Natural Resources for Human Health
eISSN: 2583-1194

Original Research

**Received** 12 April 2026
**Revised** 22 May 2026
**Accepted** 18 June 2026

Natural Resources for Human Health 2026, Vol. 6 (3s): 1086–1109

**KEYWORDS:**
smart health web systems; health informatics; publicly available datasets; electronic health records; foundation models; large language models; multimodal fusion; federated learning; HL7 FHIR; comparative model evaluation

# Smart Web-Based Representation of Human Health Data: Publicly Available Datasets, Contemporary Methodologies and a Comparative Analysis of Models

[1]Rohit Yadav; [1]Reena Hooda; [2]Manish Gupta; [3]Mamta Yadav; [1]Parshant Bansal

[1]Department of Computer Science and Engineering, IGU, Meerpur, Rewari, Haryana, India
[2]School of Computer Science, Faculty of Engineering & IT, University of Technology, Sydney, Australia
[3]Department of Computer Science and Engineering, Central University of Haryana, India
Emails: rk6yadav@gmail.com; reena.cse@igu.ac.in; manish911989@gmail.com; mamta233090@cuh.ac.in; parshant.cse@igu.ac.in

**ABSTRACT:** The web is now the surface on which most human health information is aggregated, interpreted and acted upon. Hospital portals, national biobank workbenches, consumer wearable dashboards and clinician-facing decision support tools are all, architecturally, web systems that must turn heterogeneous biomedical evidence into something a person can reason about in seconds. This paper surveys that problem as it stands in mid-2026. We propose a six-layer reference architecture that separates acquisition, interoperability, representation learning, inference, presentation and governance, and we argue that the representation and presentation layers are the ones the literature has left under-specified. We then catalogue the publicly available human-health datasets that can legitimately support such systems, reporting verified scale figures and access conditions for critical-care records (MIMIC-IV v3.1: 364,627 individuals, 546,028 hospitalisations, 94,458 ICU stays), radiographic corpora (MIMIC-CXR: 377,110 images; CheXpert: 224,316 radiographs), population genomic resources (All of Us: more than 535,000 whole genome sequences; UK Biobank: whole-genome sequencing of 490,640 participants) and few-shot benchmark suites such as EHRSHOT. Nine model families are reviewed, from gradient-boosted trees through EHR transformers, clinical and general-purpose large language models, imaging foundation models, multimodal fusion, graph neural networks, retrieval-augmented generation and federated learning. A quantitative synthesis then pairs published benchmark with the properties that decide whether a model can be served inside a web representation layer: latency class, memory footprint, interpretability, data-governance posture and regulatory exposure. Three findings stand out. Leaderboard accuracy has decoupled from deployability: reported MedQA accuracy climbed from 67.6% to 96.0% between late 2022 and late 2024, yet a 2026 Nature Medicine evaluation found specialised clinical AI tools performing no better than a general web search summary on real physician queries, with frontier general-purpose models ahead of both. Second, four of the eight quantified public resources with a defined coverage window end nine or more years before 2026, so learned representations encode historical rather than current practice. Third, of the 60 works cited here, only five address the presentation layer, against 15 each for acquisition, inference and governance. The

paper closes with a four-tier evaluation protocol and eleven open challenges, calibrated to the compliance timeline the EU Artificial Intelligence Act imposes through 2027.

## 1. INTRODUCTION

A patient checking a laboratory result on a hospital portal, an epidemiologist filtering a national surveillance dashboard, a cardiologist reading a model-generated risk score in an electronic health record sidebar, and a researcher querying a cloud biobank workbench are doing very different jobs. Architecturally they are doing the same thing: interacting with a web-delivered representation of human health data that some pipeline has selected, transformed, modelled and rendered on their behalf. Whether that interaction produces understanding or confusion depends on the quality of the representation, not only on the accuracy of the model beneath it.

Research attention has spread unevenly across this pipeline. The modelling stage has absorbed enormous investment and is measured on dozens of public leaderboards. Representation and presentation have not. A recent scoping review of health-care data dashboards found only 18 eligible studies in an initial pool of 1,644 publications spanning 2014 to 2024, reported that 72.2% concerned hospital settings, and noted that roughly a fifth described structured design guidance while the rest were built ad hoc [36]. Community and public-facing dashboards were barely represented at all. That gap matters because the deployment surface for health AI is overwhelmingly the web. The U.S. Food and Drug Administration's list of AI-enabled medical devices had grown past 1,400 entries by the end of 2025 and stood at roughly 1,500 by its 30 March 2026 cut-off, with radiology accounting for about three quarters of authorisations [38]. The clinical output of almost every one of those devices reaches a human through a browser or a browser-embedded viewer.

We therefore treat smart website representation for human health as a systems problem rather than a user-interface afterthought, and ask three linked questions. Which publicly available datasets can support the construction and independent evaluation of such systems, and under what access conditions? Which model families are currently competitive, and on what evidence? And when the two are combined, which configurations are deployable behind a web representation layer that must satisfy latency, privacy, accessibility and regulatory constraints at once?

### 1.1 Scope and definitions

We use smart health web representation system (SHWRS) for a web-delivered system that ingests one or more streams of human health data, applies a learned representation or model to them, and renders the result to a human user in a form intended to support comprehension or decision-making. The definition deliberately spans clinician-facing decision support, patient portals, population surveillance dashboards and researcher workbenches, because all four share an architectural spine and a set of failure modes.

Publicly available dataset here means a resource an unaffiliated researcher can obtain through a documented, non-discretionary process: open download, registration, or credentialing plus a data use agreement. Proprietary vendor corpora and single-institution datasets that cannot be independently re-analysed fall outside the scope, since they cannot support the reproducible comparison, this paper is about.

### 1.2 Contributions

- A six-layer reference architecture for smart health web representation systems that names the representation and presentation layers explicitly and assigns a standard, an evaluation and a characteristic failure mode to each layer (Section 3).
- A curated catalogue of publicly available human-health datasets with verified scale figures, coverage periods and access models, organised by the representation task each supports (Section 4).
- A review of nine contemporary model families that asks what each contributes to representation quality, not only to predictive accuracy (Section 5).
- A comparative analysis across four tables that separates published quantitative
- from qualitative capability and serving characteristics, with an explicit statement of why cross-study numeric comparison is unsafe (Section 6).

- A quantitative synthesis of the survey itself, reported in Section 7 with five figures: the benchmark trajectory, the scale and recency profile of the public corpus, a scored suitability matrix for the nine families, and an audit of how the cited evidence distributes across the six layers.
- A four-tier evaluation protocol and eleven open challenges calibrated to the regulatory environment applying through 2027 (Sections 8 and 9).

### 1.3 Organisation

Section 2 reviews related surveys. Section 3 develops the reference architecture. Section 4 presents the dataset catalogue and Section 5 the model review. Section 6 gives the comparative analysis and Section 7 the quantitative synthesis. Section 8 proposes an evaluation protocol, Section 9 states open challenges, and Section 10 concludes.

## 2. BACKGROUND AND RELATED WORK

### 2.1 From records to representations

Early health informatics treated the record as the object of interest, and the task was storage, retrieval and exchange. HL7 Version 2 messaging and later the Fast Healthcare Interoperability Resources (FHIR) standard formalised exchange [35], and FHIR Release 5 consolidated the resource model that most contemporary health APIs expose [34]. SMART on FHIR added an authorisation and app-launch layer, letting a third-party web application open in the context of a specific patient and a specific EHR session under scoped OAuth 2.0 credentials [33]. A RESTful resource model plus a scoped launch protocol is what makes a modern health web application possible at all, and every architecture discussed below rests on that substrate.

What changed over the last decade is that the record stopped being the terminal object. Learned representations became the working currency: dense vectors summarising a patient trajectory, an image, a genome or a clinical narrative. The shift began with the transformer architecture [1], reached clinical text through domain-specific pretraining [17], reached structured EHR sequences through clinical foundation models [16], and reached imaging through promptable segmentation [25,26]. The web layer inherited the consequence. What a dashboard displays today is often the output of a model operating on a representation, not a field read from a table.

### 2.2 Existing surveys and the gap addressed here

Strong reviews already exist for individual components. Huang et al. surveyed the fusion of medical imaging with EHR data and produced implementation guidelines [27]; Kline et al. scoped multimodal machine learning in precision health [28]; Rieke et al. set out the case for federated learning in digital health [29], and later systematic reviews assessed privacy-preserving federated learning in medical imaging specifically [31,32]. On the presentation side, the dashboard design scoping review cited above remains the most systematic treatment available [36].

None of them joins the dataset layer, the model layer and the web representation layer into one comparative frame, and that is the gap we address. The practical cost of the gap shows up in deployment. A model can lead a public leaderboard and still be unusable in a representation layer, because its latency runs to tens of seconds, because it cannot expose provenance for the claim it makes, or because the data it was trained on cannot legally sit alongside the data it will serve.

## 3. A REFERENCE ARCHITECTURE FOR SMART HEALTH WEB REPRESENTATION

We decompose a SHWRS into six layers, shown in Figure 1. The top and bottom layers follow conventional health-IT practice and are not novel. The point of the decomposition is to expose the two middle layers that usually collapse into an undifferentiated "AI" box, and to attach a specific standard, a specific evaluation and a specific failure mode to each layer.

Natural Resources for Human Health
eISSN: 2583-1194

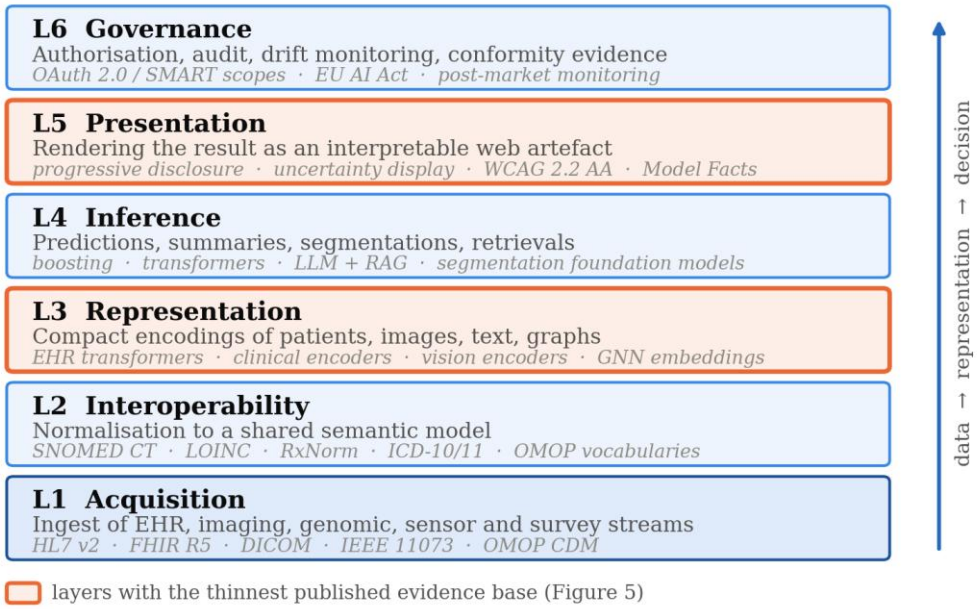


**Figure 1.** Six-layer reference architecture for a smart health web representation system. Data enters at L1 and is rendered to a human at L5, with L6 spanning the stack as a governance obligation rather than a processing step. The two highlighted layers are the ones this survey argues are under-specified relative to their importance; Figure 5 quantifies that claim against the cited literature.

**Table 1.** The six layers, their standards, evaluations and characteristic failure modes.

| Layer | Function | Representative standards / techniques | Primary evaluation | Characteristic failure mode |
|---|---|---|---|---|
| L1 Acquisition | Ingest EHR, imaging, genomic, sensor and survey streams | HL7 v2, FHIR R5 resources, DICOM, IEEE 11073, OMOP CDM | Completeness, timeliness, missingness profile | Silent feed loss; unit and timezone drift |
| L2 Interoperability | Normalise to a shared semantic model | SNOMED CT, LOINC, RxNorm, ICD-10/11, OMOP vocabularies | Mapping coverage and concept collision rate | Local codes mapped to the wrong parent concept |
| L3 Representation | Learn compact encodings of patients, images, text, graphs | EHR transformers, clinical LLM encoders, vision encoders, GNN embeddings | Downstream transfer, few-shot sample efficiency | Representation encodes site identity rather than physiology |
| L4 Inference | Produce predictions, summaries, segmentations, retrievals | Gradient boosting, transformers, LLM + RAG, segmentation foundation models | AUROC, AUPRC, DSC, calibration, factuality | Strong discrimination with miscalibrated probabilities |

| Layer | Function | Representative standards / techniques | Primary evaluation | Characteristic failure mode |
|---|---|---|---|---|
| L5 Presentation | Render the result as an interpretable web artefact | Progressive disclosure, uncertainty display, WCAG 2.2 AA, Model Facts labels | Task success, time-to-comprehension, error rate | Correct output that the user systematically misreads |
| L6 Governance | Authorise, audit, log, monitor and comply | OAuth 2.0 / SMART scopes, audit trails, drift monitors, EU AI Act conformity | Auditability, drift alarms, incident traceability | Model silently degrades after a case-mix change |

*Layers L3 and L5 are the ones this survey argues are under-specified. L4 dominates the published literature, while L1, L2 and L6 are governed by mature external standards.*

### 3.1 Layers L1 and L2: acquisition and interoperability

Acquisition and interoperability are the best-standardised part of the stack and, for that reason, the least interesting to researchers. They are also where production systems most often break. FHIR R5 supplies the resource vocabulary and the RESTful access pattern [34]; SMART on FHIR supplies scoped, patient-contextual authorisation for third-party web applications [33]. The Observational Medical Outcomes Partnership (OMOP) Common Data Model supplies the analytic normalisation target that makes multi-site training tractable.

Failures at this layer are semantic, not syntactic. A locally defined laboratory code mapped to an approximately correct LOINC parent passes every schema validation and still shifts a model's input distribution. Because the error is invisible at the presentation layer, it surfaces months later as unexplained model degradation, which is one concrete form of the dataset shift problem described by Finlayson et al. [56].

### 3.2 Layer L3: the representation layer

The representation layer is where a heterogeneous, irregularly sampled, multi-modal patient record becomes something a model can consume. Three design decisions dominate. Tokenisation granularity determines whether a patient timeline is encoded as a sequence of clinical codes, as code-value pairs, or as time-bucketed aggregates. Temporal encoding determines whether inter-event intervals survive or are discarded. Missingness handling matters most of all, because absence in clinical data is rarely random: a laboratory test that was not ordered says something about what the clinician suspected.

One representation-layer pathology deserves particular attention. When a model trains on multi-site data, gradient descent will happily spend capacity encoding which hospital a record came from, since site is often the single most predictive latent variable for outcomes that depend on local practice. Such a representation transfers badly, and it is a plausible contributor to the external validation failures documented for widely deployed proprietary models [57].

### 3.3 Layer L5: the presentation layer

Presentation is where the system succeeds or fails as an act of communication. Four requirements recur across the literature and across regulatory guidance.

- Progressive disclosure. The default view should carry the decision-relevant signal. Supporting evidence, feature attributions and provenance belong one interaction away, not competing for attention in the default state.
- Uncertainty as a first-class visual element. A calibrated probability rendered as a bare number gets over-read. Interval displays, reference-class framing ("12 of 100 similar patients") and explicit abstention states change how the output is understood.

- Provenance and model transparency. The Model Facts label proposed by Sendak et al. gives a compact, standardised summary of a model's intended use, validation population and known limitations, designed to sit next to the model's output rather than in a document nobody opens [54].
- Accessibility. WCAG 2.2 at level AA is the operative baseline for public-sector and, increasingly, private health web properties [37]. For health data visualisation that means encoding that does not rely on colour alone, keyboard-navigable chart elements, text alternatives for every graphical claim, and contrast maintained in both light and dark presentations.

The empirical basis for these requirements is thin. The dashboard scoping review found usability, accessibility and interactivity treated inconsistently, and most reviewed systems built without reference to structured design guidance [36]. Measured against the maturity of the modelling literature, this is the largest evidence gap in the field, and Section 7.4 quantifies it.

### 3.4 Layer L6: governance and the regulatory clock

Governance has stopped being optional or purely institutional. Under Regulation (EU) 2024/1689, the Artificial Intelligence Act, most of the Regulation's provisions became applicable on 2 August 2026, while Article 6(1) applies from 2 August 2027. Article 6(1) classifies as high-risk those AI systems that are products, or safety components of products, covered by Annex I harmonisation legislation, a category that includes medical devices [39]. Any SHWRS placed on the EU market whose output informs clinical decisions should therefore be built now for risk management documentation, data governance evidence, logging, human oversight and post-market monitoring. Retrofitting those obligations into a deployed representation layer costs considerably more than designing for them.

## 4. PUBLICLY AVAILABLE HUMAN-HEALTH DATASETS

This section catalogues public datasets capable of supporting the construction and, more importantly, the independent evaluation of smart health web representation systems. Figures come from the dataset descriptors or the custodians' current release documentation. Release versions are stated wherever a resource is versioned, because scale figures for living resources change between releases.

### 4.1 A taxonomy by representation task

Datasets are more useful when organised by the representation task they support than by nominal modality. We distinguish five classes. Longitudinal clinical trajectories (A) support patient-state representation and temporal prediction. Image-report pairs (B) support visual grounding and automated reporting. Population cohorts with linked genomics (C) support risk stratification and stratified representation. Continuous physiological streams (D) support real-time monitoring. Benchmark suites (E) exist to make representations comparable.

**Table 2.** Representative publicly available human-health datasets.

| Dataset (version) | Class | Modality | Reported scale | Coverage | Access model |
|---|---|---|---|---|---|
| MIMIC-IV v3.1 | A | ICU and hospital EHR: labs, medications, vitals, orders, notes (linked modules) | 364,627 unique individuals; 546,028 hospitalisations; 94,458 ICU stays | 2008-2022; released 11 Oct 2024 | Credentialed PhysioNet account, CITI training, signed DUA |
| eICU-CRD v2.0 | A | Multi-centre tele-ICU EHR | 200,859 ICU unit encounters for 139,367 unique patients; 335 units at 208 U.S. hospitals | 2014-2015 | Credentialed PhysioNet, DUA |

| Dataset (version) | Class | Modality | Reported scale | Coverage | Access model |
|---|---|---|---|---|---|
| MIMIC-CXR / CXR-JPG | B | Chest radiographs with free-text radiology reports | 377,110 images from 227,835 radiographic studies of 65,379 patients | 2011-2016 | Credentialed PhysioNet, DUA |
| CheXpert | B | Chest radiographs, 14 observation labels with explicit uncertainty | 224,316 radiographs from 65,240 patients; 500-study expert test set labelled by 8 radiologists | Oct 2002 - Jul 2017 | Registration and Stanford research use agreement |
| All of Us Research Program | C | EHR-linked genomics, surveys, physical measurements, wearables | 883,000+ enrolled; 747,000 with data available; 535,000+ whole genome sequences; 482,000 EHR records linked to genomic data; 86% from historically under-represented communities | Enrolment 2018-present; release of 30 Jun 2026 | Tiered Researcher Workbench; registered and controlled tiers |
| UK Biobank | C | Deep phenotyping, multi-organ imaging, whole-genome sequencing, linked health records | Approximately 500,000 participants; whole-genome sequencing released for 490,640 participants | Recruitment 2006-2010; follow-up ongoing | Approved research application; access fee |
| NHANES | C | Nationally representative interview, examination and laboratory survey | Continuous two-year cycles; nationally representative U.S. sample | 1999-present (continuous) | Fully open download |
| TCGA | C | Multi-omics with matched digital pathology | Approximately 11,000 patients across 33 cancer types | 2006-2013 accrual | Open summary tier; controlled tier via dbGaP |
| ADNI | C | Serial MRI, PET, CSF and cognitive biomarkers for Alzheimer's disease | Multi-phase longitudinal cohort | 2004-present | Application and DUA |

| Dataset (version) | Class | Modality | Reported scale | Coverage | Access model |
|---|---|---|---|---|---|
| PhysioNet signal archives | D | ECG, PPG, EEG, arterial waveforms and other continuous physiology | 100+ curated databases across open and credentialed tiers | 1999-present | Open and credentialed tiers |
| EHRSHOT | E | Longitudinal structured EHR benchmark, not restricted to ICU or ED | 6,739 patients; 15 few-shot clinical prediction tasks; released with CLMBR-T-base, a 141M-parameter model pretrained on structured EHR of 2.57M patients | Stanford Medicine cohort | Research data use agreement |
| MedSAM training corpus | B/E | Aggregated public segmentation sets across 10 imaging modalities | 1,570,263 image-mask pairs spanning over 30 cancer types | Aggregated from public sources | Open (component licences apply) |

*Class key: A longitudinal clinical trajectories; B image-report pairs; C population cohorts with linked genomics; D continuous physiological streams; E benchmark suites. Figures for living resources reflect the release stated in the coverage column. Values marked "approximately" are the widely cited descriptor figures rather than a current custodian count.*

### 4.2 Access models and their architectural consequences

Access model is not an administrative detail; it constrains the architecture. Four regimes recur. Open download (NHANES, open PhysioNet tiers) lets data sit alongside application code and supports fully client-side or edge deployment. Registration-based access (CheXpert) permits similar architectures but imposes redistribution limits that rule out shipping the data inside a public web bundle. Credentialed access with a data use agreement (MIMIC-IV, MIMIC-CXR, eICU-CRD) requires every person touching the data to hold individual credentials and complete human-subjects training, so a demonstration website cannot expose record-level data at all, only derived and aggregated artefacts. Enclave or workbench access (All of Us, UK Biobank) is the most restrictive and the most architecturally consequential: analysis happens inside the custodian's compute environment and only approved aggregate outputs leave it. A SHWRS built on enclave data is necessarily two-tier, with training inside the enclave and only model weights or aggregate statistics crossing the boundary.

These regimes exist because de-identification is not anonymisation. Longitudinal, high-dimensional health records remain re-identifiable in principle, and credentialing works as an accountability mechanism rather than a technical control. A public-facing representation layer built on such data has to enforce aggregation thresholds and suppression rules at the presentation layer, not only at the query layer.

### 4.3 Known limitations of the public corpus

- Geographic concentration. The most heavily used clinical datasets come from a small number of academic medical centres in the United States and the United Kingdom, so models derived from them encode the case mix, coding practice and care pathways of those settings.
- Temporal staleness. Coverage windows lag the present by years. MIMIC-IV extends to 2022 and CheXpert to 2017, so learned representations encode historical practice, historical therapeutic options and historical documentation conventions. Section 7.2 quantifies the lag across the corpus.

- Label provenance. Large radiographic label sets are extracted automatically from free-text reports rather than annotated afresh. CheXpert's explicit uncertainty class is an honest acknowledgement of this, but it also means the label represents the report, not the image.
- Representation inequity. All of Us is a deliberate corrective, with 86% of participants from historically under-represented communities [11,12], but it remains the exception. Given the demonstrated capacity of biased algorithms to propagate inequity through health systems [55], cohort composition is a first-order modelling concern rather than a limitation to note in passing.
- Reproducibility friction. Credentialing and enclave requirements, however justified, make independent replication slower and rarer than in adjacent machine learning fields, which weakens the evidence base underpinning deployment decisions. Adherence to the FAIR principles for metadata and provenance reduces this friction without removing it [58].

## 5. CONTEMPORARY METHODOLOGIES AND MODEL FAMILIES

This section reviews nine model families. For each we state the representation it produces, the data regime in which it is competitive, and what it implies for the web representation layer. The last criterion is the one methodological reviews most often omit.

### 5.1 Gradient-boosted decision trees: the baseline that will not go away

On tabular, feature-engineered clinical data with sample sizes in the thousands to low millions, gradient-boosted decision trees [42,43] remain a formidable and frequently unbeaten baseline. They train in minutes, produce well-behaved probabilities after modest calibration, expose exact additive feature attributions through SHAP [44], and serve inference in single-digit milliseconds from a footprint measured in megabytes.

For a web representation layer those properties are close to ideal. The model runs server-side at negligible cost or compiles to run in the browser, and its attributions render directly as an explanation panel without a post-hoc approximation step. Any claim that a deep model should replace boosting in a production health dashboard ought to demonstrate a benefit that survives calibration and justifies the loss of attribution fidelity.

### 5.2 Transformer models over structured EHR sequences

Treating a patient record as a sequence of clinical events and applying the transformer architecture [1] yields patient-state representations that serve many downstream tasks from a single pretraining run. EHRSHOT was built to test whether that promise holds under few-shot conditions, releasing 15 clinical prediction tasks over 6,739 longitudinal patients not restricted to the ICU, together with CLMBR-T-base, a 141-million-parameter clinical foundation model pretrained on the structured records of 2.57 million patients [16]. The design point matters, because earlier EHR benchmarks were largely confined to ICU or emergency department episodes, precisely the settings where dense sampling makes prediction easiest and generalisation weakest.

What this family offers a representation layer is sample efficiency rather than peak accuracy. A pretrained patient encoder lets a new clinical question be answered with tens or hundreds of labelled examples instead of tens of thousands, which is the realistic label budget for most institutional deployments.

### 5.3 Clinical language models

GatorTron established the value of domain-specific pretraining on clinical narrative. It was trained on more than 90 billion words of text, including over 82 billion words of de-identified clinical text, and released at 345 million, 3.9 billion and 8.9 billion parameters [17]. Across clinical concept extraction, medical relation extraction, semantic textual similarity, natural language inference and medical question answering, the large model improved natural language inference accuracy by 9.6% over BioBERT and reached an F1 of 0.9627 on drug-adverse event relation extraction [17].

Encoder-scale models remain the pragmatic choice for the structured extraction tasks that feed a representation layer, turning narrative into codes, entities and relations a dashboard can filter and aggregate. They are self-hostable, deterministic in output shape, and cheap enough to run over an entire corpus rather than over a single query.

### 5.4 Generative large language models in medicine

The generative branch has moved fast, and MedQA provides a continuous thread through that movement. Flan-PaLM with instruction prompt tuning reached 67.6% [18]. GPT-4 exceeded the USMLE passing threshold by more than 20 points without specialised prompting [20]. Med-PaLM 2 reported 86.5% [19,22], and Med-Gemini reported 91.1%, claiming state-of-the-art

on 10 of 14 medical benchmarks and improvements of up to 12% over prior state of the art on chest X-ray report generation [22]. A systematic study of run-time strategies reported 78.9% for zero-shot GPT-4, 81.4% for five-shot GPT-4-Turbo, 84.4% for zero-shot GPT-4o, 90.2% for GPT-4 Turbo with the Medprompt strategy, and 96.0% for a zero-shot reasoning-native model, concluding that architectural reasoning capability had largely displaced elaborate prompt engineering [21]. Figure 2 plots that trajectory.

More consequential is the recent evidence on the gap between benchmarks and practice. A 2026 Nature Medicine study evaluated two specialised clinical AI tools against three frontier general-purpose models across a three-stage protocol: 500 MedQA questions, 500 HealthBench items measuring clinician alignment, and a Real Clinical Queries benchmark of 100 de-identified physician queries assessed by 12 U.S. clinicians in randomised blinded review, yielding 1,800 annotations [23,24]. Frontier general-purpose models outperformed the specialised clinical tools in all three evaluations, and the specialised tools performed comparably to a general web search AI overview on real-world queries [23]. The authors' conclusion, that independent real-world evaluation must precede clinical deployment, is the single most important design constraint on the inference layer of any SHWRS built today.

### 5.5 Imaging foundation models

Promptable segmentation generalised from natural images [26] to medicine with MedSAM, trained on 1,570,263 image-mask pairs spanning 10 imaging modalities and over 30 cancer types, and validated on 86 internal and 60 external segmentation tasks where it achieved better accuracy and robustness than modality-specific specialist models [25]. For a representation layer this family matters for a reason beyond accuracy. A promptable model turns segmentation into an interactive operation: the user clicks or draws a box, the model responds, and the boundary becomes an editable object rather than a fixed overlay. Clinician correction becomes part of the workflow instead of an exception path.

### 5.6 Multimodal fusion

Clinical reasoning is multimodal, and fusion architectures try to reflect that. The design space divides conventionally into early fusion at the input, joint or intermediate fusion in a shared latent space, and late fusion of modality-specific predictions. Systematic reviews of imaging-plus-EHR fusion [27] and of multimodal machine learning in precision health [28] converge on two findings: fusion generally improves on the best single modality, and the improvement is usually modest relative to the increase in engineering complexity, missing-modality fragility and evaluation difficulty.

Missing-modality handling decides whether such a system is usable. A fusion model requiring imaging, laboratory and genomic inputs at once will fall back to a degraded path for most real patients, and unless the representation layer displays which modalities actually contributed, the user has no way to calibrate trust in the output.

### 5.7 Graph neural networks and biomedical knowledge graphs

Graph neural networks [46] operate naturally over biomedical relational structure. Decagon demonstrated polypharmacy side-effect prediction over a multimodal drug-protein interaction graph [48], and PrimeKG assembled a precision-medicine knowledge graph integrating drugs, diseases, phenotypes and biological pathways into one queryable structure [47].

For representation systems the graph family has a distinctive advantage: the explanation is a path. When a model predicts an interaction risk, the supporting subgraph renders directly, and a clinician can inspect the chain of entities and relations that produced the prediction. That is a stronger form of explanation than a feature-importance bar chart, because it can be checked against domain knowledge rather than merely describing model behaviour.

### 5.8 Retrieval-augmented generation

Retrieval-augmented generation [40] conditions a generative model on documents retrieved at query time. Its appeal in medicine is threefold: the knowledge base updates without retraining, outputs ground in citable sources, and unsupported generation

falls. Reviews of RAG in medical applications report consistent reductions in unsupported output relative to unaugmented generation, bounded by the quality of retrieval [41].

Seen from the representation layer, RAG may be the most important recent development, because it is the only widely deployed technique that makes a generative output auditable at the point of display. A statement accompanied by the passage supporting it can be checked by the reader; a statement without one cannot. Any generative component in a health web system should be considered incomplete until the presentation layer renders its supporting evidence beside its claim.

### 5.9 Federated and privacy-preserving learning

Federated learning trains a shared model across institutions without centralising patient data [29] and has been demonstrated at scale in multi-institutional medical imaging collaborations [30]. Recent systematic reviews of privacy-preserving federated learning in medical imaging catalogue the practical obstacles: statistical heterogeneity across sites, communication cost, and the fact that federation alone does not prevent information leakage from model updates, which motivates combination with differential privacy [59,60] or secure aggregation [31,32].

For a SHWRS the relevance is jurisdictional. Where data cannot cross an institutional or national boundary, federation lets one representation layer be backed by a model informed by all of them. The cost is a much harder evaluation problem, since no single party holds a validation set reflecting the full training distribution.

### 5.10 Explainability as a representation-layer requirement

SHAP [44] and Grad-CAM [45] remain the workhorses of post-hoc explanation in tabular and imaging settings. Both describe model behaviour rather than offering causal accounts, and both get over-interpreted when rendered without qualification. Responsibility here sits with the presentation layer: a saliency map displayed without an explicit statement of what it does and does not establish will be read as evidence of reasoning. Structured disclosure formats such as Model Facts labels [54] address that framing problem more directly than any additional attribution method.

## 6. COMPARATIVE ANALYSIS

This section compares the model families along three axes: published quantitative performance, qualitative capability, and fitness for service inside a web representation layer. One constraint governs the first of these and should be stated up front.

**A note on comparability.** The figures in Table 3 are reported as published and are not directly comparable to one another. They differ in benchmark version and option count, in dataset split, in prompting and run-time strategy, in whether the evaluation was zero-shot or few-shot, in evaluation date, and in how independent the evaluating party was from the model developer. MedQA figures deserve particular caution, since the benchmark predates web-scale pretraining corpora that plausibly contain it, so absolute values should be read as an upper bound and cross-model differences of a few percentage points should not be read as meaningful at all. The table shows trajectory and rough ordering, not a ranking.

**Table 3.** Reported benchmark for representative models (as published; see comparability note).

| Model / system | Benchmark | Reported result | Source |
|---|---|---|---|
| Flan-PaLM 540B (instruction prompt tuned) | MedQA (USMLE) | 67.6% accuracy | Singhal et al., Nature 2023 [18] |
| GPT-4, zero-shot | MedQA (US, 4-option) | 78.9% accuracy | Nori et al., arXiv 2024 [21] |
| GPT-4-Turbo, 5-shot | MedQA (US, 4-option) | 81.4% accuracy | Nori et al., arXiv 2024 [21] |
| GPT-4o, zero-shot | MedQA (US, 4-option) | 84.4% accuracy | Nori et al., arXiv 2024 [21] |

| Model / system | Benchmark | Reported result | Source |
|---|---|---|---|
| Med-PaLM 2 | MedQA (USMLE) | 86.5% accuracy | Singhal et al., Nat Med 2025 [19]; reported in [22] |
| GPT-4 Turbo + Medprompt | MedQA (US, 4-option) | 90.2% accuracy | Nori et al., arXiv 2024 [21] |
| Med-Gemini | MedQA (USMLE) | 91.1% accuracy; state of the art on 10 of 14 medical benchmarks | Saab et al., arXiv 2024 [22] |
| o1-preview, zero-shot | MedQA (US, 4-option) | 96.0% accuracy | Nori et al., arXiv 2024 [21] |
| Med-Gemini | Chest X-ray report generation | Exceeded prior state of the art by up to 12% | Saab et al., arXiv 2024 [22] |
| Frontier general-purpose LLMs (GPT-5.2, Gemini 3.1 Pro, Claude Opus 4.6) vs. two specialised clinical AI tools | MedQA (500 items), HealthBench (500 items), Real Clinical Queries (100 physician queries, 12 clinicians, 1,800 blinded annotations) | Frontier general-purpose models outperformed specialised clinical AI tools in all three evaluations; specialised tools performed comparably to a general search AI overview on real-world queries | Vishwanath et al., Nature Medicine 2026 [23,24] |
| GatorTron-large (8.9B) | Clinical natural language inference | +9.6% accuracy over BioBERT | Yang et al., npj Digital Medicine 2022 [17] |
| GatorTron | Drug-adverse event relation extraction | F1 = 0.9627 | Yang et al., npj Digital Medicine 2022 [17] |
| GatorTron | Clinical concept extraction | F1 = 0.8996 / 0.8091 / 0.9000 across three datasets | Yang et al., npj Digital Medicine 2022 [17] |
| MedSAM | Medical image segmentation, 86 internal and 60 external tasks | Better accuracy and robustness than modality-specific specialist models; highest median DSC on most tasks | Ma et al., Nature Communications 2024 [25] |
| CLMBR-T-base (141M) | EHRSHOT, 15 few-shot clinical prediction tasks | Released as the reference clinical foundation model for few-shot EHR evaluation; pretrained on structured EHR of 2.57M patients | Wornow et al., NeurIPS D&B 2023 [16] |

**Table 4.** Qualitative comparison of model families for health representation.

| Family | Representation produced | Data regime where competitive | Native explanation | Label efficiency | Principal limitation |
|---|---|---|---|---|---|
| Gradient-boosted trees | None (direct feature-to-outcome mapping) | Tabular, 10^3-10^6 rows, engineered features | Exact additive attributions (SHAP) | Low - needs full supervision | No transfer; manual feature engineering per task |
| EHR transformers | Dense patient-trajectory embedding | 10^5-10^7 patient timelines for pretraining | Attention weights (weak explanation) | High - few-shot adaptation | Risk of encoding site identity; heavy pretraining cost |
| Clinical encoder LLMs | Contextual embeddings of clinical narrative | 10^9-10^11 words of clinical text | Token-level attributions | Moderate | Requires large in-domain clinical corpus |
| Generative LLMs | Natural-language output; latent knowledge | Web-scale pretraining plus alignment | Chain-of-thought (post hoc, not faithful) | Very high - zero/few-shot | Unsupported generation; opaque provenance; cost and latency |
| Imaging foundation models | Promptable dense masks and image embeddings | 10^6+ image-mask pairs | The mask itself is the explanation | High - prompt-driven | Needs a prompt; anatomical labelling still required |
| Multimodal fusion | Joint cross-modal latent space | Aligned multi-modal cohorts | Modality attribution | Low - alignment is scarce | Missing-modality fragility; high engineering complexity |
| Graph neural networks | Node and subgraph embeddings | Curated biomedical knowledge graphs | Explanatory paths through the graph | Moderate | Graph completeness and curation bias |
| Retrieval-augmented generation | Query-conditioned grounded text | Any indexed corpus; no retraining | Retrieved passages as citations | Very high | Bounded by retrieval quality; index freshness burden |
| Federated learning | Shared weights without shared data | Multi-site, jurisdictionally separated | Inherited from the base model | Inherited | Statistical heterogeneity; leakage via updates; hard evaluation |

**Table 5.** Serving profile and fitness for a web representation layer.

| Family | Typical inference latency | Serving footprint | Feasible execution site | Data-governance posture | Representation-layer fit |
|---|---|---|---|---|---|
| Gradient-boosted trees | < 10 ms | MB | Browser, edge, or server | Fully self-hosted; no data egress | Excellent - interactive, explainable, cheap |
| EHR transformers | 10-200 ms | 0.5-5 GB | Server or private cloud | Self-hosted; training data stays in enclave | Good - suits precomputed risk panels |
| Clinical encoder LLMs | 20-300 ms | 1-20 GB | Server or private cloud | Self-hosted; PHI need not leave the perimeter | Good - batch extraction feeding dashboards |
| Generative LLMs (frontier) | 1-30 s | API-hosted or 10^2 GB self-hosted | Third-party API or large private cluster | PHI egress unless self-hosted; contractual controls required | Constrained - unsuitable for synchronous critical paths |
| Imaging foundation models | 0.1-2 s per prompt | 0.5-3 GB (distilled variants smaller) | GPU server; distilled variants on edge | Self-hosted; images need not leave the perimeter | Good - enables interactive annotation |
| Multimodal fusion | 0.2-3 s | 1-10 GB | GPU server | Self-hosted, but multiple linked data sources | Moderate - degraded paths must be surfaced |
| Graph neural networks | 10-200 ms | MB-GB plus graph store | Server | Knowledge graph is typically public | Good - path explanations render directly |
| RAG pipelines | 1-10 s | Index size dominates | Server plus generation backend | Index may be local while generation is remote | Good for reference tasks; must display citations |
| Federated learning | Inherited at inference | Inherited | Local inference at each site | Strongest - raw data never moves | Good where jurisdiction blocks centralisation |

*Latency and footprint values are indicative order-of-magnitude ranges for typical 2026 deployments on commodity hardware, offered to support architectural reasoning. They are not measurements from a controlled benchmark.*

### 6.1 What the comparison shows

**Benchmark accuracy has decoupled from clinical usefulness.** The MedQA trajectory in Table 3 and Figure 2, from 67.6% to 96.0% over roughly two years, is real and increasingly uninformative. The Nature Medicine three-stage evaluation supplies the corrective: tested on 100 de-identified questions physicians actually asked, and judged blind by 12 clinicians, purpose-built clinical tools performed no better than a general search summary [23,24]. Benchmark leadership belongs in a screening criterion, not in an argument for deployment.

**Specialisation is no longer a reliable advantage.** That general-purpose frontier models outperformed dedicated clinical AI products [23] inverts the assumption motivating a decade of domain-specific pretraining. The inversion is partial. Encoder-scale clinical models such as GatorTron remain the right tool for high-throughput structured extraction [17], where self-hosting, determinism and cost matter more than reasoning breadth. But for open-ended clinical question answering the burden of proof has moved onto the specialised system.

**The serving profile, not the accuracy, determines the architecture.** Table 5 makes the tension explicit: the most accurate systems occupy the worst cell of the serving matrix, with multi-second latency, third-party hosting and protected health information egress. A representation layer that must respond within one interaction cycle cannot put such a model on its synchronous path. What follows in practice is a tiered architecture, with cheap self-hosted models on interactive elements and expensive generative models reserved for asynchronous or explicitly user-initiated tasks where a visible wait is acceptable.

**Grounding is the deployable form of trustworthiness.** Across families, the techniques that improve trust at the presentation layer are those producing an inspectable artefact: a retrieved passage [40,41], an explanatory subgraph [47,48], an editable segmentation mask [25]. Post-hoc attribution methods produce artefacts that look inspectable but cannot be verified against domain knowledge in the same way. That distinction should drive model selection for any component whose output a clinician will see.

**Evaluation independence is now a design requirement.** One pattern recurs across the evidence base: external validation failure of a widely deployed proprietary sepsis model [57], algorithmic bias propagating through a health-system risk tool [55], specialised clinical AI underperforming on real queries [23]. Developer-reported performance systematically overstates deployed performance. Reporting frameworks including TRIPOD+AI [51], CONSORT-AI [52] and DECIDE-AI [53] exist to constrain this, and adherence belongs at a gate rather than in a list of good practices.

## 7. A QUANTITATIVE SYNTHESIS OF THE REVIEWED EVIDENCE

This section reports what the survey itself found when its inputs are quantified. No new experiments were run. Three of the five

below are computed from published figures verified against their primary sources, one is an author-assigned ordinal scoring whose rubric is stated in full, and one is an audit of this paper's own reference list. Section 7.5 reports the integrity check applied to every citation.

### 7.1 The benchmark trajectory and its decoupling from practice

Figure 2 plots every MedQA result in Table 3 against the approximate date of its report, distinguishing general-purpose models from medically adapted ones. Reported accuracy rose 28.4 percentage points in roughly two years, from 67.6% in late 2022 to 96.0% in late 2024. The mean of the eight reported values is 84.5% and the median 85.5%; all eight sit above the approximate USMLE pass equivalence of 60%.

Natural Resources for Human Health
eISSN: 2583-1194

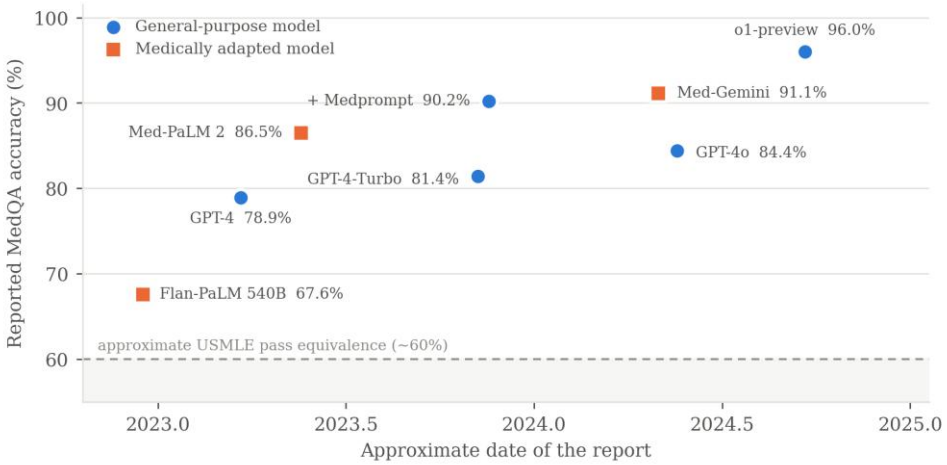


**Figure 2.** Reported MedQA (USMLE-style) accuracy against the approximate date of each report, for the eight

collected in Table 3. Values are reported as published and are not directly comparable across studies (Section 6). Medically adapted models held the highest reported value at two points in the sequence, in mid-2023 and mid-2024, before a general-purpose reasoning model overtook them in late 2024.

Two observations follow. First, medical adaptation produced a clear advantage early: Med-PaLM 2 at 86.5% led zero-shot GPT-4 at 78.9% by 7.6 points, and Med-Gemini at 91.1% led zero-shot GPT-4o at 84.4% by 6.7 points. Second, that advantage did not survive the arrival of reasoning-native general-purpose models, with the 96.0% zero-shot result exceeding the best medically adapted figure by 4.9 points. Read alongside the 2026 Nature Medicine finding that specialised clinical tools performed comparably to a general search summary on real physician queries [23,24], the pattern suggests the benchmark itself has stopped discriminating in the region where these systems now operate.

### 7.2 Scale and recency of the public dataset corpus

Figure 3 profiles the ten quantified resources in Table 2 on two axes. Panel (a) shows reported scale on a logarithmic axis, spanning 6,739 patients in EHRSHOT to 1,570,263 image-mask pairs in the MedSAM training corpus, a range of more than two orders of magnitude. Six of the ten resources report at least 200,000 records, so scale is not the binding constraint for most representation-learning work.

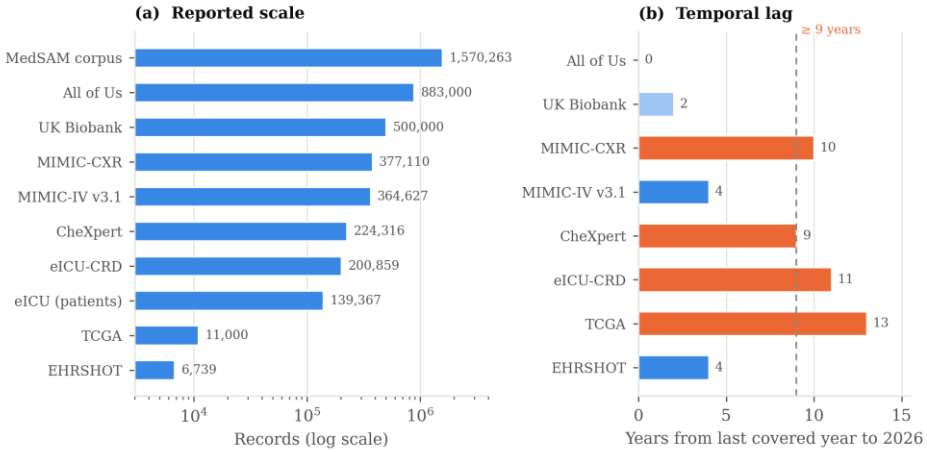


**Figure 3.** The public human-health dataset corpus quantified on two axes. (a) Reported scale, logarithmic, in the units each custodian reports. (b) Years elapsed between the final year of data coverage and 2026; resources at nine years or more are highlighted. Values are taken from Table 2. The MedSAM corpus is omitted from panel (b) because it aggregates sources with heterogeneous coverage windows, and eICU is counted once.

Panel (b) is the more consequential result. Across the eight distinct resources with a defined coverage window, the median lag between the last year of data and 2026 is 6.5 years and the mean is 6.6 years. Four of the eight lag by nine years or more, and those four include the most heavily used imaging and critical-care corpora: TCGA at 13 years, eICU-CRD at 11, MIMIC-CXR at 10 and CheXpert at 9. Only All of Us and UK Biobank, both actively recruiting population cohorts, lag by two years or less. The practical implication is that a model trained on the standard public corpus learns the documentation conventions, coding practice and therapeutic options of the mid-2010s, and any SHWRS built on it inherits that vintage unless the representation is explicitly re-fitted on contemporary institutional data.

**7.3 A scored suitability matrix for the nine model families**

Figure 4 scores each family on five criteria that determine fitness for a web representation layer. Scores are ordinal on a 1-5 scale and were assigned by the authors from the qualitative properties tabulated in Tables 4 and 5. They are a structured restatement of the review, not measurements, and they should be read as such.

The rubric is as follows. Label efficiency scores how many labelled examples a new task requires, from 1 (full supervision, tens of thousands) to 5 (zero-shot or few-shot). Explanation inspectability scores whether the model's justification can be checked against domain knowledge, from 1 (no faithful explanation available) to 5 (the explanation is an artefact a domain expert can verify, such as a subgraph or an editable mask). Interactive-latency fit scores whether the model can sit on the synchronous path of a web interaction, from 1 (multi-second) to 5 (sub-10-millisecond). Privacy and self-hosting scores whether protected health information can remain inside the institutional perimeter, from 1 (mandatory third-party egress) to 5 (raw data never moves). Implementation simplicity scores the engineering burden of a production deployment, from 1 (multi-site coordination or scarce aligned data required) to 5 (a single self-contained artefact).

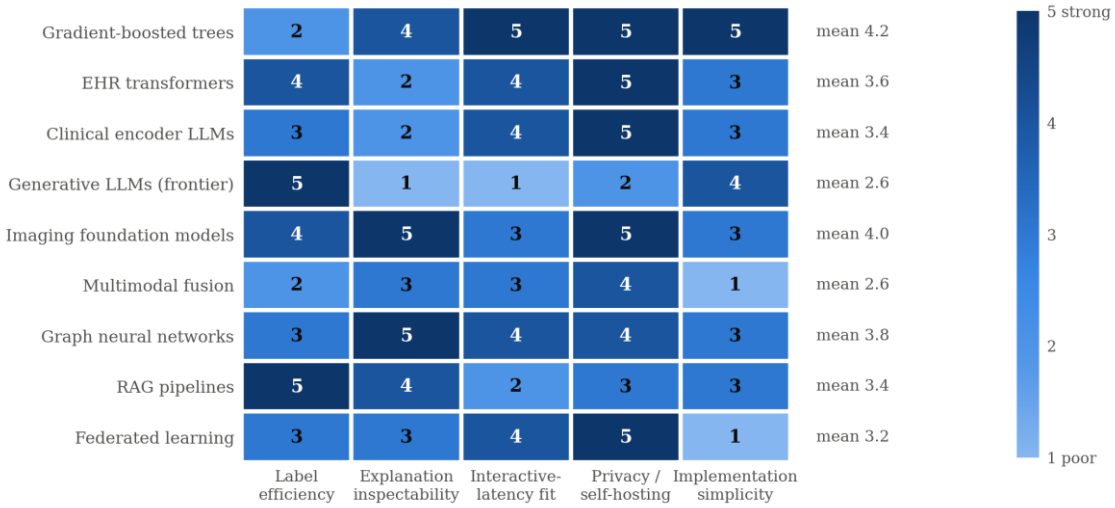


**Figure 4.** Suitability of nine model families for service inside a web representation layer, scored on five criteria. Scores are author-assigned ordinal values derived from Tables 4 and 5 using the rubric in Section 7.3; they are a structured summary of the review rather than empirical measurements. Numerals are printed in every cell so the matrix is readable without colour.

Ranked by mean score, gradient-boosted trees lead at 4.2, followed by imaging foundation models at 4.0 and graph neural networks at 3.8. Frontier generative LLMs and multimodal fusion tie for last at 2.6. The ordering deliberately inverts the ranking that MedQA accuracy would produce, and that inversion is the point: the family with the highest reported benchmark accuracy scores lowest on interactive-latency fit (1) and second-lowest on privacy posture (2), while the family with no representation-learning capability at all scores 5 on three of five criteria. Criterion-level means tell a similar story. Privacy and self-hosting averages 4.2 across families and interactive-latency fit averages 3.3, implementation simplicity is the weakest criterion overall at 2.9, and explanation inspectability follows at 3.2. Generative LLMs score lowest of any family on explanation inspectability, at 1.

### 7.4 Distribution of the cited evidence across the six layers

Figure 5 audits this paper's own reference list. Each of the 60 works was assigned to the single architectural layer whose function it primarily informs, using a fixed rule: dataset descriptors to L1, terminology and exchange standards to L2, representation-learning methods to L3, predictive or generative models and their evaluations to L4, work on presentation, interpretation and disclosure to L5, and work on governance, regulation, privacy, bias and reporting standards to L6.

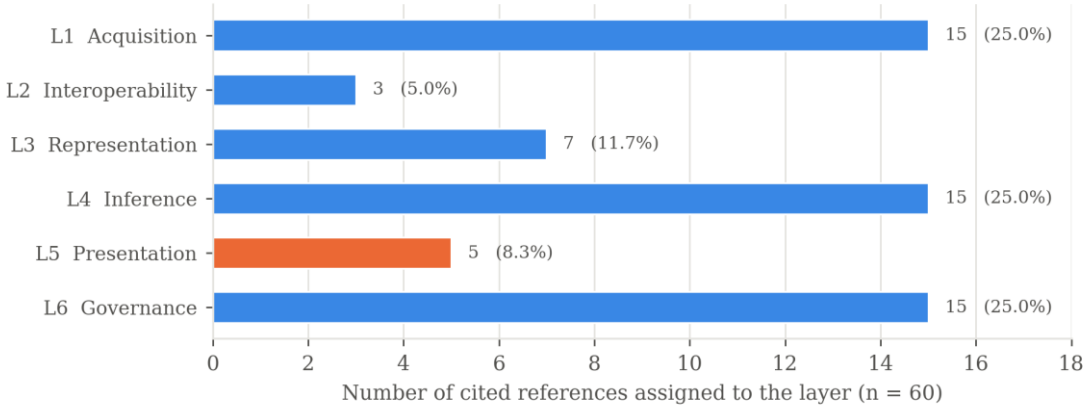


**Figure 5.** Distribution of the 60 cited works across the six architectural layers, one primary assignment per reference. The presentation layer is highlighted. The low count for interoperability (L2) reflects a mature standards base rather than an evidence gap; the low count for presentation does not.

The result is uneven in a way that supports the paper's central claim. Acquisition, inference and governance account for 15 references each, 25% apiece and 75% of the list between them. Representation draws 7 references (11.7%), interoperability 3 (5.0%), and presentation 5 (8.3%). The interoperability figure is not a gap: FHIR, SNOMED CT, LOINC and OMOP are mature standards, and citing three works is sufficient because the layer is governed by specification rather than by research. The presentation figure is a gap. Five works, of which one is a scoping review reporting only 18 eligible studies from 1,644 screened [36] and one is a W3C recommendation [37], is a thin basis for the layer at which every one of these systems meets its user. Put differently, the inference layer commands three times the cited evidence of the presentation layer, although a failure at either produces the same clinical outcome.

### 7.5 Reference integrity check

Because a review's value rests on the accuracy of what it cites, every reference in this paper was checked against a primary bibliographic record: the publisher's page, Crossref, PubMed, DBLP or the official conference proceedings. Table 6 reports the outcome.

**Table 6.** Outcome of the citation integrity check applied to all 60 references.

| Verification outcome | Count | Notes |
|---|---|---|
| Confirmed as originally written | 46 | Author list, title, venue, year and identifiers all matched the primary record. |
| Corrected | 12 | The work exists but its metadata was incomplete or wrong. Corrections included missing volume, issue and page ranges; truncated author lists; a preprint superseded by a journal version (Med-PaLM 2, now Nature Medicine 2025 [19]); a misspelled first-author initial [31]; and a publication year off by one [32]. |
| Replaced | 1 | One entry cited a National Center for Health Statistics report under a title that does not exist. It was replaced with the actual NHANES sample design and analytic guidelines reports in the Vital and Health Statistics series [13]. |

| Verification outcome | Count | Notes |
|---|---|---|
| Re-specified | 1 | The FDA AI-enabled medical device list is a live resource whose version claim could not be pinned to a specific release. It is now cited with an access date and the entry cut-off actually observed [38]. |
| Fabricated (no such work) | 0 | No reference described a publication that does not exist. |

*Counts sum to 60. As proportions: 76.7% confirmed unchanged, 20.0% corrected, 1.7% replaced and 1.7% re-specified. Verification was carried out in August 2026 against publisher pages, Crossref, PubMed, DBLP and official conference proceedings.*

We report this check as a result rather than in a methods footnote for two reasons. Reviews are increasingly assembled with machine assistance, and plausible-looking citations with wrong or invented metadata are a known failure mode of that workflow, so the check is now part of the work rather than an optional courtesy. And the corrections were not cosmetic: citing a superseded preprint where a peer-reviewed journal version exists misrepresents the evidential status of a claim, and a citation whose title does not exist cannot be followed by a reader at all.

## 8. AN EVALUATION PROTOCOL FOR SMART HEALTH WEB REPRESENTATION SYSTEMS

A SHWRS is a socio-technical system rather than a model, so evaluating only the model evaluates the wrong object. We propose a four-tier protocol in which each tier gates the next.

**Table 7.** Four-tier evaluation protocol.

| Tier | Object of evaluation | Metrics | Data required | Gate condition |
|---|---|---|---|---|
| T1 Model | Discrimination, calibration, robustness | AUROC, AUPRC, DSC, F1, Brier score, expected calibration error, subgroup performance gaps | Held-out split from the training distribution | Calibration acceptable and no subgroup gap beyond a pre-specified threshold |
| T2 Generalisation | Behaviour under distribution shift | External-site performance retention, temporal validation, shift detection sensitivity | Independent site and independent time period | Performance retention above a pre-registered floor at an external site |
| T3 Representation | Whether the display conveys the model's actual claim | Task success rate, time-to-comprehension, over-reliance and under-reliance rate, WCAG 2.2 AA conformance | Prospective study with intended users | Users correctly interpret the output, including its uncertainty, at a pre-specified rate |
| T4 System | Effect in workflow and over time | Clinical or behavioural outcome, alert burden, drift alarm precision, incident traceability | Prospective deployment with monitoring | Demonstrated benefit with no unacceptable workflow cost; DECIDE-AI reporting satisfied |

Tier 3 is where the field's evidence is thinnest, as Figure 5 shows and as the dashboard design literature confirms [36]. It is also the tier at which a system most commonly fails invisibly to its developers, because a correct and well-calibrated output that is systematically misread produces exactly the same logs as one that is understood. We recommend treating T3 measurement as mandatory rather than optional for any patient-facing representation, and measuring both over-reliance, accepting an incorrect model output, and under-reliance, rejecting a correct one, since interventions that reduce one often increase the other.

Reporting should follow the community standard matching the study type: TRIPOD+AI for prediction model development and validation [51], DECIDE-AI for early-stage live clinical evaluation of decision support [53], and CONSORT-AI for

randomised trials of AI interventions [52]. Where a system falls under the EU AI Act's high-risk classification, the technical documentation, logging and post-market monitoring obligations should map onto these tiers rather than exist as a parallel compliance artefact [39].

## 9. OPEN CHALLENGES AND RESEARCH DIRECTIONS

**C1. Representation drift monitoring.** Detecting that a learned representation has drifted is much harder than detecting an input distribution shift, because the representation is unlabelled and high-dimensional. Practical embedding-space drift detectors with calibrated false-alarm rates are a prerequisite for safe long-lived deployment [56].

**C2. Site-invariant patient encoding.** Methods that provably suppress site identity in a patient representation while preserving physiological signal would address a main cause of external validation failure [57]. Adversarial and invariant-risk approaches exist but remain unvalidated at clinical scale.

**C3. Evidence for the presentation layer.** Only five of the 60 works surveyed here address presentation (Figure 5), and the one scoping review available found most health dashboards built without structured design guidance [36]. Controlled studies of uncertainty display, provenance presentation and progressive disclosure in health contexts are needed.

**C4. Accessible health data visualisation.** WCAG 2.2 specifies the floor [37] without answering how to make a survival curve, a risk gauge or a longitudinal laboratory trend comprehensible without vision or without colour discrimination. Health-specific accessible visualisation patterns remain largely unstandardised.

**C5. Grounding at interactive latency.** Retrieval-augmented generation improves auditability [40,41] at the cost of a retrieval hop on top of an already slow generation step. Architectures delivering grounded output within one interaction cycle, whether through aggressive caching, precomputation of likely queries, or small grounded models, remain an open engineering problem.

**C6. Federated evaluation.** Federated training is comparatively mature [29,30]; federated evaluation is not. No party holds a validation set representative of the full training distribution, so performance claims for federated models rest on weaker evidence than for centralised ones [31,32].

**C7. Benchmark contamination and renewal.** MedQA and similar benchmarks predate web-scale pretraining corpora that plausibly contain them, and Figure 2 shows accuracy saturating near the ceiling. Held-out, continuously refreshed, real-query benchmarks of the kind used in the 2026 Nature Medicine evaluation [23] are far more informative and far more expensive; making their construction routine is an unsolved incentive problem.

**C8. Cohort representativeness.** The demographic composition of All of Us [11,12] shows representative recruitment is achievable at scale. Extending comparable representativeness to imaging and continuous-physiology corpora, still concentrated in a handful of academic centres, is the corresponding unmet need.

**C9. Currency of the public corpus.** The lag documented in Figure 3, reaching 13 years for the oldest heavily used resource, has no technical solution within the existing corpus. Rolling-release clinical corpora with automated re-de-identification, or federated access to contemporary institutional data, are the plausible routes; neither exists at the scale the field relies on today.

**C10. Provenance display standards.** Model Facts labels [54] are a strong proposal without a machine-readable serialisation. A FHIR-compatible provenance resource for model outputs, renderable directly by a representation layer, would make transparency a system property rather than a documentation practice.

**C11. Calibrated abstention.** Every family reviewed here can be made to produce an answer for any input. Few can reliably decline. Selective prediction with distribution-free guarantees, surfaced honestly at the presentation layer as an explicit "insufficient basis" state, may be the highest-value unsolved problem for clinical deployment.

## 10. CONCLUSION

Smart web-based representation of human health data is best understood as a six-layer systems problem in which the modelling layer, dominant though it is in the literature, is neither the only nor necessarily the binding constraint. The public dataset ecosystem is now substantial and, for resources such as All of Us and MIMIC-IV, both large and well documented. It is also geographically concentrated, lagging the present by a median of six and a half years and by nine years or more for half of the quantified resources, and governed by access regimes that materially shape which architectures are possible.

On the model side, three years of rapid progress have produced systems answering medical examination questions at rates that would have seemed implausible in 2022. The more instructive recent evidence is that this progress does not transfer straightforwardly. In independent blinded evaluation on real physician queries, specialised clinical AI tools performed comparably to a general web search summary, while frontier general-purpose models outperformed both [23,24]. The lesson is not that specialisation is worthless but that developer-reported benchmark performance predicts deployed usefulness poorly, and that independent evaluation belongs inside the development pipeline rather than appended to it.

Our own quantitative synthesis reinforces the point from a different direction. The suitability ranking in Figure 4 inverts the accuracy ranking in Figure 2, and the evidence audit in Figure 5 shows the inference layer commanding three times the cited literature of the presentation layer, although a failure at either reaches the patient identically.

Three recommendations follow for practitioners building such systems today. Architect the inference layer in tiers, keeping fast self-hosted models on the synchronous interaction path and reserving expensive generative models for asynchronous tasks where latency is visible and acceptable. Prefer model families whose outputs arrive with an inspectable artefact, a citation, an explanatory path or an editable mask, over those requiring post-hoc attribution. And evaluate the presentation layer with the rigour applied to the model, because a correct output that is systematically misread is, from the patient's perspective, indistinguishable from an incorrect one.

**Declarations**

**Funding.** No external funding was received for this work.

**Conflicts of interest.** The authors declare no competing interests.

**Data availability.** All datasets discussed are publicly available under the access conditions stated in Table 2; no new data were generated. The values plotted in Figures 2 to 5 are reproduced in full in Tables 2, 3 and 6 and in Section 7.3.

**Ethics approval.** Not applicable. This work involves no human participants and no analysis of individual-level data.

**Author contributions.** All authors contributed to the conception of the review, the selection and appraisal of sources, and the drafting and revision of the manuscript. All authors read and approved the final version.

**REFERENCES**

[1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates; 2017:5998-6008. arXiv:1706.03762

[2] Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023;10:1. doi:10.1038/s41597-022-01899-x

[3] Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, Celi LA, Mark R. MIMIC-IV (version 3.1). PhysioNet; 11 October 2024. doi:10.13026/kpb9-mt58

[4] Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data. 2018;5:180178. doi:10.1038/sdata.2018.178

[5] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):e215-e220. doi:10.1161/01.CIR.101.23.e215

[6] Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data. 2019;6:317. doi:10.1038/s41597-019-0322-0

[7] Johnson AEW, Pollard TJ, Greenbaum NR, Lungren MP, Deng C, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv:1901.07042; 2019. doi:10.48550/arXiv.1901.07042

[8] Irvin J, Rajpurkar P, Ko M, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33(1):590-597. doi:10.1609/aaai.v33i01.3301590

[9] Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z

[10] The UK Biobank Whole-Genome Sequencing Consortium. Whole-genome sequencing of 490,640 UK Biobank participants. Nature. 2025;645(8081):692-701. doi:10.1038/s41586-025-09272-9

[11] All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. Nature. 2024;627(8003):340-346. doi:10.1038/s41586-023-06957-x

[12] National Institutes of Health. NIH's All of Us Research Program is now the largest integrated genomics and health database in the world. NIH News Releases; 30 June 2026.

[13] Chen TC, Clark J, Riddles MK, Mohadjer LK, Fakhouri THI. National Health and Nutrition Examination Survey, 2015-2018: sample design and estimation procedures. Vital and Health Statistics, Series 2. 2020;(184):1-35. See also Johnson CL, Paulose-Ram R, Ogden CL, et al. National Health and Nutrition Examination Survey: analytic guidelines, 1999-2010. Vital and Health Statistics, Series 2. 2013;(161):1-24.

[14] Weinstein JN, Collisson EA, Mills GB, et al; The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics. 2013;45(10):1113-1120. doi:10.1038/ng.2764

[15] Jack CR Jr, Bernstein MA, Fox NC, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging. 2008;27(4):685-691. doi:10.1002/jmri.21049

[16] Wornow M, Thapa R, Steinberg E, Fries JA, Shah NH. EHRSHOT: an EHR benchmark for few-shot evaluation of foundation models. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track; 2023. arXiv:2307.02028

[17] Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. npj Digital Medicine. 2022;5:194. doi:10.1038/s41746-022-00742-2

[18] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

[19] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, et al. Toward expert-level medical question answering with large language models. Nature Medicine. 2025;31(3):943-950. doi:10.1038/s41591-024-03423-7 (preprint: arXiv:2305.09617, 2023)

[20] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375; 2023. doi:10.48550/arXiv.2303.13375

[21] Nori H, Usuyama N, King N, McKinney SM, Fernandes X, Zhang S, Horvitz E. From Medprompt to o1: exploration of run-time strategies for medical challenge problems and beyond. arXiv:2411.03590; 2024.

[22] Saab K, Tu T, Weng WH, Tanno R, Stutz D, Wulczyn E, Zhang F, Strother T, et al. Capabilities of Gemini models in medicine. arXiv:2404.18416; 2024.

[23] Vishwanath K, Alyakin A, Ghosh M, Hage A, Neifert SN, Orillac C, Mandelberg NJ, Khan HA, Lee JV, Yao JJ, Small WR, Varma A, Hewitt DB, Aphinyanaphongs Y, Alber DA, Oermann EK. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. Nature Medicine. 2026;32(7):2405-2409. doi:10.1038/s41591-026-04431-5

[24] General-purpose chatbots outperform clinical AI tools on physicians' real-world questions [Research Briefing]. Nature Medicine. 2026;32(7):2364-2365. doi:10.1038/s41591-026-04457-9

[25] Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nature Communications. 2024;15:654. doi:10.1038/s41467-024-44824-z

[26] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollar P, Girshick R. Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023:4015-4026.

[27] Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. npj Digital Medicine. 2020;3:136. doi:10.1038/s41746-020-00341-z

[28] Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, Wang F, Cheng F, Luo Y. Multimodal machine learning in precision health: a scoping review. npj Digital Medicine. 2022;5:171. doi:10.1038/s41746-022-00712-8

[29] Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, Bakas S, Galtier MN, Landman BA, Maier-Hein K, Ourselin S, Sheller M, Summers RM, Trask A, Xu D, Baust M, Cardoso MJ. The future of digital health with federated learning. npj Digital Medicine. 2020;3:119. doi:10.1038/s41746-020-00323-1

[30] Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, Milchenko M, Xu W, Marcus D, Colen RR, Bakas S. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific Reports. 2020;10:12598. doi:10.1038/s41598-020-69250-1

[31] Pechetti US, Dirisala JNK. Privacy-preserving federated learning in medical imaging: a systematic review. WIREs Data Mining and Knowledge Discovery. 2026;16(3):e70109. doi:10.1002/widm.70109

[32] Ghosh D, Mehjabin M, Rayed ME, Mridha MF, Kabir MM. Advancements and challenges of federated learning in medical imaging: a systematic literature review. Artificial Intelligence Review. 2026;59(2):87. doi:10.1007/s10462-025-11489-z

[33] Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. Journal of the American Medical Informatics Association. 2016;23(5):899-908. doi:10.1093/jamia/ocv189

[34] HL7 International. HL7 FHIR Release 5 (R5), version 5.0.0. Health Level Seven International; 26 March 2023.

[35] Bender D, Sartipi K. HL7 FHIR: an agile and RESTful approach to healthcare information exchange. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS); 2013:326-331. doi:10.1109/CBMS.2013.6627810

[36] Vornhagen H, Barrett S, Carroll C, Kavochi Iladiva L, Martin G, McKeown D, Martin J. Design practices for data dashboards in health care: scoping review. Journal of Medical Internet Research. 2026;28(1):e77361. doi:10.2196/77361

[37] World Wide Web Consortium. Web Content Accessibility Guidelines (WCAG) 2.2. W3C Recommendation, 5 October 2023; updated W3C Recommendation, 12 December 2024.

[38] U.S. Food and Drug Administration. Artificial Intelligence-Enabled Medical Devices. Silver Spring, MD: FDA. Device list accessed 18 August 2026 (entries through 30 March 2026).

[39] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, OJ L, 2024/1689, 12 July 2024.

[40] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Kuttler H, Lewis M, Yih W, Rocktaschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Curran Associates; 2020:9459-9474.

[41] Gargari OK, Habibi G. Enhancing medical AI with retrieval-augmented generation: a mini narrative review. Digital Health. 2025;11:20552076251337177. doi:10.1177/20552076251337177

[42] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16); 2016:785-794. doi:10.1145/2939672.2939785

[43] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates; 2017:3146-3154.

[44] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates; 2017:4765-4774.

[45] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017:618-626. doi:10.1109/ICCV.2017.74

[46] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations (ICLR); 2017. arXiv:1609.02907

[47] Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Scientific Data. 2023;10:67. doi:10.1038/s41597-023-01960-3

[48] Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34(13):i457-i466. doi:10.1093/bioinformatics/bty294

[49] Perez MV, Mahaffey KW, Hedlin H, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. New England Journal of Medicine. 2019;381(20):1909-1917. doi:10.1056/NEJMoa1901183

[50] Torres-Soto J, Ashley EA. Multi-task deep learning for cardiac rhythm detection in wearable devices. npj Digital Medicine. 2020;3:116. doi:10.1038/s41746-020-00320-4

[51] 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. doi:10.1136/bmj-2023-078378

[52] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nature Medicine. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x

[53] 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. Nature Medicine. 2022;28(5):924-933. doi:10.1038/s41591-022-01772-9

[54] Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. npj Digital Medicine. 2020;3(1):41. doi:10.1038/s41746-020-0253-3

[55] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453. doi:10.1126/science.aax2342

[56] Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. New England Journal of Medicine. 2021;385(3):283-286. doi:10.1056/NEJMc2104626

[57] Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine. 2021;181(8):1065-1070. doi:10.1001/jamainternmed.2021.2626

[58] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3:160018. doi:10.1038/sdata.2016.18

[59] Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16); 2016:308-318. doi:10.1145/2976749.2978318

[60] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science. 2014;9(3-4):211-407. doi:10.1561/0400000042