both can be useful, but they answer different questions and can disagree without either being wrong (60). In disease-agnostic settings, global explanations often align better with mechanism-oriented narratives because they can be compared across cohorts and

linked to biological programs. Local explanations are often more helpful for error analysis, identifying atypical cases, and understanding when a model is extrapolating beyond the data (61). Discordance between global and local explanations should be treated as a diagnostic signal rather than an immediate failure. When local attributions such as SHAP values conflict with global feature-importance rankings, researchers can stratify cases into clinically or molecularly similar subgroups, evaluate feature-interaction effects, and test whether distinct pathological drivers operate in different patient subsets.

Mechanism-aware interpretation across omics, pathology, imaging, and clinical data

How a model's signal connects to a mechanism depends strongly on the modality. In multi-omics settings, gene-level attributions are frequently unstable because features are correlated and redundant. For mechanistic interpretation, it is often more credible to elevate the analysis to higher biological organization, such as pathways, gene programs, cell states, regulatory modules, or interaction networks (62). These summaries reduce sensitivity to idiosyncratic features and make it easier to test whether the signal is consistent across cohorts and across layers of measurement (63). When several omic layers converge on the same process, the argument becomes more resilient, and the experimental path to validation becomes clearer.

In pathology and imaging, interpretation is commonly presented as spatial evidence, such as attention or attribution maps. These can be informative, but they can also reflect confounding, including scanner or staining signatures and slide preparation

artifacts (64). Mechanism-aware interpretation is stronger when spatial signals are translated into interpretable phenotypes, such as immune infiltration patterns, stromal architecture, tumor-stroma boundaries, necrosis, vascular features, or other tissue-level structures that can be reviewed by experts and validated by orthogonal assays. A particularly strong validation strategy is to physically co-register AI-generated histology heatmaps with spatial transcriptomic or proteomic clusters from adjacent tissue sections. In tumor microenvironment analyses, this type of *in situ* multi-omics anchoring can help determine whether attention-highlighted regions correspond to immune-excluded niches, stromal remodeling, hypoxic regions, or invasive tumor fronts, thereby moving interpretation beyond visual heatmaps toward molecularly grounded evidence of pathogenic mechanisms (65).

For EHR and clinical text, the main interpretability challenge is that strong predictors may represent the care process rather than disease biology (23). Orders, billing codes, and documentation features often capture clinician suspicion, institutional pathways, and resource availability. This does not prevent useful prediction, but it changes how explanations should be read (66). Mechanism claims should be restrained, and interpretation should focus on temporality, portability, and clinical coherence. When EHR-derived phenotypes are used as discovery tools, they are best treated as clinical strata or trajectories that require biological anchoring through targeted molecular or tissue studies.

Across modalities, mechanism-oriented narratives are strongest when model-derived signals are stable, biologically coherent, and supported by orthogonal evidence rather

than by a single explanation method.

Causal inference as a complement to predictive modeling

Prediction answers what is likely to occur given observed data. Causal inference addresses what would occur under intervention. Precision medicine increasingly depends on the second question because treatment selection is central to care. This is why causal framing is valuable even in studies that are primarily predictive. It makes assumptions explicit, clarifies which variables are confounders versus mediators, and helps separate prognostic prediction from treatment-effect estimation (67).

Causal tools are most relevant in two recurring situations. The first is individualized treatment decisions, where models aim to estimate heterogeneous treatment effects or to support choices among therapies (68). The second is mechanistic interpretation, where authors seek to argue that an AI-highlighted signal is not only correlated with outcome but plausibly upstream of disease progression or resistance. In both cases, causal reasoning provides a structured way to connect observational patterns to intervention-relevant claims. Study designs such as target trial emulation can improve clarity for treatment questions, but only when eligibility assessment, treatment assignment, covariate measurement, and follow-up origin are aligned to the same clinically meaningful time zero (69). If an intervention group must survive or remain event-free until treatment initiation while the comparator group is followed from an earlier point, immortal time bias can spuriously inflate apparent survival or remission benefits. Studies evaluating chronic disease interventions or therapeutic strategies should therefore define the index date, treatment grace period, follow-up window, and

censoring rules before estimating treatment effects.

Causal inference should not be used as a rhetorical upgrade. Identification assumptions remain central, and unmeasured confounding is common in biomedical data. The most responsible use is to treat causal analyses as evidence that helps prioritize hypotheses and guide experiments or prospective evaluation rather than as definitive proof (58).

An evidence ladder linking association to mechanism and clinical utility

Biomedical AI produces a spectrum of outputs, from correlational signals to decision-support tools. Organizing these outputs by evidence strength helps align conclusions with what the data justify. At the base is association, where a model identifies features correlated with an endpoint within a dataset. This level can be valuable for hypothesis generation, but it is also vulnerable to shortcut learning and leakage. The next level is robustness, where signals persist across resampling, reasonable pipeline variations, and independent cohorts (43). Robustness reduces the likelihood that a finding is a dataset artifact.

Mechanistic relevance requires additional layers, and each layer has boundaries.

Biological coherence means that a signal aligns with plausible pathways, cell states, or tissue phenotypes, ideally with concordance across modalities or molecular layers, but it can still reflect bias in prior knowledge. Causal support reflects evidence consistent with an intervention-based interpretation, whether through principled observational designs, genetic instruments or other strategies that address confounding, but it depends on explicit assumptions and sensitivity analyses.

Perturbation evidence strengthens a mechanistic claim by showing that manipulating a candidate driver changes a relevant phenotype in experimental systems, yet experimental context may not translate directly to patients. At the top sits clinical utility, where outputs translate into calibrated, actionable decisions and demonstrate benefit under realistic validation, ideally including prospective assessment and decision-oriented evaluation (70). To organize evidence strength, we propose an evidence ladder that situates AI findings along a hierarchy from association to clinical utility (Fig. 4).

This ladder is not a demand that all work reach the top. It is a language for honest positioning: studies remain valuable at lower rungs when claims are framed accordingly, while stronger mechanistic or clinical claims require progressively stronger evidence.

Validation, generalization, and reliability

Evaluation is the scientific core, not an afterthought

In biomedical AI, the credibility of a claim is determined less by the sophistication of the model than by the realism of the evaluation. This is now a broad consensus across modalities: reported improvements are often fragile when they rely on permissive data splitting, extensive reuse of the same cohort for model selection, or benchmarking that does not reflect how the system would be used. The field has also learned that “internal validation” is frequently a statement about how well a model fits an institutional dataset, not a guarantee of performance elsewhere (71). As a result, strong studies increasingly treat evaluation as part of study design rather than a final

reporting step, and they position performance numbers as conditional on clearly stated data, endpoints, and deployment assumptions.

A second consensus is that many high-profile failures trace back to subtle information pathways rather than obvious coding errors. These include contamination during preprocessing, non-independent samples appearing across splits, and outcome-adjacent proxies that are available only because of downstream care processes(43). The consequence is not merely inflated metrics; it is misinterpretation. When a model's success depends on shortcuts, explanations become misleading, and downstream biological narratives become difficult to reproduce. For a disease-agnostic review, the key message is that preventing these pathways is foundational to both translational prediction and mechanistic inference.

External and temporal validation as the practical test of transportability

Generalization is the central barrier to clinical translation, and external validation remains the most direct test of whether a model has learned disease-relevant signals rather than institution-specific signatures (72, 73). In practice, external validation should be interpreted as a family of tests rather than a single checkbox.

Cross-institution validation probes differences in patient mix and workflows; cross-platform validation probes differences in measurement technology; temporal validation probes drift as guidelines, assays, and treatments change. Across diseases, the pattern is consistent: models that remain stable across these settings are rare, but when they do, their conclusions become substantially more credible (74).

The literature also suggests that transportability should be framed as a spectrum.

Some shifts are superficial and can be mitigated with careful calibration or harmonization; others reflect deeper changes in clinical practice or disease definition that fundamentally alter the relationship between inputs and outcomes (75). This is why it is increasingly viewed as insufficient to report a single pooled metric. Studies that aim for clinical relevance often include site- and time-stratified performance, discuss where degradation occurs, and make the boundaries of applicability explicit.

Calibration and clinical usefulness: Performance is not the endpoint

Another cross-cutting lesson is that discrimination is not clinically useful (76, 77). Many biomedical AI papers continue to emphasize ranking metrics because they are familiar and easy to report, yet decision-making requires well-calibrated probabilities, clearly defined time horizons, and thresholds that map to action. Miscalibration can persist even when the AUC is high, and it is frequently exacerbated by dataset shift. When models are proposed for decision support, calibration and threshold behavior become as important as discrimination because they determine whether predictions translate into appropriate interventions.

A parallel development is the increasing use of decision-oriented evaluation. Threshold selection should not be based solely on statistically convenient cutoffs such as the Youden index. Decision curve analysis and disease-specific net-benefit functions can help quantify the asymmetric costs of false positives and false negatives (70, 76). A threshold suitable for triggering low-risk surveillance may be inappropriate for selecting an invasive procedure or withholding therapy. Clinically meaningful thresholding therefore requires explicit consideration of decision context,

expected benefit, harm, and the patient population in which the model will be used.

Reliability under uncertainty and shift

Biomedical AI operates in settings where uncertainty is inevitable: measurements are noisy, phenotypes are heterogeneous, and real-world inputs differ from curated development datasets (78). The field's experience has shown that models can become dangerously overconfident under shift, producing plausible outputs with little warning that the input lies outside what the model has effectively learned (79). Reliability therefore increasingly refers to a broader set of behaviors beyond average accuracy, including how confidence aligns with correctness and how models behave when inputs are atypical or incomplete (80).

Reliability therefore needs to be demonstrated rather than assumed. Practical checks include sensitivity to common perturbations, robustness to missingness patterns, uncertainty behavior under shift, and explicit failure-mode analysis. In translational contexts, systems should support conservative behavior when uncertainty is high because safe deployment often depends on knowing when not to trust a prediction.

Subgroup performance and fairness as part of scientific validity

Across diseases and modalities, subgroup performance has emerged as both an ethical and a scientific issue (30). Models trained on imbalanced data can perform well on average while failing for underrepresented groups, and these failures can remain invisible if reporting is limited to pooled metrics (81). Importantly, subgroup disparities are not only fairness concerns; they are also signals that the model has learned patterns that do not generalize evenly across populations, which weakens

mechanistic interpretation and undermines clinical transfer (46).

Accordingly, contemporary best practice is moving toward reporting performance across clinically meaningful strata and describing how subgroup definitions relate to the intended use of the model. In a mechanistic setting, subgroup inconsistency can indicate confounding, measurement artifacts, or effect modification that should shape interpretation and guide follow-up studies. In a clinical setting, subgroup reporting clarifies where a model may require additional data, recalibration, or restrictions on use.

Translation, implementation, and governance

Clinical workflows and human–AI collaboration

For biomedical AI to matter beyond publications, it must connect to decisions (82, 83). This connection is often the missing link in otherwise technically strong studies. In practice, “clinical deployment” rarely means replacing clinicians; it means inserting an AI system into an existing workflow so that it changes timing, prioritization, or selection in a way that is both feasible and defensible (84). The most transferable work therefore begins by specifying the use case with discipline: who uses the model, at what point in the care pathway, what information is available at that moment, and what action the output is intended to support (85). Without this framing, performance and interpretability become hard to interpret because it is unclear what the model is being judged against (86).

Human – AI collaboration introduces its own constraints. Predictions must be presented in a way that supports appropriate reliance rather than automation bias.

Interfaces should communicate uncertainty, confidence, and known limitations clearly rather than showing only a single deterministic output. Explanations should be tailored to the decision context: clinicians often need concise, clinically legible reasons, uncertainty ranges, and warning signals for situations in which the model may be unreliable, whereas researchers pursuing mechanisms need biological coherence and triangulation. For LLM-generated summaries, users should be able to trace outputs back to auditable source documents, and domain experts should verify clinically consequential statements before they influence care. Models that require extensive manual preprocessing, special hardware, or inputs that are rarely available will struggle to move beyond research settings, regardless of headline metrics. As a result, clinical translation favors systems that are robust to missingness, resilient under operational variability, and compatible with the cadence and incentives of routine care (87, 88). Clinical translation should be viewed as a lifecycle process that spans standardized reporting, deployment integration, continuous monitoring, and accountable updating (Fig. 5).

Monitoring, drift, updating, and auditability

Even when a model generalizes at launch, biomedical environments do not remain static (89). Diagnostic criteria change, assays and scanners are upgraded, treatment regimens evolve, and patient populations shift (90). These changes can alter both input distributions and outcome relationships, which means that model performance and calibration can degrade silently (91). For this reason, translation is increasingly understood as a lifecycle problem rather than a one-time validation problem. A

clinically credible system requires a plan for monitoring, for detecting drift, and for deciding when and how to update the model.

Governance mechanisms become essential at this stage (92). Version control and audit trails support traceability, allowing users to know which model generated which output and under what data conditions (93). Updating strategies must be accountable because improving performance in one setting can worsen it in another. The shared principle is simple: deployment without monitoring is experimentation without feedback.

Accountability, regulation, and ethical constraints

Clinical AI sits within frameworks of responsibility. When a model influences triage, diagnosis, or treatment selection, questions of accountability cannot be deferred to technical appendices (56). Who is responsible for the model's recommendations, who is authorized to override them, and how adverse events are investigated all shape whether a system is acceptable in practice. This is one reason why transparency and documentation are repeatedly emphasized in the translational AI literature. They support not only scientific reproducibility but also clinical safety and legal defensibility.

Ethical considerations extend beyond privacy. Models can amplify inequities when trained on data that underrepresent certain populations or reflect unequal access to care. They can also encode historical practice patterns that are not optimal. Moreover, the introduction of AI can change behavior, including clinician ordering patterns and patient pathways, which in turn changes the data generating process. This feedback

loop is often overlooked in retrospective studies but becomes important when interpreting model performance in real-world settings. A trustworthy translational posture therefore includes explicit limits of applicability, careful communication about uncertainty, and ongoing evaluation in the populations where the model will be used (10, 94).

Reporting practices as a bridge from innovation to reuse

A practical obstacle to progress is that many AI studies cannot be faithfully reproduced or compared because critical details are missing (95–97). Essential reporting elements include data provenance and inclusion criteria, endpoint definitions with time anchoring, leakage-safe splitting, preprocessing performed within evaluation folds, and external or temporal validation when available (98). For decision-support models, calibration, threshold behavior, and subgroup performance should be treated as standard reporting rather than optional additions (94).

We propose presenting these practices in a compact model-and-data-card format that functions both as a reporting template and as a reviewer checklist (92). By documenting the target population, prediction horizon, validation regime, uncertainty, calibration, threshold behavior, and subgroup performance, the card also creates a bridge to deployment: what is not documented during development cannot be reliably monitored, audited, or updated after implementation.

Discussion

The rapid expansion of biomedical AI has shifted the central challenge from building models to substantiating claims. A model may be useful for operational prediction,

mechanism generation, treatment-effect reasoning, or workflow support, but each use requires a different evidentiary standard (82, 99–101). The disease-agnostic framework proposed here is therefore intended to make those standards explicit: data and labels define what can be learned, modeling choices shape what is discoverable, interpretation and causal reasoning constrain what can be claimed, and validation and governance determine whether a result can travel beyond its development setting. Self-supervised and foundation-model approaches will likely accelerate this shift by making reusable representations available across imaging, pathology, text, and omics (102–104). Their value, however, should be judged less by benchmark breadth than by downstream behavior under external, temporal, and subgroup evaluation (13, 105, 106). In other words, foundation models change the scale of representation learning, but they do not change the evidence required for trustworthy use.

A second direction is tighter integration of predictive modeling with causal reasoning (107). This is especially important for precision medicine questions involving treatment choice, sequencing, and heterogeneous response (108, 109). Causal framing helps distinguish prognostic associations from intervention-relevant effects, clarifies assumptions, and identifies which AI-derived signals deserve experimental or prospective follow-up (110, 111).

Multimodal learning offers a complementary path by testing whether signals converge across orthogonal measurement systems (112, 113). Its practical limitation is modality availability: outside curated cohorts, missingness often reflects care pathways and access rather than random absence (36, 114, 115). Strong multimodal studies should

therefore report performance under both complete- and partial-modality scenarios and reserve mechanistic language for signals that remain stable across those conditions. Finally, clinical translation should be treated as a maintenance problem as much as an implementation problem (116, 117). Real-world use can change patient pathways, clinician behavior, and the data-generating process itself (118, 119). Monitoring, drift detection, accountable updating, and transparent documentation are therefore not administrative additions; they are the conditions under which AI evidence can remain valid after deployment (120–124). Taken together, the next phase of biomedical AI will be judged less by incremental gains on static benchmarks than by transportability, evidence integration, and maintainability. A disease-agnostic roadmap that links modeling choices to validation, causal reasoning, and lifecycle governance can help align technical innovation with the demands of disease pathogenesis research and precision medicine.

Conflict of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the Medical Research Foundation of Chongqing General Hospital (China) (No. Y2023YXYJMSXM04), the Science and Technology Research Program of the Chongqing Municipal Education Commission (China) (No.

KJQN202400116), the National Natural Science Foundation of China (No. 12475343), the Collaborative Academic Innovation Project of Shandong Cancer Hospital (China) (No. TS004), the Taishan Scholars Program of Shandong Province, China (No. ts202408375), and the Natural Science Foundation of Shandong Province, China (No. ZR2025QB37).

Acknowledgements

The authors thank all collaborators and supporting institutions for their contributions to this work. The authors also thank BioRender.com for providing graphical resources that supported the preparation of the figures.

Data availability

No new data were generated or analyzed in this study. Data sharing is not applicable to this article.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used generative AI and AI-assisted technologies solely for language polishing and readability improvement. No AI tools were used to generate scientific content, interpret data, draw scientific conclusions, or make research decisions. All content was critically reviewed and revised by the authors, and the authors take full responsibility for the accuracy, integrity, and originality of the manuscript.

749

750 **References**

- 751 1. Zhou SK, Greenspan H, Davatzikos C, et al. A review of deep learning in medical
752 imaging: Imaging traits, technology trends, case studies with progress highlights, and
753 future promises. *Proc IEEE Inst Electr Electron Eng*. 2021;109(5):820-838.
- 754 2. Wang X, Xiong Z, Zou K, et al. Reasoning-driven large language models in
755 medicine: opportunities, challenges, and the road ahead. *Lancet Digit Health*.
756 2026;8(1):100931.
- 757 3. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on
758 prediction models developed using supervised machine learning techniques:
759 systematic review. *BMJ*. 2021:n2281.
- 760 4. Matta S, Lamard M, Zhang P, et al. A systematic review of generalization research
761 in medical image classification. *Comput Biol Med*. 2024;183:109256.
- 762 5. Hu F, Chen AA, Horng H, et al. Image harmonization: a review of statistical and
763 deep learning methods for removing batch effects and evaluation metrics for effective
764 harmonization. *Neuroimage*. 2023;274:120125.
- 765 6. Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital
766 histology signatures on deep learning model accuracy and bias. *Nat Commun*.
767 2021;12(1):4423.
- 768 7. Ong Ly C, Unnikrishnan B, Tadic T, et al. Shortcut learning in medical AI hinders
769 generalization: method for estimating AI model generalization without external data.
770 *NPJ Digit Med*. 2024;7(1):124.

8. Ojeda FM, Jansen ML, Thiéry A, et al. Calibrating machine learning approaches for probability estimation: a comprehensive comparison. *Stat Med*. 2023;42(29):5451-5478.
9. Davis SE, Matheny ME, Balu S, Sendak MP. A framework for understanding label leakage in machine learning for health care. *J Am Med Inform Assoc*. 2023;31(1):274-280.
10. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
11. Hamdan S, Love BC, von Polier GG, et al. Confound-leakage: confound removal in machine learning leads to leakage. *GigaScience*. 2022;12:giad071.
12. Chicco D, Cumbo F, Angione C. Ten quick tips for avoiding pitfalls in multi-omics data integration analyses. *PLoS Comput Biol*. 2023;19(7):e1011224.
13. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505.
14. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J*. 2021;19:3735-3746.
15. Da-ano R, Masson I, Lucia F, et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Sci Rep*. 2020;10(1):10248.

- 793 16. Adam E, Zanoaga MD, Rota R, Cominetti O. A Comprehensive Protocol and
794 Step-by-Step Guide for Multi-Omics Integration in Biological Research. *J Vis Exp*.
795 2025;(222):10.3791/66995.
- 796 17. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using
797 deep learning. *Brief Bioinform*. 2022;23(1):bbab454.
- 798 18. Vásquez-Venegas C, Wu C, Sundar S, et al. Detecting and mitigating the clever
799 hans effect in medical imaging: a scoping review. *J Digit Imaging Inform Med*.
800 2025;38(4):2563-2579.
- 801 19. Dehkharghanian T, Bidgoli AA, Riasatian A, et al. Biased data, biased AI: deep
802 networks predict the acquisition site of TCGA images. *Diagn Pathol*. 2023;18(1):67.
- 803 20. Baxi V, Edwards R, Montalto M, Saha S. Digital pathology and artificial
804 intelligence in translational medicine and clinical practice. *Mod Pathol*.
805 2022;35(1):23-32.
- 806 21. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F.
807 Data-efficient and weakly supervised computational pathology on whole-slide images.
808 *Nat Biomed Eng*. 2021;5(6):555-570.
- 809 22. Soliman MAS, Kelahan LC, Magnosta M, et al. A framework for harmonization of
810 radiomics data for multicenter studies and clinical trials. *JCO Clin Cancer Inform*.
811 2022(6):e2200023.
- 812 23. Ramadan B, Liu MC, Burkhart MC, Parker WF, Beaulieu-Jones BK. Diagnostic
813 codes in AI prediction models and label leakage of same-admission clinical outcomes.
814 *JAMA Netw Open*. 2025;8(12):e2550454.

- 815 24. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing
816 clinical prediction models: a step-by-step guide. *BMJ*. 2024;386:e078276.
- 817 25. Pacheco JA, Rasmussen LV, Wiley K Jr, et al. Evaluation of the portability of
818 computable phenotypes with natural language processing in the eMERGE
819 network. *Sci Rep*. 2023;13(1):1971.
- 820 26. Orcutt X, Chen K, Mamtani R, Long Q, Parikh RB. Evaluating generalizability of
821 oncology trial results to real-world patients using machine learning-based trial
822 emulations. *Nat Med*. 2025;31(2):457-465.
- 823 27. Ackerman B, Gan RW, Meyer CS, et al. Measurement error and bias in real-world
824 oncology endpoints when constructing external control arms. *Front Drug Saf Regul*.
825 2024;4:1423493.
- 826 28. Collins GS, Chester-Jones M, Gerry S, et al. Clinical prediction models using
827 machine learning in oncology: challenges and recommendations. *BMJ Oncol*.
828 2025;4(1):e000914.
- 829 29. Yu Y, Mai Y, Zheng Y, Shi L. Assessing and mitigating batch effects in large-scale
830 omics studies. *Genome Biol*. 2024;25(1):254.
- 831 30. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial
832 intelligence for medicine and healthcare. *Nat Biomed Eng*. 2023;7(6):719-742.
- 833 31. Liu L, Bustamante R, Earles A, Demb J, Messer K, Gupta S. A strategy for
834 validation of variables derived from large-scale electronic health record data. *J*
835 *Biomed Inform*. 2021;121:103879.
- 836 32. Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in

- healthcare: review and recommendations. *Jpn J Radiol.* 2024;42(1):3-15.
33. Hui HWH, Kong W, Bin Goh WW. Thinking points for effective batch correction on biomedical data. *Brief Bioinform.* 2024;25(6):bbae515.
34. Jafari E, Blackman MH, Karnes JH, et al. Using electronic health records for clinical pharmacology research: Challenges and considerations. *Clin Transl Sci.* 2024;17(7):e13871.
35. Digitale J, Franzon D, Pletcher MJ, McCulloch CE, Gennatas ED. Methods for addressing missingness in electronic health record data for clinical prediction models: comparative evaluation. *JMIR Med Inform.* 2025;13:e79307.
36. Tan ALM, Getzen EJ, Hutch MR, et al. Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record?. *J Biomed Inform.* 2023;139:104306.
37. Ballard J, Wang Z, Li W, Shen L, Long Q. Deep learning-based approaches for multi-omics data integration and analysis. *BioData Min.* 2024;17(1):38.
38. Li M, Xu P, Hu J, Tang Z, Yang G. From challenges and pitfalls to recommendations and opportunities: Implementing federated learning in healthcare. *Med Image Anal.* 2025;101:103497.
39. Shah ST, Ali Z, Waqar M, Kim A. Federated learning in public health: a systematic review of decentralized, equitable, and secure disease prevention approaches. *Healthcare (Basel).* 2025;13(21):2760.
40. Zantvoort K, Nacke B, Görlich D, Hornstein S, Jacobi C, Funk B. Estimation of minimal data sets sizes for machine learning predictions in digital mental health

interventions. *NPJ Digit Med*. 2024;7(1):361.

41. O'Donnell A, Cronin M, Moghaddam S, Wolsztynski E. A systematic review on machine learning techniques for survival analysis in cancer. *Cancer Med*. 2025;14(22):e71375.

42. Pickett KL, Suresh K, Campbell KR, Davis S, Juarez-Colunga E. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. *BMC Med Res Methodol*. 2021;21(1):216.

43. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (N Y)*. 2023;4(9):100804.

44. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*. 2024;384:e074820.

45. Ghaffari Laleh N, Muti HS, Loeffler CML, et al. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Med Image Anal*. 2022;79:102474.

46. Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. *Nat Med*. 2024;30(10):2838-2848.

47. Nielsen C, Souza R, Wilms M, Forkert ND. Foundation model-driven distributed learning for enhanced retinal age prediction. *J Am Med Inform Assoc*. 2024;31(11):2550-2559.

48. Xu H, Usuyama N, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*. 2024;630(8015):181-188.

49. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models

- 881 can outperform medical experts in clinical text summarization. *Nat Med*.
882 2024;30(4):1134-1142.
- 883 50. Gaudet T, Day B, Jamasb AR, et al. Utilizing graph machine learning within
884 drug discovery and development. *Brief Bioinform*. 2021;22(6):bbab159.
- 885 51. Liu J, Capurro D, Nguyen A, Verspoor K. Attention-based multimodal fusion with
886 contrast for robust clinical prediction in the face of missing modalities. *J Biomed*
887 *Inform*. 2023;145:104466.
- 888 52. Reza MK, Prater-Bennette A, Asif MS. Robust multimodal learning with missing
889 modalities *via* parameter-efficient adaptation. *IEEE Trans Pattern Anal Mach Intell*.
890 2025;47(2):742-754.
- 891 53. Ktena I, Wiles O, Albuquerque I, et al. Generative models improve fairness of
892 medical classifiers under distribution shifts. *Nat Med*. 2024;30(4):1166-1173.
- 893 54. Dong W, Fong DYT, Yoon JS, et al. Generative adversarial networks for imputing
894 missing data for big data clinical research. *BMC Med Res Methodol*. 2021;21(1):78.
- 895 55. Pezoulas VC, Zaridis DI, Mylona E, et al. Synthetic data generation methods in
896 healthcare: a review on open-source tools and methods. *Comput Struct Biotechnol J*.
897 2024;23:2892-2910.
- 898 56. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus
899 guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*.
900 2025;388:e081554.
- 901 57. Chen V, Yang M, Cui W, Kim JS, Talwalkar A, Ma J. Applying interpretable
902 machine learning in computational biology-pitfalls, recommendations and

- 903 opportunities for new developments. *Nat Methods*. 2024;21(8):1454-1461.
- 904 58. Hernán MA, Dahabreh IJ, Dickerman BA, Swanson SA. The target trial
905 framework for causal inference from observational data: why and when is it helpful?
906 *Ann Intern Med*. 2025;178(3):402-407.
- 907 59. Ciobanu-Caraus O, Aicher A, Kernbach JM, Regli L, Serra C, Staartjes VE. A
908 critical moment in machine learning in medicine: on reproducible and interpretable
909 learning. *Acta Neurochir (Wien)*. 2024;166(1):14.
- 910 60. Champendal M, Müller H, Prior JO, Dos Reis CS. A scoping review of
911 interpretability and explainability concerning artificial intelligence methods in
912 medical imaging. *Eur J Radiol*. 2023;169:111159.
- 913 61. Arun N, Gaw N, Singh P, et al. Assessing the trustworthiness of saliency maps for
914 localizing abnormalities in medical imaging. *Radiol Artif Intell*. 2021;3(6):e200267.
- 915 62. Vimbi V, Shaffi N, Mahmud M. Interpreting artificial intelligence models: a
916 systematic review on the application of LIME and SHAP in Alzheimer's disease
917 detection. *Brain Inform*. 2024;11(1):10.
- 918 63. Colliot O, Thibeau-Sutre E, Burgos N. Reproducibility in Machine Learning for
919 Medical Imaging. In: Colliot O, ed. *Machine Learning for Brain Disorders*. New York,
920 NY: Humana; July 23, 2023.631-653.
- 921 64. Mahmood U, Shrestha R, Bates DDB, et al. Detecting spurious correlations with
922 sanity tests for artificial intelligence guided radiology systems. *Front Digit Health*.
923 2021;3:671015.
- 924 65. Ng CW, Wong KK, Lawson BC, Ferri-Borgogno S, Mok SC. Spatial

- transcriptome reveals histology-correlated immune signature learnt by deep learning
attention mechanism on H&E-stained images for ovarian cancer prognosis. *J Transl Med.* 2025;23(1):113.
66. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage
clinical evaluation of decision support systems driven by artificial intelligence:
DECIDE-AI. *Nat Med.* 2022;28(5):924-933.
67. Smit JM, Krijthe JH, Kant WMR, et al. Causal inference using observational
intensive care unit data: a scoping review and recommendations for future
practice. *NPJ Digit Med.* 2023;6(1):221.
68. Ling Y, Upadhyaya P, Chen L, Jiang X, Kim Y. Emulate randomized clinical trials
using heterogeneous treatment effect estimation for personalized treatments:
Methodology review and benchmark. *J Biomed Inform.* 2023;137:104256.
69. Hernán MA, Wang W, Leaf DE. Target trial emulation: a framework for causal
inference from observational data. *JAMA.* 2022;328(24):2446.
70. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating
prediction models. *Med Decis Making.* 2006;26(6):565-574.
71. Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part
1): from development to external validation. *BMJ.* 2024:e074819.
72. de Hond AAH, Shah VB, Kant IMJ, Van Calster B, Steyerberg EW,
Hernandez-Boussard T. Perspectives on validation of clinical predictive
algorithms. *NPJ Digit Med.* 2023;6(1):86.
73. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing

- 947 as a validated prediction model. *BMC Med.* 2023;21(1):70.
- 948 74. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning:
949 implications and potential remedies. *Br J Radiol.* 2023;96(1150):20220878.
- 950 75. Cabanillas Silva P, Sun H, Rezk M, et al. Longitudinal model shifts of machine
951 learning-based clinical risk prediction models: evaluation study of multiple use cases
952 across different hospitals. *J Med Internet Res.* 2024;26:e51409.
- 953 76. Piovani D, Sokou R, Tsantes AG, Vitello AS, Bonovas S. Optimizing clinical
954 decision making with decision curve analysis: insights for clinical investigators.
955 *Healthcare (Basel).* 2023;11(16):2244.
- 956 77. Verbakel JY, Steyerberg EW, Uno H, et al. ROC curves for clinical prediction
957 models part 1. ROC plots showed no added value above the AUC when evaluating the
958 performance of clinical prediction models. *J Clin Epidemiol.* 2020;126:207-216.
- 959 78. Seoni S, Jahmunah V, Salvi M, Barua PD, Molinari F, Acharya UR. Application of
960 uncertainty quantification to artificial intelligence in healthcare: a review of last
961 decade (2013–2023). *Comput Biol Med.* 2023;165:107441.
- 962 79. Kore A, Abbasi Babil E, Subasri V, et al. Empirical data drift detection
963 experiments on real-world medical imaging data. *Nat Commun.* 2024;15(1):1887.
- 964 80. Lindenmeyer A, Blattmann M, Franke S, Neumuth T, Schneider D. Towards
965 trustworthy AI in healthcare: epistemic uncertainty estimation for clinical decision
966 support. *J Pers Med.* 2025;15(2):58.
- 967 81. Griffin AC, Wang KH, Leung TI, Facelli JC. Recommendations to promote
968 fairness and inclusion in biomedical AI research and clinical use. *J Biomed Inform.*

- 969 2024;157:104693.
- 970 82. Wang F, Beecy A. Implementing AI models in clinical workflows: a roadmap.
971 *BMJ Evid Based Med.* 2025;30(5):285-287.
- 972 83. DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the
973 development-to-implementation gap in clinical artificial intelligence. *Nat Med.*
974 2021;27(2):186-187.
- 975 84. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage
976 clinical evaluation of decision support systems driven by artificial intelligence:
977 DECIDE-AI. *BMJ.* 2022;377:e070904.
- 978 85. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and
979 CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for
980 interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet*
981 *Digit Health.* 2020;2(10):e537-e548.
- 982 86. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical
983 trial protocols for interventions involving artificial intelligence: the SPIRIT-AI
984 Extension. *BMJ.* 2020:m3210.
- 985 87. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: Implications for clinical
986 decision-making. *PLOS Digit Health.* 2024;3(11):e0000651.
- 987 88. Marko JGO, Neagu CD, Anand PB. Examining inclusivity: the use of AI and
988 diverse populations in health and social care: a systematic review. *BMC Med Inform*
989 *Decis Mak.* 2025;25(1):57.
- 990 89. Andersen ES, Birk-Korch JB, Hansen RS, et al. Monitoring performance of

- 991 clinical artificial intelligence in health care: a scoping review. *JBIM Evid Synth.*
992 2024;22(12):2423-2446.
- 993 90. Subasri V, Krishnan A, Kore A, et al. Detecting and remediating harmful data
994 shifts for the responsible deployment of clinical AI models. *JAMA Netw Open.*
995 2025;8(6):e2513685.
- 996 91. Davis SE, Dorn C, Park DJ, Matheny ME. Emerging algorithmic bias: fairness
997 drift as the next dimension of model maintenance and sustainability. *J Am Med Inform*
998 *Assoc.* 2025;32(5):845-854.
- 999 92. Gilbert S, Adler R, Holoyad T, Weicken E. Could transparent model cards with
1000 layered accessible information drive trust and safety in health AI?. *NPJ Digit Med.*
1001 2025;8(1):124.
- 1002 93. Mennella C, Maniscalco U, de Pietro G, Esposito M. Ethical and regulatory
1003 challenges of AI technologies in healthcare: a narrative review. *Heliyon.*
1004 2024;10(4):e26297.
- 1005 94. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for
1006 diagnostic accuracy studies using artificial intelligence. *Nat Med.*
1007 2025;31(10):3283-3289.
- 1008 95. Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in
1009 medical artificial intelligence: a systematic review and meta-analysis. *Commun Med*
1010 *(Lond).* 2024;4(1):71.
- 1011 96. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about
1012 clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med.*

2020;26(9):1320-1324.

97. Park SH, Suh CH. Reporting guidelines for artificial intelligence studies in healthcare (for both conventional and large language models): what's new in 2024.

Korean J Radiol. 2024;25(8):687.

98. Rosenblatt M, Tejavibulya L, Jiang R, Noble S, Scheinost D. Data leakage inflates prediction performance in connectome-based machine learning models. *Nat Commun.* 2024;15(1):1829.

99. Joshi S, Urteaga I, van Amsterdam WAC, et al. AI as an intervention: improving clinical outcomes relies on a causal approach to AI development and validation. *J Am Med Inform Assoc.* 2025;32(3):589-594.

100. Sanchez P, Voisey JP, Xia T, Watson HI, O'Neil AQ, Tsafaris SA. Causal machine learning for healthcare and precision medicine. *R Soc Open Sci.* 2022;9(8):220638.

101. Wang JW, Meng M, Dai MW, Liang P, Hou J. Correlation does not equal causation: the imperative of causal inference in machine learning models for immunotherapy. *Front Immunol.* 2025;16:1630781.

102. Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digit Med.* 2023;6(1):74.

103. Ryu JS, Kang H, Chu Y, Yang S. Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomed Eng Lett.* 2025;15(5):809-830.

104. Teoh JR, Dong J, Zuo X, Lai KW, Hasikin K, Wu X. Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. *Peerj Comput Sci.* 2024;10:e2298.
105. Medetalibeyoglu A, Velichko YS, Hart EM, Bagci U. Foundational artificial intelligence models and modern medical practice. *BJR Artif Intell.* 2025;2(1):ubae018.
106. Owens K, Griffen Z, Damaraju L. Managing a “responsibility vacuum” in AI monitoring and governance in healthcare: a qualitative study. *BMC Health Serv Res.* 2025;25(1):1217.
107. Feuerriegel S, Frauen D, Melnychuk V, et al. Causal machine learning for predicting treatment outcomes. *Nat Med.* 2024;30(4):958-968.
108. Abécassis J, Dumas É, Alberge J, Varoquaux G. From prediction to prescription: machine learning and causal inference for the heterogeneous treatment effect. *Annu Rev Biomed Data Sci.* 2025;8(1):381-404.
109. Inoue K, Adomi M, Efthimiou O, et al. Machine learning approaches to evaluate heterogeneous treatment effects in randomized controlled trials: a scoping review. *J Clin Epidemiol.* 2024;176:111538.
110. Dang LE, Gruber S, Lee H, et al. A causal roadmap for generating high-quality real-world evidence. *J Clin Transl Sci.* 2023;7(1):e212.
111. Lamsaf A, Carrilho R, Neves JC, Proença H. Causality, Machine Learning, and Feature Selection: A Survey. *Sensors (Basel).* 2025;25(8):2373.
112. Zubair M, Hussain M, Albashrawi MA, Bendeche M, Owais M. A comprehensive review of techniques, algorithms, advancements, challenges, and

- clinical applications of multi-modal medical image fusion for improved diagnosis.
Comput Methods Programs Biomed. 2025;272:109014.
113. Banerji CRS, Bhardwaj Shah A, Dabson B, et al. Clinicians must participate in the development of multimodal AI. *EClinicalMedicine.* 2025;84:103252.
114. Velez T, Ibrahim Z, Duru K, et al. Predicting hospital admissions, ICU utilization, and prolonged length of stay among febrile pediatric emergency department patients using incomplete and imbalanced electronic health record (EHR) data strategies. *Int J Med Inform.* 2025;200:105905.
115. Bhandari A, Tyagi S. A comparative evaluation of handling missing data points and modalities in electronic health records. *Int J Med Inform.* 2026;210:106302.
116. Elechi U, Orobator ET, Udoh K, et al. Artificial Intelligence in Healthcare: A Narrative Review of Recent Clinical Applications, Implementation Strategies, and Challenges. *J Healthc Leadersh.* 2025;17:863-876.
117. Rajagopal A, Ayanian S, Ryu AJ, et al. Machine Learning Operations in Health Care: A Scoping Review. *Mayo Clin Proc Digit Health.* 2024;2(3):421-437.
118. Ansari S, Baur B, Singh K, Admon AJ. Challenges in the Postmarket Surveillance of Clinical Prediction Models. *NEJM AI.* 2025;2(5):10.1056/aip2401116.
119. Yan C, Zhang X, Kannampallil TG, Adler-Milstein J, Chen Y. Reimagining clinical AI: from clickstreams to clinical insights with EHR use metadata. *Npj Health Syst.* 2025;2(1):33.
120. Wells BJ, Nguyen HM, McWilliams A, et al. A practical framework for appropriate implementation and review of artificial intelligence (FAIR-AI) in

healthcare. *NPJ Digit Med.* 2025;8(1):514.

121. Hassan M, Borycki EM, Kushniruk AW. Artificial intelligence governance

framework for healthcare. *Healthc Manage Forum.* 2025;38(2):125-130.

122. Mehta V, Komanduri A, Bhadouriya RS, et al. Evaluating transparency in AI/ML

model characteristics for FDA-reviewed medical devices. *NPJ Digit Med.*

2025;8(1):673.

123. Shiferaw KB, Roloff M, Balaur I, Welter D, Waltemath D, Zeleke AA.

Guidelines and standard frameworks for artificial intelligence in medicine: a

systematic review. *JAMIA Open.* 2025;8(1):ooae155.

124. Scott IA, Reddy S, Kelly T, Miller T, van der Vegt A. Using generative artificial

intelligence in clinical practice: a narrative review and proposed agenda for

implementation. *Med J Aust.* 2025;223(11):664-672.

Figure legends

Figure 1 An end-to-end roadmap for trustworthy AI in disease pathogenesis and precision medicine. The framework delineates the biomedical AI lifecycle across five stages. **(A)** Data integration across multi-omics, imaging, EHR, and longitudinal sources. **(B)** Model selection ranging from classical ML to generative AI. **(C)** Explain & causal reasoning via interpretable features and causal modeling. **(D)** Validate through temporal/external cohorts and calibration. **(E)** Deploy for mechanistic discovery and clinically actionable precision medicine. The entire pipeline is

grounded in foundational principles of reliability, safety, validity, fairness, and transparency.

Figure 2 Data-generating processes and modality-specific validity threats. The integration of diverse biomedical data requires careful alignment with endpoint design (diagnosis, prognosis, response, mechanism) to mitigate inherent risks. **(A)** Multi-omics data face batch effects, sparsity, and feature leakage. **(B)** Imaging & pathology are susceptible to site signatures, scanner drift, and shortcut learning. **(C)** EHR & text encounter label leakage, template drift, and care pathway confounding. **(D)** Longitudinal & real-world data are challenged by informative missingness, temporal leakage, and censoring bias.

Figure 3 Evolution and taxonomy of AI models in biomedicine. The landscape progresses from classical ML (*e.g.*, logistic regression and SVMs for risk stratification) to deep learning (*e.g.*, CNN/RNNs for automatic feature extraction). Emerging paradigms include self-supervised & foundation models (*e.g.*, BERT/GPT) for transfer learning, graph/knowledge-guided networks for reasoning and drug repurposing, multimodal fusion models for joint clinical phenotyping, and generative AI for data augmentation and image synthesis. Developing these models requires balancing global strengths (interpretability, multimodal integration, scalability, and data efficiency) against inherent risks and drift (bias, overfitting, and data privacy).

Figure 4 The Evidence Ladder: Escalating AI findings from association to clinical utility. A hierarchical pyramid illustrating the increasing evidence strength and reliability of AI-derived findings. The foundation begins with association (initial correlations susceptible to confounding) and advances to robustness (cohort consistency and cross-validation). Evidence is further strengthened through biological coherence (molecular networks and pathway consistency) and causal support (DAGs and confounding control). The pinnacle of validation requires perturbation evidence (experimental validation and candidate driver testing) to ultimately achieve clinical utility delivering actionable insights.

Figure 5 The translational lifecycle and governance of biomedical AI. Clinical translation is a continuous lifecycle. **(A)** Model & data card: A standardized reporting framework functioning as a prerequisite for deployment. It documents critical metadata (provenance, endpoints, validation regimes, and calibration) to ensure transparency and reusability. **(B)** The deployment lifecycle: A dynamic, closed-loop system integrating models into clinical workflows. It requires continuous monitoring for dataset shift and accountable updating via safe recalibration and audit trails. The entire process is underpinned by transparent reporting, governance, and ethical constraints.









