

Review

From Algorithmic Performance to Hospital Value: A Narrative Review of Clinical Artificial Intelligence

Christina Morfaki

Department of Economic Science, University of Piraeus, 18534 Piraeus, Greece; morfakic@unipi.gr

Abstract

Clinical artificial intelligence (AI) has demonstrated performance across diagnostic, predictive, monitoring, and decision-support tasks, yet hospital adoption depends on whether that capability can produce sustainable value under local conditions. Informed by structured searches of PubMed/MEDLINE and Scopus, supplemented by Google Scholar and citation tracking (2015–July 2026), this narrative review synthesizes evidence on clinical AI methods, applications, implementation, evaluation, and sustainability in hospitals. Evidence remains uneven: image-based applications have substantial evidence for diagnostic performance, whereas prospective clinical utility, workflow effects, economic consequences, and organizational sustainability are evaluated less consistently. Established approaches—including efficacy hierarchies, the Radiology AI Deployment and Assessment Rubric (RADAR), hospital-based health technology assessment, implementation science, and AI governance frameworks—offer perspectives on performance, transferability, implementation, and value. Building on these traditions, this review proposes institutional efficacy as the warranted and revisable hospital-level judgment that an AI-enabled service configuration can generate and sustain patient-centered value under local priorities, constraints, and adaptive capacity. The review further proposes that deeper organizational integration increases the proportion of realized-value evidence requiring local generation or verification. Two additional properties—context and data dependence, and the consequence profile—separately shape that burden. External evidence remains essential, but hospital-specific appraisal is necessary to determine whether value can be realized and sustained.

Keywords: artificial intelligence; clinical decision support; hospital management; health technology assessment; institutional efficacy; implementation science; healthcare governance; health services administration



Academic Editor: Sandra C. Buttigieg

Received: 19 June 2026

Revised: 24 September 2026

Accepted: 24 September 2026

Published: 28 September 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)[Attribution \(CC BY\)](#) license.

1. Introduction

Clinical artificial intelligence (AI) has developed from a predominantly experimental field into a broad set of diagnostic, predictive, monitoring, and decision-support applications. Contemporary clinical AI draws on machine learning, deep learning, natural language processing, large language models, and multimodal approaches to analyze structured clinical records, medical images, biosignals, temporal data, and free-text documentation. These methods can transform heterogeneous information into classifications, risk estimates, alerts, summaries, or recommendations intended to support clinical judgment and care delivery [1,2]. In selected diagnostic tasks, particularly in medical imaging, AI systems have achieved performance comparable to that of clinical experts under defined evaluation conditions [3,4].

The available evidence, however, remains concentrated largely at the levels of technical performance and diagnostic accuracy. Metrics such as sensitivity, specificity, discrimination, calibration, and area under the receiver operating characteristic curve are essential for determining whether a model functions as intended under specified conditions. They provide much less information about whether its introduction changes professional decisions, improves patient outcomes, modifies workflow, reduces or redistributes workload, or produces sustainable value for the organization responsible for its use. Comparative studies have also frequently relied on retrospective and curated datasets, while prospective evaluation under routine clinical conditions remains less common [3,5]. Algorithmic capability therefore represents an indispensable starting point for evaluation, rather than a sufficient basis for hospital adoption.

This distinction is particularly important because a hospital is a sociotechnical organization rather than a neutral setting into which a validated model can be inserted unchanged. AI systems interact with local data-generating practices, information infrastructure, professional roles, staffing levels, clinical pathways, financing arrangements, and established structures of responsibility. Their effects depend on when and where an output enters the workflow, who receives it, whether the organization has the capacity to respond, and how performance and adverse consequences are monitored over time [5–7]. The same model may therefore perform differently, create a different workload, or generate different operational consequences across institutions. Evidence maturity is correspondingly specific to the system, task, population, and intended use rather than a general property of “AI”.

The contrast between selected image-based applications and applications that depend more extensively on institution-specific data and coordinated workflow illustrates this variability. Selected imaging tools benefit from clearly bounded tasks, established digital infrastructures, and comparatively well-defined reference standards, and have consequently accumulated a more extensive evidence base. Even in this domain, however, evidence remains weighted toward diagnostic accuracy, with fewer studies addressing effects on clinical decisions, patient outcomes, workflow, or socio-economic value [8,9]. Applications based on longitudinal electronic health records, real-time alerts, triage, language models, or multimodal decision support are often more dependent on institution-specific data and coordinated responses across multiple professional groups. Their value is therefore more sensitive to local integration, professional use, organizational capacity, and governance.

Recent foundation models and large language models further intensify this evaluative challenge. Their capacity to generate, retrieve, summarize, and combine information across clinical text, images, laboratory data, and other modalities creates opportunities for more integrated forms of decision support [10,11]. At the same time, these systems may produce fluent yet unsupported outputs, reproduce biases, and create errors whose significance depends heavily on the task, the degree of human mediation, and the consequences of reliance. Their evaluation requires task-specific evidence, transparent reporting, clearly defined human oversight, and monitoring under the conditions in which they are used [12]. The expansion of AI capability thus increases the importance of evaluating the institutional setting through which that capability is translated into action.

Several established traditions address the hospital-evaluation problem from complementary perspectives, including efficacy hierarchies, the local efficacy level of the Radiology AI Deployment and Assessment Rubric (RADAR), hospital-based health technology assessment, and implementation science [13–16]. Together, they provide approaches to technical and clinical performance, local transferability, organizational appraisal, implementation outcomes, and sustainability. The organizational level is therefore not unexplored. The contribution developed in this review is to formalize institutional efficacy as the warranted and revisable hospital-level judgment that a defined AI-enabled service configuration can

generate and sustain patient-centered value under the adopting hospital's actual conditions, priorities, and constraints. The review further proposes that deeper organizational integration increases the proportion of evidence concerning realized value that must be generated or verified locally, while the local evidence burden is separately shaped by two additional properties: context and data dependence, and the consequence profile of the application. Section 7 develops the conceptual genealogy, definition, and implications of these propositions.

The review has three objectives. First, it synthesizes the principal methodological capabilities of clinical AI and clarifies the kinds of evidentiary claims that they can support. Second, it examines clinical application domains, evidence profiles, explainability, human oversight, data integration, and institutional readiness. Third, it develops institutional efficacy and the integration-depth proposition as a conceptual basis for evaluating AI from the perspective of hospital governance and decision-making. The operational specification of indicators, measurement methods, weighting principles, and decision thresholds is reserved for subsequent framework development and empirical evaluation.

2. Approach and Scope of the Review

This article was designed as a narrative review informed by a structured literature search. Narrative reviews are appropriate when the objective is to integrate heterogeneous bodies of literature, clarify conceptual relationships, and develop an interpretive synthesis across a broad field [17–19]. To strengthen methodological transparency, the review process was informed by the Scale for the Assessment of Narrative Review Articles (SANRA), particularly its criteria concerning the statement of aims, description of the literature search, referencing, scientific reasoning, and presentation of relevant evidence [20]. SANRA was used to guide the conduct and reporting of the review rather than as a formal risk-of-bias instrument. A SANRA self-assessment of the revised manuscript is provided in Table S2 (Supplementary Materials).

The review examined the relationship between clinical artificial intelligence and hospital value. The principal literature search was conducted in PubMed/MEDLINE and Scopus, supplemented by backward citation tracking from key publications and authoritative reports. Google Scholar was used as a supplementary source to identify authoritative reports and recent publications rather than as a reproducible database search. The principal search period covered English-language publications from 2015 through July 2026. Earlier foundational publications were retained where necessary to define established methodological or evaluative concepts, including machine learning, reinforcement learning, causal inference, explainable artificial intelligence, health technology assessment, implementation science, and narrative-review methodology.

Search terms were organized into four thematic blocks. The first concerned AI methods and included terms such as “artificial intelligence”, “machine learning”, “deep learning”, “natural language processing”, “large language models”, “foundation models”, “multimodal AI”, “explainable AI”, “causal inference”, and “reinforcement learning”. The second concerned clinical application domains and included “clinical decision support”, “medical imaging”, “radiology”, “pathology”, “ophthalmology”, “endoscopy”, “electronic health records”, “risk prediction”, “precision medicine”, and “patient monitoring”. The third concerned evaluation and translation and included “algorithmic performance”, “diagnostic accuracy”, “external validation”, “generalizability”, “clinical utility”, “real-world implementation”, “workflow integration”, “interoperability”, and “cost-effectiveness”. The fourth concerned institutional and governance issues and included “hospital management”, “health services management”, “AI governance”, “accountability”, “patient safety”, “human oversight”, “procurement”, “auditability”, “post-deployment monitoring”, and

“regulation”. Terms from these blocks were combined iteratively according to the topic under examination rather than through a single fixed Boolean query.

Eligible sources included peer-reviewed empirical studies, systematic reviews and meta-analyses, methodological and conceptual papers, implementation studies, health-services and management literature, and authoritative policy or regulatory documents. Priority was given to sources addressing one or more dimensions directly relevant to the review question: algorithmic performance, external validation, clinical utility, prospective evaluation, real-world implementation, workflow effects, economic implications, governance, accountability, sustainability, or hospital decision-making. Sources were excluded when they concerned non-healthcare applications, lacked a clear connection to clinical or institutional use, duplicated evidence already represented by more comprehensive sources, or addressed purely technical model development without relevance to clinical translation or hospital-level evaluation.

Selection was purposive and iterative rather than based on a prespecified exhaustive screening protocol. Where several publications addressed the same topic, preference was given to primary empirical studies, systematic or authoritative syntheses, methodological contributions, and sources that added a distinct evaluative perspective. Apparently conflicting findings were retained where they reflected differences in study design, setting, intended use, comparator, or evaluated endpoint rather than being resolved through formal meta-analysis.

The synthesis followed an interpretive and integrative logic [21,22]. Evidence was organized around the central question of the review: under what conditions can clinical AI move from algorithmic performance to sustainable hospital value? Findings from AI methodology, clinical applications, validation studies, implementation research, health-services management, economic evaluation, and governance literature were therefore examined according to the evidentiary claims they support and the additional information required when an AI-enabled service is considered for adoption within a hospital.

The narrative design imposes corresponding limitations. The review does not claim exhaustive retrieval of all eligible studies, and purposive source selection may introduce selection effects. It did not apply a uniform formal risk-of-bias instrument across the heterogeneous evidence included, nor can the search process be reproduced as a systematic-review search history. These limitations are acknowledged explicitly and are considered when interpreting statements about evidence maturity and when distinguishing established findings from the conceptual propositions developed in this review.

3. Algorithmic Capabilities and Evidentiary Limits of Clinical AI

Clinical AI uses heterogeneous data—including structured clinical records, medical images, physiological signals, and clinical text—to support prediction, pattern recognition, information processing, and treatment planning [2,23]. Learning paradigms, model architectures, and application domains overlap rather than constitute mutually exclusive categories. The following sections distinguish their capabilities according to the outputs they produce and the evidence needed to interpret those outputs. This provides the methodological background for examining clinical applications and hospital-level evaluation without treating technical capability as a uniform indicator of readiness for adoption.

3.1. Machine Learning: From Data to Prediction

Machine learning (ML) develops models that learn patterns from data to support prediction, classification, or decision-making [2,24–26]. In supervised learning, models learn relationships between inputs and labeled targets, such as diagnoses or subsequent clinical outcomes. Their evidentiary meaning depends on the quality and clinical rele-

vance of those targets: diagnostic conventions, coding practices, and outcome definitions can influence what a model learns and what its reported performance represents [2,27]. Evaluation therefore requires scrutiny of the reference standard and validation population, together with discrimination, calibration where applicable, and subgroup performance. Strong results on development data do not establish reliable performance in a different care setting [5].

Unsupervised learning identifies structures in data without predefined outcome labels. Clustering can suggest patient subgroups, dimensionality reduction can simplify high-dimensional biomedical data, and anomaly detection can identify atypical observations [28–31]. These outputs are primarily exploratory: a statistically distinct cluster is not automatically a clinically meaningful disease subtype, and an unusual observation is not necessarily a pathological finding. Stability, reproducibility, and clinical relevance must therefore be examined before the findings inform patient stratification or monitoring [2,31].

Reinforcement learning (RL) addresses sequential decisions by learning a policy intended to maximize cumulative reward over time [32–34]. It has been investigated for treatment selection in sepsis and adaptive radiotherapy, where successive actions depend on an evolving patient state [35,36]. Such studies examine candidate treatment strategies; they do not by themselves establish the benefit of implementing those strategies in routine care. Evaluation must address whether the reward represents clinically appropriate objectives, whether observational treatment trajectories contain relevant confounding, and whether the proposed actions are adequately supported by the available data. These issues distinguish policy evaluation from predictive validation and require particular caution before learned recommendations influence treatment [33,37].

3.2. Learning with Limited or Unlabeled Data

Self-supervised and semi-supervised learning address the limited availability of expert-labeled clinical data. Self-supervised learning derives supervisory signals from the data themselves, using tasks such as predicting masked content or learning relationships between different views of a sample [38]. The resulting representations can subsequently be adapted to specific clinical tasks using labeled examples. In medical imaging, this approach has been explored to improve classification when annotations are scarce, including through masked-autoencoder pre-training on chest radiographs [39,40].

Semi-supervised learning combines labeled and unlabeled examples during training. Methods include pseudo-labeling, which assigns provisional model-generated labels; consistency regularization, which encourages stable predictions under suitable input perturbations; and graph-based learning, which exploits relationships between samples [41,42]. Applications include chest-radiograph classification under limited supervision and colorectal tissue classification using a hypergraph neural network [41,43].

These approaches can reduce annotation requirements, but their contribution must be distinguished from evidence of clinical reliability. Learned representations and provisional labels remain sensitive to the quality and composition of the source data, while the usefulness of unlabeled examples depends on their relationship to the target task [39,42]. For hospital evaluation, the relevant question is whether gains achieved with fewer annotations persist on independently assessed, clinically representative data. Reduced labeling requirements during model development do not remove the need for reliable reference standards when evaluating the resulting predictions.

3.3. Deep Learning and the Analysis of Unstructured Clinical Data

Deep learning (DL), a subfield of machine learning, uses multilayered neural networks to learn hierarchical representations while reducing reliance on manually engineered

features [44]. In healthcare, it supports analysis of medical images, physiological signals, longitudinal clinical records, and text for diagnostic and prognostic tasks [23,45].

Convolutional neural networks extract spatial features from magnetic resonance imaging (MRI) and computed tomography (CT) scans, retinal images, and digitized histopathological slides [46,47]. Recurrent neural networks, including long short-term memory architectures, model temporal dependencies in physiological signals and sequential measurements [48,49]. Transformers use attention mechanisms to capture relationships among input elements and underpin many language and multimodal systems [10,50].

Generative approaches extend these capabilities to data synthesis and reconstruction. They include generative adversarial networks and variational autoencoders [51,52], as well as diffusion and related score-based models [53,54]. Applications of adversarial models in medical imaging include data augmentation, denoising, reconstruction, and image synthesis [55]. Their evaluation requires attention to the clinical fidelity of the output, not only its visual quality: synthetic or reconstructed images may introduce artifacts or distortions. Data generated for model training and images intended for direct clinical interpretation therefore require validation appropriate to their respective uses [55].

For hospital use, evaluation must also address the resources required for data curation, expert annotation where needed, computational support, maintenance, and monitoring [2,5]. Changes in acquisition equipment, protocols, patient populations, or clinical workflows can affect model performance and should inform local validation and reassessment [5].

3.4. Natural Language Processing, Large Language Models, and Multimodal AI

Natural language processing (NLP) extracts and organizes information from clinical notes, discharge summaries, and diagnostic reports, supporting coding, concept and relation extraction, and document classification [56]. Transformer architectures underpin biomedical and clinical language models such as BioBERT and ClinicalBERT, which learn representations that can be adapted to specific language-processing tasks [50,57,58].

Large language models (LLMs) support different forms of language processing. GatorTron was evaluated on clinical language-understanding tasks [59], whereas generative models such as Med-PaLM and GPT-4 have been evaluated on medical question-answering benchmarks [11,60]. These results do not by themselves demonstrate benefit in routine clinical use. Generated summaries and answers require assessment of factual errors, clinically important omissions, bias, and potential harm [11,61], supported by task-specific reporting and clearly defined human oversight [12].

Retrieval-augmented generation (RAG) conditions generated responses on information retrieved from external sources [62]. Clinical guidelines, scientific literature, and institutional protocols can provide grounding, but reliability still depends on source quality and currency, retrieval relevance, and the model's use of retrieved content [63]. Multimodal foundation models may combine clinical text with images, laboratory measurements, and other data, broadening decision support while introducing validation and data-governance requirements across modalities [10].

For hospital evaluation, these capabilities must be assessed in relation to the intended task and the consequences of reliance. Coding, retrieval, and documentation support should be distinguished from uses that directly influence diagnosis or treatment, without assuming that the former are risk-free. Assessment should address output quality, source traceability where applicable, the workload of human verification, workflow integration, responsibility for acting on outputs, and monitoring under routine conditions [5,12,56]. Table S1 (Supplementary Materials) summarizes the principal model families and their properties.

3.5. Causal Inference and Causal Machine Learning in Healthcare: From Prediction to Intervention Effects

Causal inference provides methods for estimating the effects of interventions rather than merely predicting outcomes from observed associations [64–66]. When combined with flexible machine-learning methods, it can support estimation of average treatment effects and variation in effects across patient profiles [67–69]. A predictive model may identify a patient at increased risk of deterioration without establishing which intervention would reduce that risk [67,68].

For each patient, only the outcome under the treatment received is observed; the corresponding outcome under an alternative treatment remains counterfactual [64]. Causal interpretation therefore depends on study design, explicit assumptions, clinically meaningful definitions of treatments and outcomes, and appropriate handling of confounding [65,66]. Flexible models do not remove these requirements, and estimates from observational data remain vulnerable to unmeasured confounding and other sources of bias [67,69].

For hospital evaluation, this distinction matters when assessing whether a change in treatment, staffing, or service organization improves outcomes or resource use. Such questions require randomized or otherwise credible causal designs; causal machine learning may contribute by estimating heterogeneous intervention effects [66,68,69]. An estimated treatment benefit, however, is not equivalent to evidence that deploying an AI system to recommend that treatment will improve care. Evaluation must also examine how clinicians use the recommendations, whether the service can act on them, and what implementation demands arise under routine conditions [5].

4. Clinical Application Domains and Evidence Maturity

Clinical AI applications differ in their intended tasks, deployment arrangements, and supporting evidence [2]. This section examines selected applications through the relationship between study design, evaluated endpoints, and implications for care delivery. Diagnostic performance, effects on clinical decisions, patient outcomes, and organizational consequences are distinguished rather than treated as interchangeable indicators of readiness for adoption [3,5].

4.1. Image-Based Diagnosis and Screening

Mammography illustrates the progression from algorithmic comparison to prospective evaluation of an AI-supported service. McKinney et al. [70] evaluated diagnostic performance retrospectively using screening datasets and reader comparisons. The randomized Mammography Screening with Artificial Intelligence (MASAI) trial subsequently compared AI-supported screen reading with standard double reading. Its interim analysis reported reduced screen-reading workload [71], while the primary-outcome analysis demonstrated noninferiority for the interval-cancer rate and higher screening sensitivity [72]. These findings concern a defined screening strategy, rather than mammography AI as an undifferentiated technology category.

Ophthalmic applications likewise require distinction by task and mode of use. Retinal-photograph classification for diabetic retinopathy [73] and diagnosis and referral based on optical coherence tomography [74] address different clinical questions. A prospective pivotal trial of an autonomous diabetic-retinopathy screening system in primary care evaluated sensitivity, specificity, and imageability against an independent reading-center reference standard [75]. This provides evidence of diagnostic performance in the tested setting, not a direct demonstration of improved visual outcomes. Referral completion and access to subsequent assessment and treatment remain separate service-level questions.

In digital pathology, whole-slide analysis has demonstrated strong discrimination for selected tissue-classification tasks and supports the investigation of histological biomarkers [76,77]. Implementation also depends on slide digitization, image quality, scanner and staining variability, storage, and integration into pathology workflows [77,78]. Campanella et al. [76], for example, estimated that algorithmic prescreening could reduce the number of slides requiring review. Such estimates should be distinguished from prospectively observed reductions in reporting time or staffing requirements.

Earlier systematic reviews identified substantial variation in reference standards, external validation, reporting quality, and the realism of algorithm–clinician comparisons [3,4]. The prospective examples above demonstrate progress beyond those limitations without establishing equivalent evidence across products or specialties. Hospital appraisal should therefore identify the specific task, comparator, and demonstrated endpoint, then examine the additional effects on diagnostic safety, reporting or referral pathways, workload, and local costs [5].

4.2. Prediction and Monitoring from Longitudinal Clinical Data

Electronic health records (EHRs) support prediction of deterioration, complications, readmission, and length of stay using diagnoses, laboratory results, medication histories, and physiological measurements [2,79]. Missing data, differences in recording practices, and limited validation across institutions can complicate the interpretation and transferability of these predictions [79].

Physiological monitoring provides a related but distinct example. Hannun et al. [80] evaluated a deep-learning model for classifying cardiac rhythms from single-lead ambulatory electrocardiograms against cardiologist reference labels. This established performance in rhythm classification, rather than demonstrating that deploying the model prevents acute events or improves patient outcomes.

Sepsis prediction illustrates the importance of evaluating both discrimination and potential workload. In a retrospective external validation of the Epic Sepsis Model, Wong et al. [81] reported poor calibration and a hospitalization-level area under the receiver operating characteristic curve of 0.63. At the evaluated threshold, the model would have flagged 18% of hospitalizations while failing to identify 67% of sepsis cases. These estimates describe potential alert burden, not prospectively observed effects on clinician workload or patient outcomes. Lyons et al. [82] subsequently found substantial variation in discrimination across nine networked hospitals, with scores calculated without being displayed to clinicians. These findings concern the evaluated model versions and settings and should not be generalized to all sepsis-alert systems.

Prospective clinical-impact evidence is available for selected deterioration-support systems. A one-year pragmatic trial randomized 74 clinical units across two health systems to usual care or care informed by a machine-learning early warning system using nursing documentation patterns. Across 60,893 hospital encounters, the intervention group had a lower adjusted hazard of in-hospital death (hazard ratio, 0.64; 95% confidence interval, 0.53–0.78) and a higher adjusted hazard of unanticipated intensive-care transfer (hazard ratio, 1.25; 95% confidence interval, 1.09–1.43) [83]. This evaluates an implemented decision-support strategy rather than isolated predictive accuracy. For hospital appraisal, the combination of improved survival and increased escalation of care makes downstream capacity and resource use relevant additional questions; it does not automatically establish financial savings.

4.3. Real-Time and Workflow-Embedded Decision Support

Computer-aided detection (CADe) during colonoscopy allows evaluation of endoscopist performance with versus without real-time AI assistance. A meta-analysis of 21 randomized trials involving 18,232 patients reported a higher adenoma detection rate with CADe than with standard colonoscopy (44.0% versus 35.9%; relative risk, 1.24; 95% confidence interval, 1.16–1.33), although the certainty of evidence for this outcome was rated low. CADe also increased removal of nonneoplastic polyps and modestly prolonged inspection time. The review did not establish effects on colorectal-cancer incidence or cancer-related mortality [84].

A subsequent pragmatic randomized trial involving 1627 patients across 12 German institutions found no statistically significant improvement in adenoma detection with CADe: examiner-adjusted rates were 40.6% with assistance and 38.3% without assistance. Detection of hyperplastic polyps was higher in the CADe group [85]. This result illustrates that benefits observed across pooled trials are not uniformly reproduced in individual clinical settings; it does not establish that CADe is ineffective across all systems or populations.

For an endoscopy service, the evaluative question therefore extends to what additional detection changes in care delivery. Relevant measures include the detection of advanced lesions, nonneoplastic resections, procedure duration, adverse events, pathology workload, and incremental cost per examination. Additional resections may generate downstream activity rather than release resources, while potential reductions in future cancer burden require longer-term evidence. These distinctions translate the broader implementation requirement—examining changes in clinical work and resource use alongside diagnostic performance—into application-specific hospital outcomes [5,6].

4.4. Uneven Evidence Maturity and Implications for Hospital Adoption

Evidence maturity is considered here for a specific system, task, population, and intended use. It is described through the evidence available for external validity, prospective performance in care, comparative clinical impact, workflow and resource effects, and sustained use. Interpretation also considers reported methodological limitations, comparator relevance, and follow-up duration [3–5]. These considerations provide a descriptive evidence profile rather than a numerical rating or a uniform maturity label for an entire specialty.

A recurring limitation identified in earlier clinical-AI research is the mismatch between performance-centered endpoints and claims about improved care. Retrospective testing on curated datasets and isolated clinician-comparison studies may not reproduce routine case selection, access to clinical information, or competing workflow demands [3,4]. For treatment-support systems evaluated through offline reinforcement learning, estimates of policy benefit also depend on assumptions that require scrutiny before clinical use [33,37]. Such designs contribute to development and validation but leave the consequences of implementation incompletely examined.

The prospective studies discussed above address parts of this gap, while longer-term organizational questions require separate evaluation. A short pilot may leave maintenance costs, performance following model changes, and effects under routine staffing unexamined. These outcomes cannot be inferred from initial accuracy or uptake alone [5,6]. Reviews of commercial radiology products likewise document more extensive evidence of diagnostic performance than of clinical or socioeconomic consequences [8,9]. Regulatory status should be interpreted within the applicable approval pathway rather than treated as a uniform indicator of clinical or organizational evidence [86].

For hospital appraisal, outcomes that have not been evaluated remain unresolved rather than demonstrably favorable or unfavorable. The reviewed findings provide a

literature-informed rationale for the discussion of institutional efficacy in Section 7; they do not establish a general relationship between integration depth and local evidence requirements. Table 1 summarizes application-specific evidence alongside the additional outcomes and local assessments relevant to hospital use.

Table 1. Selected clinical AI applications, demonstrated evidence, and hospital-level evaluation needs.

Application/Use Case	What Published Evidence Principally Demonstrates	Outcomes Not Established by Performance Evidence Alone	Principal Local Evaluation Needs
AI-supported breast-cancer screening	Retrospective diagnostic performance and reader comparisons [70]; prospective randomized evidence on AI-supported screening, workload, sensitivity, and interval-cancer outcomes [71,72]	Long-term service effects, downstream resource use, implementation and maintenance costs, and sustainability across screening settings	Workflow compatibility; reader workload; recall and referral pathways; equipment and population fit; downstream capacity and budget impact
Ophthalmic screening and referral	Diagnostic classification and referral performance [73,74], including prospective evaluation of autonomous diabetic-retinopathy screening [75]	Referral completion, access to treatment, effects on visual outcomes, service capacity, and long-term resource implications	Imageability and population performance; referral completion; specialist follow-up capacity; integration with primary-care or ophthalmology pathways
Digital pathology	Whole-slide classification and biomarker-related performance in selected tasks [76,77]	Routine reporting-time effects, staffing consequences, economic sustainability, and system-wide workflow benefit	Digitization and quality control; scanner and staining variability; storage and computational capacity; workflow integration; pathology workload and maintenance [77,78]
EHR-based deterioration and sepsis support	Predictive discrimination, calibration, and cross-site variability [79,81,82]; prospective evidence of clinical impact for selected early-warning strategies [83]	Transferability of benefit across hospitals, alert burden, downstream escalation of care, resource use, and sustained effects over time	Local calibration and subgroup performance; alert frequency; response capacity; escalation pathways; workload; ICU and downstream service capacity
Real-time endoscopy decision support	Randomized and pooled evidence on adenoma detection and related procedural outcomes [84]; variable effectiveness across pragmatic settings [85]	Effects on cancer incidence and mortality, downstream pathology burden, incremental cost, and long-term service value	Lesion mix and examiner performance; procedure duration; additional resections; pathology workload; adverse events; incremental cost per examination
Clinical NLP and LLM-supported information use	Performance in information extraction, question answering, summarization, retrieval, and selected language tasks [11,56,59–61,63]	Reliable clinical utility under routine use, effects on decisions and patient outcomes, verification burden, organizational sustainability, and consequences of model updates	Factuality and omission rates; source traceability; human-verification workload; privacy and data governance; workflow integration; responsibility; version and performance monitoring

Note. Evidence profiles are application- and use-case-specific and should not be interpreted as maturity ratings for an entire specialty or model class. The table distinguishes demonstrated endpoints from additional outcomes that require local or prospective evaluation. Abbreviations: AI, artificial intelligence; EHR, electronic health record; ICU, intensive care unit; LLM, large language model; NLP, natural language processing.

5. Explainability, Human Oversight, and Institutional Accountability

Explainability refers to methods and system properties that make the basis of an AI output more intelligible to relevant users. In clinical settings, its significance extends beyond technical transparency to the ability of health professionals to scrutinize algorithmic outputs, assess their plausibility, communicate relevant uncertainty, and retain responsibility for decisions. The form and level of explanation required depend on the intended use, the associated risk, and the stakeholder concerned: developers, clinicians, patients, regulators, and hospital governance bodies require different kinds of information [87,88].

Explainability methods are commonly distinguished between post hoc approaches applied to an already trained model and intrinsically interpretable models whose operation is designed to be more transparent. SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) estimate the contribution of features

to individual predictions, while gradient-weighted class activation mapping (Grad-CAM) highlights image regions that influence the output of convolutional neural networks [89–91]. These techniques can support model inspection and help identify implausible dependencies, but their outputs do not necessarily provide a faithful account of the model's decision function and do not establish causal relationships. In high-stakes settings, post hoc explanation should therefore not be treated as a substitute for clinical validation or, where comparable alternatives are available, for the use of inherently interpretable models [92,93].

The appropriate objective is not to maximize trust in an AI system, but to support appropriately calibrated reliance. Human oversight must therefore be designed as an operational process rather than assumed from the presence of a clinician who formally retains the final decision. Meaningful oversight requires clarity regarding who reviews the output, what supporting information is available, under which conditions the recommendation may be challenged or overridden, how disagreement is documented, and where responsibility lies when the system contributes to harm. Explanations should also be understandable in relation to the clinical decision at issue and, where relevant, support communication with patients without transferring professional responsibility to the algorithm [88,94].

From the perspective of hospital adoption, explainability contributes to accountability but does not establish it on its own. A plausible explanation cannot compensate for inadequate external validation, poor calibration, dataset shift, or the absence of post-deployment monitoring. Institutional evaluation should therefore distinguish among model-level explanation, system-level transparency—including the intended use, validation population, known limitations, and model updates—and process-level auditability, including the capacity to reconstruct the data received, the output produced, the action taken, and the resulting consequence. These elements support incident investigation and clearer lines of responsibility, but acquire practical value only when incorporated into the hospital's wider governance and patient-safety structures [88,93,94].

Explainability should consequently be understood neither as a purely technical add-on nor as a universal threshold that is identical across all AI applications. It is one component of risk-proportionate institutional governance, whose importance increases when outputs are difficult to verify, when they directly influence high-stakes decisions, or when responsibility is distributed across clinicians, managers, and technology providers. In this sense, explainability and auditability form part of the institutional evidence required to determine whether an AI system can be used safely and accountably within a specific hospital.

6. Precision Medicine, Data Integration, and Institutional Readiness

Precision medicine seeks to improve prevention, diagnosis, prognosis, and treatment by combining molecular, genomic, clinical, and environmental information to identify clinically relevant patient subgroups and differences in treatment response. It is related to, but not synonymous with, personalized care, which also incorporates comorbidities, patient preferences, social context, and the therapeutic relationship [95,96]. Artificial intelligence can contribute by integrating heterogeneous and high-dimensional data, identifying phenotypes and biomarkers, estimating risk, and supporting predictions of treatment response [97,98]. These capabilities, however, do not by themselves establish that an AI-supported recommendation is clinically useful or implementable.

The translation of precision medicine into routine hospital care depends on an end-to-end sociotechnical pathway that includes data generation and curation, laboratory and bioinformatics capacity, interoperable clinical and genomic information systems, clinical decision support, specialist interpretation, workforce training, and governance of consent, privacy, data reuse, and system updates. Pharmacogenomics and precision oncology illustrate this dependence: potentially actionable information must be delivered at the

appropriate point in the clinical workflow and linked to an evidence-supported action, while laboratories, clinicians, information systems, and financing arrangements remain coordinated [99,100]. A technically valid result may therefore have limited institutional value if the hospital cannot incorporate it into care, sustain the required infrastructure, or monitor its downstream consequences.

Precision medicine also sharpens questions of equity. Genomic and clinical datasets may underrepresent particular populations, while access to testing, specialist interpretation, targeted interventions, and follow-up varies across institutions and patient groups. Without explicit implementation strategies, advances in genomics and precision medicine may widen existing disparities rather than reduce them [101]. Institutional evaluation should therefore consider not only molecular or predictive validity, but also population representativeness, subgroup performance, access to downstream interventions, affordability, and opportunity costs.

Precision medicine is thus a useful test case for the central argument of this review. As AI-supported care becomes more dependent on multiple data sources, specialized expertise, coordinated workflows, and downstream treatment capacity, institutional conditions increasingly determine whether technical capability becomes sustainable clinical value. Evidence must therefore address not only whether a model, biomarker, or recommendation performs, but whether the adopting hospital can deliver the associated care pathway safely, equitably, and sustainably.

7. From Algorithmic Performance to Institutional Value

7.1. *The Indirectness of AI Value*

Most current diagnostic, predictive, and decision-support AI systems affect patient care indirectly. Their immediate output is information—a classification, risk estimate, alert, or recommendation—that must influence professional judgment, workflow, or action before clinical benefit can occur. Technical performance therefore represents only the beginning of an evaluative chain: a system may generate accurate outputs without changing clinical decisions, improving patient outcomes, or producing value at the level of the hospital. Evidence of success at one stage does not establish success at the stages that follow [5,13]. The route from algorithmic output to realized benefit depends on the clinical context, the person receiving the information, the action it triggers, and the organizational capacity to respond.

Two qualifications sharpen this argument. First, when AI moves from informing decisions toward acting through more autonomous or closed-loop functions, institutional requirements for safety assurance, monitoring, and accountability intensify, because the opportunity for contemporaneous professional scrutiny may be reduced. Second, the formal presence of a human decision-maker does not itself guarantee effective oversight. Automation bias, unclear allocation of responsibility, and limited practical capacity to challenge or override algorithmic outputs may weaken human control even when final authority nominally remains with the clinician [94,102]. The pathway from information to benefit is therefore not only indirect; it is contingent on the quality of clinical and organizational mediation at each stage.

7.2. *Evaluative Traditions and Their Genealogy*

The need to move beyond technical performance has been addressed through several partially overlapping evaluative traditions. The first originates in the hierarchy of efficacy formulated by Fryback and Thornbury [13] for diagnostic imaging. The hierarchy distinguishes six levels—technical efficacy, diagnostic-accuracy efficacy, diagnostic-thinking efficacy, therapeutic efficacy, patient-outcome efficacy, and societal efficacy—and rests on

the principle that evidence at a lower level is necessary, but not sufficient, to establish value at a higher one. Subsequent work on clinical AI has similarly emphasized the need to move from development-stage performance toward external validation and assessment of clinical effects [103]. Applying the efficacy hierarchy to 100 commercially available radiology AI products, van Leeuwen et al. [8] found that peer-reviewed evidence was limited and concentrated predominantly at the lower levels. RADAR subsequently adapted the hierarchy specifically for radiology AI and added a seventh level, “local efficacy,” to assess whether value demonstrated at preceding levels translates to, and remains valid within, a particular hospital environment [16].

More recent evidence indicates maturation without closure of this gap. Among 173 CE-certified radiology AI products available in 2023, the proportion supported by at least one peer-reviewed publication had increased from 36% in 2020 to 66%. Nevertheless, studies addressing higher efficacy levels accounted for approximately one quarter of the published evidence, and only 31% of products were supported by evidence reaching levels concerned with clinical or socioeconomic effects [9]. These findings suggest that the evidence base is expanding while remaining weighted toward technical feasibility and diagnostic accuracy rather than demonstrated consequences for care and health-service performance.

A second evaluative tradition is hospital-based health technology assessment (HB-HTA), which adapts health technology assessment to the hospital context to support decisions concerning technology adoption, procurement, continued use, and disinvestment. HB-HTA explicitly considers local clinical needs, economic and organizational consequences, and implementation requirements within the institution [15]. A third tradition is implementation science. Proctor et al. [14] distinguished implementation outcomes—including acceptability, adoption, appropriateness, feasibility, fidelity, implementation cost, penetration, and sustainability—from clinical and service outcomes. The Nonadoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework further demonstrated that these outcomes are shaped by interactions among the technology, its users, the organization, the wider institutional context, and their evolution over time [7].

A related, more recent strand of work addresses the organizational setting directly. Place-based evidence for clinical AI implementation emphasizes that evidence used to weigh an application’s benefits against its risks may have limited generalizability beyond the setting in which it was produced and calls for pragmatic approaches to local evidence generation [104]. This perspective further foregrounds the adopting organization, but does not explicitly formulate the overall hospital-level evaluative judgment that such locally generated evidence is intended to support.

More recently, AI-specific approaches have further developed translational evaluation and health-system governance. The Translational Evaluation of Healthcare AI (TEHAI) framework provides a structured appraisal of capability, utility, and adoption across development and deployment [105]. The Fair, Useful, and Reliable AI Models (FURM) framework combines ethical assessment, simulated usefulness, financial sustainability, information-technology feasibility, deployment planning, and prospective monitoring [106]. The international consensus guideline for trustworthy and deployable AI in healthcare (FUTURE-AI) provides lifecycle guidance extending from design and validation through deployment and monitoring [107]. The Framework for Appropriate Implementation and Review of Artificial Intelligence (FAIR-AI) offers a prescriptive health-system process for pre-implementation review, qualitative risk classification, safe implementation planning, and post-implementation monitoring [108]. Although these approaches differ in purpose, structure, and output, each organizes decision-relevant evidence beyond algorithmic performance.

The organizational level is therefore not unexplored. Taken together, these traditions and AI-specific approaches provide complementary procedures, criteria, and evidence for assessing clinical performance, local transferability, implementation, trustworthiness, organizational fit, and value. The contribution proposed here is not the addition of another assessment domain, but the explicit formalization of the all-things-considered hospital-level judgment that these heterogeneous inputs may collectively warrant and for which the adopting institution remains accountable. Institutional efficacy is proposed to denote that judgment rather than to compete with or replace the frameworks through which relevant evidence is generated, organized, and governed.

A terminological clarification is necessary. The term efficacy is retained to signal continuity with the Fryback–Thornbury tradition of differentiating evaluative claims, rather than to invoke the narrower conventional distinction between efficacy under controlled conditions and effectiveness in routine practice [13].

7.3. Institutional Efficacy as an Integrative Organizational Claim

Institutional efficacy is proposed as a conceptual construct that takes practical form as a warranted and revisable hospital-level evaluative judgment. Its unit of analysis is a defined AI-enabled service configuration—that is, the technology as deployed within a particular clinical or administrative service, workflow, workforce, infrastructure, financing, and governance arrangement—rather than the algorithm in isolation. The judgment concerns whether this configuration can generate and sustain patient-centered value under the adopting hospital’s actual priorities, constraints, and realistic capacity for adaptation. Relevant knowledge is distributed across clinical, operational, technical, administrative, and patient perspectives; responsibility for appraising and integrating these inputs, reconciling competing priorities, and reaching, justifying, and revising the overall judgment rests with hospital governance.

This judgment encompasses clinical and safety consequences; workflow, workforce, and organizational dynamics; economic and resource implications; governance and accountability; and stakeholder and system-level policy conditions insofar as they shape what the hospital can implement and sustain. The relative salience of these considerations is context-dependent and must be justified in relation to the use case and institutional setting. Institutional appraisal therefore requires not only the aggregation of evidence, but also appraisal of the validity, local relevance, completeness, and uncertainty of each input and of the interactions among them.

Net benefit is understood here in a multidimensional rather than narrowly financial sense. Patient benefit remains the normative endpoint, while the hospital is the locus through which adoption is coordinated, resources are committed, responsibilities are allocated, and benefit is sustained. Sustainability is therefore a condition for the continuing delivery of value rather than a competing objective. Institutional efficacy cannot be considered supported where operational or economic gains depend on clinically unacceptable harm or on the disproportionate transfer of burden or cost to patients, professionals, or other parts of the health system. Failures to meet applicable clinical, legal, or ethical minimums—including those concerning safety, data protection, equity, and fundamental rights—cannot be offset by gains elsewhere.

Existing frameworks may generate, organize, and govern substantial parts of the evidence required for this judgment. Their outputs nevertheless require institutional appraisal: a domain-specific assessment may be methodologically sound without being decision-sufficient. FAIR-AI illustrates this relationship. A favorable FAIR-AI review or low-risk designation may contribute evidence to an institutional-efficacy judgment, but does not by itself establish that adoption or continued use can deliver sustained patient-centered

value under the hospital's wider operating conditions and priorities [108]. Risk tolerability is therefore necessary but not sufficient for institutional efficacy. Place-based evidence similarly clarifies why evidence may need to be generated locally, whereas institutional efficacy specifies the hospital-level judgment that such evidence is intended to warrant [104]. The relevant comparator and time horizon are context-specific and should be stated for each appraisal; this review does not prescribe universal weights, thresholds, or a single decision rule.

RADAR extends the efficacy hierarchy by adding a further sequential level. Institutional efficacy instead operates across the hierarchy, integrating evidence from all seven levels with organization-specific evidence concerning implementation, workforce, economics, and governance (Figure 1).

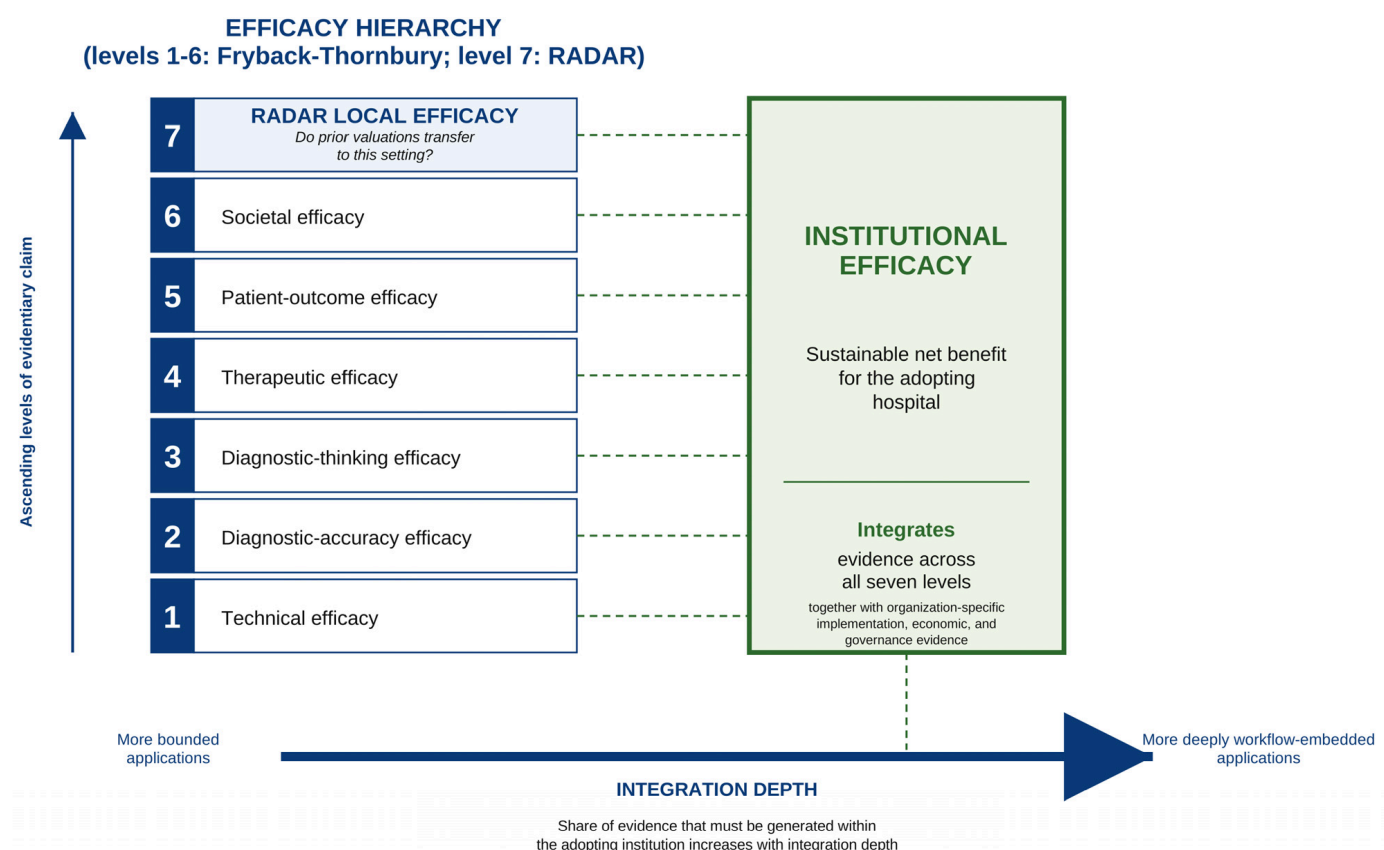


Figure 1. Institutional efficacy as an integrative organizational claim rather than an additional efficacy level. Levels 1–6 reproduce the efficacy hierarchy of Fryback and Thornbury [13], whereas level 7 represents the local efficacy level introduced by the Radiology AI Deployment and Assessment Rubric (RADAR) [16]. Institutional efficacy integrates evidence across all seven levels with organization-specific evidence concerning implementation, workforce, economic, and governance conditions. The lower arrow depicts the proposed relationship between deeper organizational integration and a greater need for locally generated or locally verified evidence; two additional properties may further modify the local evidence burden: context and data dependence, and the consequence profile of the application.

This perspective also distinguishes institutional from societal value. Under many payment and budget arrangements, hospitals bear the costs of acquisition, integration, workforce adaptation, maintenance, and monitoring, while part of the resulting benefit accrues to patients, payers, other providers, or the health system more broadly [109]. At the same time, sustained organizational adoption is a principal pathway through which demonstrated clinical benefit becomes population-level value. When technologies are not

adopted, are abandoned, or fail to become embedded in routine practice, their anticipated benefits may remain unrealized [7].

Institutional efficacy is a life-cycle judgment. Before deployment, it is necessarily prospective and provisional; following deployment, it must be substantiated, revised, restricted, or withdrawn in light of locally generated evidence. It informs decisions concerning adoption, conditional implementation, continuation, modification, restriction, or discontinuation. The operational specification of indicators, measurement methods, weighting principles, and decision thresholds requires a dedicated assessment framework and empirical validation and lies beyond the scope of the present review.

7.4. Integration Depth and the Local Evidence Burden

The proportion and type of evidence that must be generated locally are unlikely to be constant across AI applications. For the purposes of this review, integration depth is conceptualized as a multidimensional deployment profile rather than as a single additive scale. It comprises three organizational characteristics: organizational reach, referring to the number and diversity of professional roles and organizational units whose work is affected; workflow reconfiguration, referring to the extent to which tasks, sequences, handoffs, and responsibilities are altered; and decisional embedding, referring to the degree to which the system's outputs inform, trigger, or automate clinical or organizational action. These characteristics may vary independently, and the same technical system may therefore exhibit a different integration profile depending on how it is deployed. They should not be reduced to a universal integration score. This formulation synthesizes recurring implementation and sociotechnical concerns in the clinical-AI literature [5–7].

Two further properties shape the local evidence burden without defining integration depth itself. Context and data dependence concern the extent to which system performance and use rely on institution-specific data-generating practices, infrastructure, casemix, coding, care pathways, and response capacity [5,81,82]. The consequence profile concerns the severity, immediacy, reversibility, and scale of potential error or unintended effects. This distinction is consistent with risk-proportionate governance approaches that vary the level of scrutiny according to the magnitude and manageability of potential adverse consequences [108]. Thus, a documentation assistant may become deeply embedded across several workflows while carrying comparatively limited immediate clinical consequences, whereas a bounded autonomous screening application may alter relatively little of the surrounding workflow but require stringent evidence because errors may be consequential and less readily intercepted.

The proposition advanced here is that, as organizational integration deepens, the realized value of an AI-enabled service becomes increasingly contingent on local workflows, professional use, response capacity, allocation of responsibility, and the organization's capacity to sustain use. External studies may establish technical performance and provide evidence of clinical utility, but they cannot fully represent these institution-specific interactions [5–7]. Deeper integration therefore increases the share of evidence concerning realized value that must be generated or verified within the adopting institution. Context and data dependence, together with the consequence profile, separately influence the type, amount, and rigor of evidence required.

7.5. Producing Institutional Evidence: Governance as an Evidence-Generating Infrastructure

Producing institutional evidence begins with distinguishing findings that can reasonably be transferred from uncertainties that require examination in the adopting setting. Published technical validation, comparative clinical studies, and economic evaluations provide a basis for appraisal when their populations, intended uses, workflows, and an-

alytical assumptions are relevant to the proposed deployment. The extent of additional local evaluation should follow from an assessment of this applicability, rather than from an assumption that all evidence must be generated anew [5,15].

Local assessment addresses the functioning of the installed system and its interaction with actual care delivery. Technical verification establishes whether the system and its data pipeline operate as intended; validation, including calibration and subgroup assessment where relevant, examines whether performance remains adequate in the population served. Organizational evaluation addresses professional use, response capacity, workload, resource requirements, and implementation costs. Observation of practice, staff and patient feedback, and usage records can reveal informal workarounds, resistance, and redistribution of work that technical performance reports may not capture [5,7,15].

These tasks differ from establishing that AI-supported care improves outcomes. Shadow-mode evaluation can examine prospective performance and potential alert volumes without exposing outputs to clinical decision-making, but cannot demonstrate the consequences of professionals acting on those outputs [16]. Where existing evidence is insufficient for the proposed use, prospective workflow evaluation may be needed; questions about incremental clinical benefit require an appropriate comparative design. For assistive systems, the relevant comparison is generally care with AI support versus the corresponding care pathway without that support, rather than isolated algorithm-versus-clinician performance. Early clinical studies can examine safety, feasibility, and human–AI interaction, while stronger claims about clinical effectiveness require designs capable of distinguishing intervention effects from other changes in care [5,110].

Governance can connect these sources of evidence through an ongoing cycle of observation, analysis, and organizational response. Monitoring, audit, and incident review provide records whose evaluative usefulness depends on data quality, consistent measures, explicit questions, and appropriate baselines or comparators. Interpreting these records alongside evidence about how the service is used and adapted links routine practice to institutional learning, consistent with the learning-health-system approach [7,111]. Within this review, governance is therefore understood as an infrastructure for sustaining inquiry and informing revision, rather than as a substitute for the study designs required to establish clinical effects.

Local responsibility does not require every hospital to reproduce complete patient-outcome evidence independently. Collaborative evaluation can concentrate research and analytical resources while retaining institution-specific examination of applicability and service performance, consistent with the emphasis on partnerships and contextualized appraisal in hospital-based HTA [15]. The proposed division of work combines shared generation of transferable evidence with local verification of the conditions under which it is used. This is particularly important for resource-constrained hospitals: locally grounded evaluation can be supported by external expertise and shared infrastructure without requiring each institution to maintain a self-sufficient research team.

Table 2 situates institutional efficacy in relation to the established efficacy hierarchy and RADAR's local efficacy, distinguishing it from an additional sequential level.

Table 3 complements the efficacy architecture presented in Table 2 by positioning institutional efficacy in relation to selected established and AI-specific evaluative approaches and identifying the contribution each may make to the hospital-level judgment proposed here.

Table 2. Established efficacy levels, RADAR local efficacy, and institutional efficacy as an integrative organizational claim.

Evaluative Level or Construct	Core Evaluative Question	Illustrative Evidence
1. Technical efficacy	Does the system function reliably under defined technical conditions?	Technical verification; reliability, robustness, and quality-control testing
2. Diagnostic-accuracy efficacy	How accurately does the system identify, classify, or predict the target condition?	Sensitivity, specificity, discrimination, calibration, and external validation
3. Diagnostic-thinking efficacy	Does the information produced by the system alter clinical reasoning or diagnostic confidence?	Decision-impact studies; clinician-response and diagnostic-confidence studies
4. Therapeutic efficacy	Does use of the system alter treatment or clinical-management decisions?	Prospective studies of changes in management and therapeutic action
5. Patient-outcome efficacy	Does use of the system improve outcomes that matter to patients?	Comparative outcome studies; pragmatic trials; randomized studies where appropriate
6. Societal efficacy	Does the technology produce net value at the population or health-system level?	Cost-effectiveness and cost-benefit analyses; health technology assessment; population-level evaluation
7. Local efficacy—RADAR	Do valuations established at preceding levels translate to, and remain valid within, the particular hospital environment?	Local budget-impact analysis; multi-criteria decision analysis; workflow or shadow-mode evaluation; prospective monitoring
Institutional efficacy—integrative organizational construct proposed in this review	Can this defined AI-enabled service configuration generate and sustain patient-centered value under the hospital's actual priorities, constraints, and capacity for adaptation?	External clinical and technical evidence; local validation or verification where required; implementation and workflow evidence; workforce and resource implications; economic and budget-impact evidence; governance and accountability arrangements; post-deployment monitoring and organizational learning

Note. Levels 1–6 reproduce the efficacy hierarchy of Fryback and Thornbury [13]. Local efficacy is the seventh level introduced by the Radiology AI Deployment and Assessment Rubric (RADAR) for radiology AI [16]. Institutional efficacy is proposed as an integrative organizational construct rather than an eighth sequential level. It draws on evidence across the efficacy hierarchy together with organization-specific evidence concerning implementation, workforce, economics, governance, and sustained use. The evidentiary approaches shown are illustrative rather than exhaustive.

Table 3. Positioning institutional efficacy in relation to selected established and AI-specific evaluative approaches.

Approach	Primary Unit and Purpose	Main Evaluative Content	Typical Output	Relationship to Institutional Efficacy
Efficacy hierarchy [13]	Diagnostic technology; orders evidence from technical function to societal value	Technical performance; diagnostic accuracy; clinical reasoning; therapeutic decisions; patient outcomes; societal value	Level-specific profile of available evidence	Provides the evidentiary architecture across which institutional efficacy integrates evidence
RADAR [16]	Radiology AI in a local setting; extends the hierarchy through local efficacy	Levels 1–6 plus local adaptability, budget impact, workflow evaluation, and monitoring	Seven-level assessment including local transferability	Provides local-transferability evidence; institutional efficacy integrates across, rather than extends, the sequence

Table 3. Cont.

Approach	Primary Unit and Purpose	Main Evaluative Content	Typical Output	Relationship to Institutional Efficacy
Hospital-based HTA [15]	Health technology within a hospital; supports adoption, continued use, and disinvestment	Clinical, economic, organizational, strategic, and implementation considerations	Contextualized multidisciplinary appraisal for hospital decision-makers	May serve as a principal appraisal process and source of local evidence
Implementation science and NASSS [7,14]	Implementation process and technology-supported program in context; explains adoption and sustainability	Implementation outcomes and interactions among technology, users, organization, and context over time	Implementation-outcome profile and account of contextual complexity	Informs whether anticipated benefits are organizationally realizable and sustainable
Place-based evidence [104]	Emerging perspective on the evidence-setting relationship; addresses limited transferability and local evidence generation	Context specificity of implementation evidence	Rationale and approaches for producing evidence locally	Explains why local evidence may be required; institutional efficacy specifies the judgment it informs
TEHAI [105]	AI system across development and deployment; appraises translational readiness	Capability, utility, and adoption across 15 subcomponents	Scored appraisal of weighted subcomponents across the three domains	Provides structured evidence on translational readiness
FURM [106]	Candidate AI-guided workflow or use case; estimates achievable benefit and implementation feasibility	Ethical fit, simulated usefulness, financial sustainability, IT feasibility, deployment, and monitoring	Integrated assessment and prospective monitoring plan	Provides evidence on achievable usefulness, feasibility, sustainability, and organizational capacity to act
FUTURE-AI [107]	AI tool across its lifecycle; specifies conditions for trustworthy and deployable AI	Fairness, universality, traceability, usability, robustness, and explainability	Lifecycle principles and 30 best-practice recommendations	Provides trustworthiness, usability, local-validity, and monitoring requirements relevant to institutional efficacy
FAIR-AI [108]	AI solution and intended use within a health system; structures review, risk governance, and monitoring	Validation, usefulness and benefit-risk, workflow, resource use, equity, compliance, and monitoring	Risk classification, summary statement, and Safe AI Plan; governance disposition for high-risk cases	Contributes structured risk-governance and implementation evidence to the institutional judgment
Institutional efficacy (proposed in this review)	Defined AI-enabled service configuration in a specific hospital; assesses whether patient-centered value is achievable and sustainable	Clinical and safety consequences; workflow and workforce effects; economic and resource implications; governance, stakeholder, and policy conditions	Collectively informed, context-sensitive, and revisable hospital-level judgment for which hospital governance remains accountable	Integrates and adjudicates heterogeneous evidence in relation to the hospital's mission, priorities, constraints, and capacity for adaptation

Note. The selected approaches differ in purpose and output and are presented as complementary rather than mutually exclusive. The comparison is an interpretive synthesis of the cited sources. Abbreviations: AI, artificial intelligence; FAIR-AI, Framework for Appropriate Implementation and Review of Artificial Intelligence; FURM, Fair, Useful, and Reliable AI Models; FUTURE-AI, international consensus guideline for trustworthy and deployable AI in healthcare; HB-HTA, hospital-based health technology assessment; IT, information technology; NASSS, Nonadoption, Abandonment, Scale-up, Spread, and Sustainability; RADAR, Radiology AI Deployment and Assessment Rubric; TEHAI, Translational Evaluation of Healthcare AI.

8. Limitations of the Evidence Base and of the Present Review

The evidence base for clinical AI remains uneven across tasks, settings, and evaluative levels. A substantial proportion of studies continue to emphasize retrospective performance, curated datasets, and performance-centered endpoints, while prospective comparative evaluation and evidence concerning patient outcomes, workflow effects, resource use, and organizational sustainability remain less common [3,4,9]. The prospective studies discussed in Section 4 demonstrate progress beyond these limitations, but they do not establish equivalent evidence across products, specialties, or modes of deployment.

A second limitation concerns transferability. Models evaluated in research-intensive or highly digitized environments may not perform or function in the same way when deployed across different populations, data-generating practices, staffing configurations, workflows, infrastructures, and response capacities. External validation can reduce this uncertainty

but cannot eliminate it: calibration, alert burden, subgroup performance, professional response, and organizational consequences may continue to vary across institutions and over time [5,81,82]. Evidence relevant to hospital adoption must therefore be interpreted in relation to the specific use case and deployment configuration rather than as a fixed property of the technology.

The conceptual synthesis advanced in this review also has limitations. Institutional efficacy is proposed as a conceptual construct expressed through a hospital-level evaluative judgment; it has not yet been operationalized or empirically validated as a formal assessment framework. The dimensions through which such a judgment should be measured, the relative salience of different forms of evidence, and the conditions under which competing considerations should be reconciled remain to be established. Similarly, integration depth is proposed as a multidimensional deployment profile rather than a validated scale, drawing on recurring implementation and sociotechnical concerns identified in the literature [5–7,16,104]. Its relationship with local evidence requirements, context and data dependence, and the consequence profile of AI applications requires empirical examination across technologies, hospitals, and health systems.

The review design imposes further limitations. Source selection was purposive rather than exhaustive, and the review did not apply a uniform formal risk-of-bias instrument across the heterogeneous methodological, clinical, implementation, economic, governance, and policy literature included. The search and synthesis were conducted by a single author, and the absence of a prospectively archived systematic-review screening log limits reproducibility of source selection. Relevant empirical studies, implementation reports, non-indexed evidence, or recent developments may therefore have been omitted. These limitations should be considered when interpreting broad statements about evidence maturity and when evaluating the conceptual propositions advanced in this review.

9. Discussion: Implications of Institutional Efficacy for Hospital Governance

The central implication of institutional efficacy is that hospital evaluation should focus on the specific AI-enabled service configuration rather than on the AI product in isolation. The relevant configuration includes the intended use, patient population, data environment, workflow position, professional users, supporting infrastructure, financing arrangements, and governance conditions through which the technology is deployed. Evidence generated elsewhere may therefore be highly informative without being automatically sufficient for a particular adoption. The hospital-level task is to determine which findings are reasonably transferable and which uncertainties require local verification under the proposed conditions of use [5,7,15].

The extent of that local component should be proportionate rather than uniform. Integration depth, context and data dependence, and the consequence profile of the application affect the types of uncertainty that remain after external evidence has been considered. A bounded application operating within an established workflow may require relatively limited local verification if relevant external evidence is strong, whereas systems that depend heavily on institution-specific data, alter multiple professional workflows, or trigger consequential actions may require more extensive prospective evaluation and monitoring [5–7]. This does not imply that every hospital must reproduce the complete clinical evidence base independently.

The application examples reviewed in Section 4 illustrate why this distinction matters. In mammography, prospective trials evaluate an AI-supported screening strategy rather than algorithmic accuracy alone [71,72]. For deterioration support, prospective evidence may indicate clinical benefit while simultaneously revealing greater use of downstream

intensive-care resources [83]. In colonoscopy, pooled randomized evidence supporting computer-aided detection has not been reproduced uniformly in every pragmatic clinical setting [84,85]. These examples do not imply that local context overrides external evidence; rather, they show that the consequences relevant to hospital adoption extend beyond the endpoint on which the AI system was initially evaluated.

Implementation research likewise indicates that adoption is a sequence of organizational activities rather than a single acquisition event [112]. The governance implications of institutional efficacy can therefore be considered across pre-adoption appraisal, post-deployment evaluation, and the distribution of evidentiary responsibilities.

Table 4 summarizes selected approaches to evaluating model performance, clinician–AI interaction, and service performance [5,16,103,110], linking their findings to institutional appraisal. Consistent with the proposition developed in Section 7.4, deeper integration is proposed to increase the share of realized-value evidence requiring local generation or verification. Context and data dependence affect the applicability of external evidence, while the consequence profile informs the type and rigor of evidence required.

Table 4. Selected evaluation approaches, evidentiary limits, and local assessment needs for institutional appraisal.

Evaluation Approach	Question and Comparison Basis	Illustrative Evaluation Measures	What It Cannot Establish Alone	Transferability and Local Assessment Needs
Standalone model-performance validation	How well does the model perform on independent validation data against a reference standard or observed outcomes?	Accuracy; discrimination; calibration; technical failure rate	Clinical benefit, workflow effects, or performance in an untested setting	External results inform appraisal; applicability depends on model version, population, data acquisition, and local data practices.
Local current-care baseline	What are the outcomes and resource demands of the hospital's current care pathway without AI?	Diagnostic safety; decision quality; turnaround time; workload; cost per case	Whether AI can address the identified need, or what its incremental effects would be	Local by definition; external benchmarks inform comparison but cannot establish the hospital's own baseline.
AI-versus-clinician benchmarking	How does standalone AI compare with clinicians on the same task against a common reference standard?	Sensitivity; specificity; error patterns	Effects of AI assistance or performance of the combined care pathway	Applicability depends on comparable tasks, cases, clinician expertise, and evaluation conditions.
Clinician-plus-AI reader study	Does AI assistance improve reader performance compared with unaided reading under controlled conditions?	Diagnostic accuracy; decision quality; reading time	Routine workflow effects, patient benefit, or long-term sustainability	Applicability depends on reader expertise, case mix, interface, and study conditions
Shadow-mode evaluation	How does the installed system perform on prospective local data against reference outcomes, without influencing care?	Pipeline reliability; calibration; subgroup performance; potential alert volume	Professional responses or consequences of acting on AI outputs	Examines the installed configuration; external findings inform evaluation but cannot verify its local operation.
Prospective workflow and clinical evaluation	How does the AI-supported pathway perform in practice? Usual care or a justified alternative provides the comparator for estimating incremental effects.	Feasibility; safety; patient outcomes; response time; workload; downstream utilization; implementation and incremental costs	Causal benefit from prospective observation alone; effects beyond the observed follow-up	Applicability depends on comparable care pathways, case mix, and staffing; local workflows, capacity, and costs need site-specific appraisal.
Post-deployment monitoring	Do performance, benefits, and burdens remain acceptable against baselines, predefined thresholds, or relevant comparators?	Drift; subgroup outcomes; incidents; workload; ongoing costs; sustained use	Causal benefit when monitoring is uncontrolled	Local monitoring can use shared infrastructure; pooled findings support learning without replacing site-specific assessment.

Note. Approaches may overlap and need not follow a fixed sequence; selection depends on intended use, existing evidence, and unresolved uncertainty.

9.1. Pre-Adoption Evaluation and Procurement

Pre-adoption appraisal begins with a precise definition of the proposed use: the clinical or operational task, target population, users, decision point, expected action, and surrounding service pathway. External evidence should then be assessed for its relevance to this configuration, including study design, comparator, validation population, equipment or data environment, calibration where applicable, subgroup performance, and duration of follow-up. The relevant comparator and time horizon should be explicit. For assistive systems, the appropriate comparison will usually concern the AI-supported care pathway versus the corresponding pathway without AI support, rather than the isolated performance of an algorithm against that of a clinician.

Where important uncertainties remain, targeted local validation, shadow-mode evaluation, or prospective workflow assessment may be warranted before routine clinical reliance. The Developmental and Exploratory Clinical Investigations of Decision-support systems driven by Artificial Intelligence (DECIDE-AI) guideline provides a useful reference for the early-stage live clinical evaluation and reporting of AI-based decision-support systems [110]. Such evaluation should address the uncertainty relevant to the proposed adoption rather than reproduce evidence that is already sufficiently transferable.

Regulatory conformity is necessary where applicable but should not be interpreted as evidence of institutional value. Under the Medical Device Regulation, conformity assessment and CE marking concern compliance with the requirements applicable to the device and its declared intended purpose [113]. They do not by themselves establish that implementation will improve patient outcomes, reduce workload, be cost-effective, or generate sustainable value within a particular hospital [113,114]. Regulatory evidence may therefore contribute to institutional appraisal without determining its overall conclusion.

Procurement should consequently extend beyond acquisition price and headline performance. Relevant considerations include interoperability; data access, ownership, retention, and cybersecurity; model versioning and updates; vendor responsibilities for maintenance and corrective action; audit access; and exit arrangements that reduce avoidable technological dependence. Workflow assessment should identify where the system enters care, who receives and acts on its output, how disagreement or override is managed, and whether the service has sufficient capacity to respond. Staffing, training, and the continuing availability of appropriate human expertise form part of the deployment configuration rather than ancillary implementation concerns [94,115].

Economic appraisal similarly requires attention to the perspective of the adopting institution. Budget impact, implementation and maintenance costs, workforce adaptation, cybersecurity, monitoring, displaced activity, and opportunity costs may differ materially from societal cost-effectiveness [109,116]. Benefits and burdens may also fall on different actors: a hospital may bear implementation costs while benefits accrue partly to patients, payers, or other services. Pre-adoption appraisal should therefore consider both the expected incremental benefit of the AI-supported pathway and the organizational resources required to sustain it [15,109]. The existing care pathway provides the relevant baseline; non-adoption should not be treated as a consequence-free alternative when current care itself carries costs, delays, risks, or unmet need.

9.2. Post-Adoption Governance, Monitoring, and Discontinuation

Deployment converts assumptions made during pre-adoption appraisal into propositions that can be examined under routine conditions. Monitoring should therefore be aligned with the mechanism through which value is expected to occur. Depending on the application, this may include discrimination and calibration, subgroup performance, data and performance drift, alert burden, response times, override patterns, clinical outcomes,

workflow effects, resource use, and unintended consequences [5,7]. Monitoring only algorithmic performance may miss failure arising from professional response, inadequate downstream capacity, or changes in the surrounding service.

Material changes in the model, data pipeline, equipment, patient population, clinical pathway, or organizational conditions should prompt reassessment where they alter the basis on which the original judgment was made. Auditability is important for reconstructing the model version, relevant input data, output, professional response, and subsequent consequence. AI-related adverse events and near misses should enter established patient-safety and quality-improvement processes, with clear routes for investigation and corrective action. Retaining a clinician nominally “in the loop” does not in itself establish effective oversight if authority, information, time, or practical capacity to challenge the system are inadequate [94,115].

These requirements increasingly intersect with formal legal obligations. For high-risk AI systems falling within the relevant provisions of the EU Artificial Intelligence Act, deployers are subject to obligations concerning use in accordance with instructions, appropriately competent human oversight, monitoring, and retention of automatically generated logs under their control; where the conditions specified in the Regulation are met, use must be suspended when relevant risks emerge. A fundamental-rights impact assessment is additionally required for the categories of deployer and system to which Article 27 applies [114,117]. These obligations do not establish institutional efficacy, but they create governance and documentation processes that can contribute to the evidence on which continuing institutional appraisal depends.

Institutional efficacy is therefore necessarily revisable. Continuing use should depend not simply on the absence of major incidents, but on whether the evidence supporting the expected patient-centered value remains credible under current conditions. Modification, restriction, suspension, or discontinuation may become justified when performance deteriorates, burdens or harms emerge, the technology changes materially, anticipated benefits are not sustained, or the organization can no longer support appropriate use. Discontinuation should in this sense be understood as a possible outcome of responsible life-cycle governance rather than as evidence of implementation failure alone.

9.3. Evidentiary Responsibility, Institutional Accountability, and Shared Capacity

Evidence generation for hospital AI is distributed across multiple actors, but responsibility for the overall institutional judgment should not become correspondingly diffuse. Developers and vendors are positioned to provide technical documentation, validation evidence, known limitations, subgroup analyses, update procedures, and sufficient information for audit and monitoring. Regulators establish requirements for market access, safety, and performance, while health technology assessment bodies and related evaluative structures can contribute clinical, economic, organizational, and comparative evidence. The AI Act further distinguishes provider and deployer obligations, giving legal form to part of this distribution of responsibility [114,117].

Hospitals nevertheless remain the locus at which these heterogeneous inputs are applied to a particular service configuration. They must determine local compatibility, organize workflow and oversight, assess whether the service can respond to AI-generated information, evaluate resource implications, and revisit use as conditions change. Institutional efficacy is therefore collectively informed, but the responsibility for integrating, warranting, and revising the hospital-level judgment rests with hospital governance.

This responsibility should not be interpreted as requiring every hospital to reproduce complete clinical evaluations independently. Hospital-based HTA and AI implementation research emphasize the need for adequate resources, expertise, and technical infrastruc-

ture [15,112], while work on AI in resource-constrained settings highlights the importance of context-appropriate implementation [118]. These requirements suggest that research-intensive and highly digitized hospitals are likely to be better positioned to undertake extensive local validation, drift analysis, and prospective service evaluation, whereas institutions with more limited analytical and digital capacity may face comparable governance responsibilities with fewer resources. This comparison is a literature-informed inference rather than a comparative empirical finding of the review. Requiring each institution to duplicate the same evaluation could therefore increase inequity in access to well-evaluated AI rather than improve governance.

Shared evidence infrastructures offer a more proportionate response. National or regional evidence standards, collaborative validation, hospital-based HTA networks, common monitoring indicators, pooled post-market surveillance, and mechanisms for exchanging evidence among sufficiently comparable institutions can reduce unnecessary duplication while preserving local examination of applicability and service performance [15]. The distinction is therefore not between external and local evidence as competing alternatives, but between evidence that can reasonably be transferred and evidence whose meaning depends materially on the adopting setting.

These implications do not constitute a validated scoring system or decision algorithm. Rather, they specify the governance consequences of treating institutional efficacy as a warranted and revisable hospital-level judgment: heterogeneous evidence must be assembled in relation to a defined deployment, uncertainty must be made explicit, local evaluation should be proportionate to what remains unresolved, and responsibility for the resulting judgment must remain identifiable.

10. Conclusions

Clinical AI evidence remains uneven across applications and evaluative levels. Technical performance and diagnostic accuracy are increasingly well documented for selected uses, whereas prospective clinical utility, workflow consequences, resource implications, and organizational sustainability are evaluated less consistently. Hospital adoption therefore requires interpretation of external evidence in relation to the specific AI-enabled service configuration in which the technology is intended to operate.

This review proposes institutional efficacy as a conceptual construct for expressing the warranted and revisable hospital-level judgment that such a configuration can generate and sustain patient-centered value under actual institutional conditions. It operates across, rather than beyond, established efficacy levels and draws on evidence concerning clinical benefit and safety, implementation, workflow and workforce effects, economic consequences, governance, and organizational capacity. Existing evaluative frameworks provide substantial parts of this evidence; institutional efficacy concerns how these heterogeneous inputs are integrated into an accountable judgment about a particular deployment.

The integration-depth proposition further suggests that deeper organizational embedding increases the proportion of evidence concerning realized value that must be generated or verified locally. Two additional properties—context and data dependence, and the consequence profile of the application—separately shape the type and rigor of evidence required. External evidence therefore remains indispensable, while local validation, workflow evaluation, monitoring, audit, and incident learning address uncertainties that cannot be resolved through transferability alone. Shared evidence infrastructures may reduce duplication and support hospitals with unequal analytical capacity.

Future research should operationalize and empirically test institutional efficacy, examine the relationship between integration depth and local evidence requirements across technologies and hospital settings, and determine how different forms of evidence can be

combined without assuming universal weights or thresholds. Comparative research should also establish whether the construct provides additional explanatory or decision value beyond RADAR, hospital-based health technology assessment, implementation-science frameworks, and emerging AI-specific governance approaches. The value of clinical AI is ultimately realized not by algorithmic performance alone, but through the institutional conditions that convert capability into safe, sustainable, and accountable care.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/hospitals3040019/s1>, Table S1: Main deep-learning, generative, multimodal, and interpretable model families in clinical AI: properties, applications, and limitations; Table S2: SANRA self-assessment of the narrative review.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: This work has been partly supported by the University of Piraeus Research Center. During the preparation of this manuscript, the author used Claude 5.5 (Anthropic) to assist with English-language editing and translation of author-drafted content, with formatting of the manuscript and tables, and with organizing and cross-checking bibliographic references. The author reviewed and verified all content, including the accuracy of all references, and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Wang, F.; Preininger, A. AI in health: State of the art, challenges, and future directions. *Yearb. Med. Inform.* **2019**, *28*, 16–26. [[CrossRef](#)] [[PubMed](#)]
2. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **2019**, *380*, 1347–1358. [[CrossRef](#)] [[PubMed](#)]
3. Nagendran, M.; Chen, Y.; Lovejoy, C.A.; Gordon, A.C.; Komorowski, M.; Harvey, H.; Topol, E.J.; Ioannidis, J.P.A.; Collins, G.S.; Maruthappu, M. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **2020**, *368*, m689. [[CrossRef](#)] [[PubMed](#)]
4. Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D.S.W.; Karthikesalingam, A.; King, D.; Ashrafian, H.; Darzi, A. Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *npj Digit. Med.* **2021**, *4*, 65. [[CrossRef](#)] [[PubMed](#)]
5. Kelly, C.J.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **2019**, *17*, 195. [[CrossRef](#)] [[PubMed](#)]
6. Shaw, J.; Rudzicz, F.; Jamieson, T.; Goldfarb, A. Artificial intelligence and the implementation challenge. *J. Med. Internet Res.* **2019**, *21*, e13659. [[CrossRef](#)] [[PubMed](#)]
7. Greenhalgh, T.; Wherton, J.; Papoutsis, C.; Lynch, J.; Hughes, G.; A’Court, C.; Hinder, S.; Fahy, N.; Procter, R.; Shaw, S. Beyond adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *J. Med. Internet Res.* **2017**, *19*, e367. [[CrossRef](#)] [[PubMed](#)]
8. van Leeuwen, K.G.; Schalekamp, S.; Rutten, M.J.C.M.; van Ginneken, B.; de Rooij, M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur. Radiol.* **2021**, *31*, 3797–3804. [[CrossRef](#)] [[PubMed](#)]
9. Antonissen, N.; Tryfonos, O.; Houben, I.B.; Jacobs, C.; de Rooij, M.; van Leeuwen, K.G. Artificial intelligence in radiology: 173 commercially available products and their scientific evidence. *Eur. Radiol.* **2026**, *36*, 526–536. [[CrossRef](#)] [[PubMed](#)]
10. Moor, M.; Banerjee, O.; Abad, Z.S.H.; Krumholz, H.M.; Leskovec, J.; Topol, E.J.; Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature* **2023**, *616*, 259–265. [[CrossRef](#)] [[PubMed](#)]
11. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. [[CrossRef](#)] [[PubMed](#)]

12. Gallifant, J.; Afshar, M.; Ameen, S.; Aphinyanaphongs, Y.; Chen, S.; Cacciamani, G.; Demner-Fushman, D.; Dligach, D.; Daneshjou, R.; Fernandes, C.; et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **2025**, *31*, 60–69. [[CrossRef](#)] [[PubMed](#)]
13. Fryback, D.G.; Thornbury, J.R. The efficacy of diagnostic imaging. *Med. Decis. Mak.* **1991**, *11*, 88–94. [[CrossRef](#)] [[PubMed](#)]
14. Proctor, E.; Silmere, H.; Raghavan, R.; Hovmand, P.; Aarons, G.; Bunker, A.; Griffey, R.; Hensley, M. Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Adm. Policy Ment. Health Ment. Health Serv. Res.* **2011**, *38*, 65–76. [[CrossRef](#)] [[PubMed](#)]
15. Sampietro-Colom, L.; Lach, K.; Pasternack, I.; Wasserfallen, J.-B.; Cicchetti, A.; Marchetti, M.; Kidholm, K.; Arentz-Hansen, H.; Rosenmöller, M.; Wild, C.; et al. Guiding principles for good practices in hospital-based health technology assessment units. *Int. J. Technol. Assess. Health Care* **2015**, *31*, 457–465. [[CrossRef](#)] [[PubMed](#)]
16. Boverhof, B.J.; Redekop, W.K.; Bos, D.; Starmans, M.P.A.; Birch, J.; Rockall, A.; Visser, J.J. Radiology AI Deployment and Assessment Rubric (RADAR) to bring value-based AI into radiological practice. *Insights Imaging* **2024**, *15*, 34. [[CrossRef](#)] [[PubMed](#)]
17. Green, B.N.; Johnson, C.D.; Adams, A. Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. *J. Chiropr. Med.* **2006**, *5*, 101–117. [[CrossRef](#)] [[PubMed](#)]
18. Grant, M.J.; Booth, A. A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Inf. Libr. J.* **2009**, *26*, 91–108. [[CrossRef](#)] [[PubMed](#)]
19. Ferrari, R. Writing narrative style literature reviews. *Med. Writ.* **2015**, *24*, 230–235. [[CrossRef](#)]
20. Baethge, C.; Goldbeck-Wood, S.; Mertens, S. SANRA—A scale for the quality assessment of narrative review articles. *Res. Integr. Peer Rev.* **2019**, *4*, 5. [[CrossRef](#)] [[PubMed](#)]
21. Torracco, R.J. Writing integrative literature reviews: Guidelines and examples. *Hum. Resour. Dev. Rev.* **2005**, *4*, 356–367. [[CrossRef](#)]
22. Snyder, H. Literature review as a research methodology: An overview and guidelines. *J. Bus. Res.* **2019**, *104*, 333–339. [[CrossRef](#)]
23. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nat. Med.* **2019**, *25*, 24–29. [[CrossRef](#)] [[PubMed](#)]
24. Mitchell, T.M. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997.
25. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
26. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.
27. Eloranta, S.; Boman, M. Predictive models for clinical decision making: Deep dives in practical machine learning. *J. Intern. Med.* **2022**, *292*, 278–295. [[CrossRef](#)] [[PubMed](#)]
28. Naeem, S.; Ali, A.; Anam, S.; Ahmed, M.M. An unsupervised machine learning algorithms: Comprehensive review. *Int. J. Comput. Digit. Syst.* **2023**, *13*, 911–921. [[CrossRef](#)] [[PubMed](#)]
29. Kabir, M.F.; Chen, T.; Ludwig, S.A. A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthc. Anal.* **2023**, *3*, 100125. [[CrossRef](#)]
30. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 15. [[CrossRef](#)]
31. Samariya, D.; Ma, J.; Aryal, S.; Zhao, X. Detection and explanation of anomalies in healthcare data. *Health Inf. Sci. Syst.* **2023**, *11*, 20. [[CrossRef](#)] [[PubMed](#)]
32. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 1998.
33. Yu, C.; Liu, J.; Nemati, S.; Yin, G. Reinforcement learning in healthcare: A survey. *ACM Comput. Surv.* **2021**, *55*, 5. [[CrossRef](#)]
34. Al-Hamadani, M.N.A.; Fadhel, M.A.; Alzubaidi, L.; Harangi, B. Reinforcement learning algorithms and applications in healthcare and robotics: A comprehensive and systematic review. *Sensors* **2024**, *24*, 2461. [[CrossRef](#)] [[PubMed](#)]
35. Komorowski, M.; Celi, L.A.; Badawi, O.; Gordon, A.C.; Faisal, A.A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* **2018**, *24*, 1716–1720. [[CrossRef](#)] [[PubMed](#)]
36. Tseng, H.-H.; Luo, Y.; Cui, S.; Chien, J.-T.; Ten Haken, R.K.; El Naqa, I. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Med. Phys.* **2017**, *44*, 6690–6705. [[CrossRef](#)] [[PubMed](#)]
37. Gottesman, O.; Johansson, F.; Komorowski, M.; Faisal, A.; Sontag, D.; Doshi-Velez, F.; Celi, L.A. Guidelines for reinforcement learning in healthcare. *Nat. Med.* **2019**, *25*, 16–18. [[CrossRef](#)] [[PubMed](#)]
38. Gui, J.; Chen, T.; Zhang, J.; Cao, Q.; Sun, Z.; Luo, H.; Tao, D. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 9052–9071. [[CrossRef](#)] [[PubMed](#)]
39. Rani, V.; Kumar, M.; Gupta, A.; Sachdeva, M.; Mittal, A.; Kumar, K. Self-supervised learning for medical image analysis: A comprehensive review. *Evol. Syst.* **2024**, *15*, 1607–1633. [[CrossRef](#)]
40. Xing, X.; Liang, G.; Wang, C.; Jacobs, N.; Lin, A.-L. Self-supervised learning application on COVID-19 chest X-ray image classification using masked autoencoder. *Bioengineering* **2023**, *10*, 901. [[CrossRef](#)] [[PubMed](#)]

41. Aviles-Rivero, A.I.; Papadakis, N.; Li, R.; Sellars, P.; Fan, Q.; Tan, R.T.; Schönlieb, C.-B. GraphXNET—Chest X-ray classification under extreme minimal supervision. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*; Springer: Cham, Switzerland, 2019; pp. 504–512. [\[CrossRef\]](#)
42. Shakya, K.S.; Alavi, A.; Porteous, J.; Khatri, P.; Laddi, A.; Jaiswal, M. A critical analysis of deep semi-supervised learning approaches for enhanced medical image classification. *Information* **2024**, *15*, 246. [\[CrossRef\]](#)
43. Bakht, A.B.; Javed, S.; AlMarzouqi, H.; Khandoker, A.; Werghi, N. Colorectal cancer tissue classification using semi-supervised hypergraph convolutional network. In Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 13–16 April 2021; pp. 1306–1309. [\[CrossRef\]](#)
44. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep learning for healthcare: Review, opportunities and challenges. *Brief. Bioinform.* **2018**, *19*, 1236–1246. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Cireşan, D.C.; Giusti, A.; Gambardella, L.M.; Schmidhuber, J. Mitosis detection in breast cancer histology images with deep neural networks. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 411–418. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [\[CrossRef\]](#) [\[PubMed\]](#)
48. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#) [\[PubMed\]](#)
49. Rim, B.; Sung, N.-J.; Min, S.; Hong, M. Deep learning in physiological signal data: A survey. *Sensors* **2020**, *20*, 969. [\[CrossRef\]](#) [\[PubMed\]](#)
50. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. [\[CrossRef\]](#)
51. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
52. Kingma, D.P.; Welling, M. Auto-encoding variational Bayes. *arXiv* **2013**. [\[CrossRef\]](#)
53. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851. [\[CrossRef\]](#)
54. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021. [\[CrossRef\]](#)
55. Yi, X.; Walia, E.; Babyn, P. Generative adversarial network in medical imaging: A review. *Med. Image Anal.* **2019**, *58*, 101552. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Klug, K.; Beckh, K.; Antweiler, D.; Chakraborty, N.; Baldini, G.; Laue, K.; Hosch, R.; Nensa, F.; Schuler, M.; Giesselbach, S. From admission to discharge: A systematic review of clinical natural language processing along the patient journey. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 238. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [\[CrossRef\]](#) [\[PubMed\]](#)
58. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; McDermott, M.B.A. Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 6–7 June 2019; pp. 72–78. [\[CrossRef\]](#)
59. Yang, X.; Chen, A.; PourNejatian, N.; Shin, H.C.; Smith, K.E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A.B.; Flores, M.G.; et al. A large language model for electronic health records. *npj Digit. Med.* **2022**, *5*, 194. [\[CrossRef\]](#) [\[PubMed\]](#)
60. Nori, H.; King, N.; McKinney, S.M.; Carignan, D.; Horvitz, E. Capabilities of GPT-4 on medical challenge problems. *arXiv* **2023**. [\[CrossRef\]](#)
61. Asgari, E.; Montaña-Brown, N.; Dubois, M.; Khalil, S.; Balloch, J.; Au Yeung, J.; Pimenta, D. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digit. Med.* **2025**, *8*, 274. [\[CrossRef\]](#) [\[PubMed\]](#)
62. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474. [\[CrossRef\]](#)
63. Amugongo, L.M.; Mascheroni, P.; Brooks, S.; Doering, S.; Seidel, J. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLoS Digit. Health* **2025**, *4*, e0000877. [\[CrossRef\]](#) [\[PubMed\]](#)
64. Holland, P.W. Statistics and causal inference. *J. Am. Stat. Assoc.* **1986**, *81*, 945–960. [\[CrossRef\]](#)
65. Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009. [\[CrossRef\]](#)
66. Hernán, M.A.; Robins, J.M. *Causal Inference: What If*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2020.
67. Prosperi, M.; Guo, Y.; Sperrin, M.; Koopman, J.S.; Min, J.S.; He, X.; Rich, S.; Wang, M.; Buchan, I.E.; Bian, J. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat. Mach. Intell.* **2020**, *2*, 369–375. [\[CrossRef\]](#)
68. Sanchez, P.; Voisey, J.P.; Xia, T.; Watson, H.I.; O’Neil, A.Q.; Tsiftaris, S.A. Causal machine learning for healthcare and precision medicine. *R. Soc. Open Sci.* **2022**, *9*, 220638. [\[CrossRef\]](#) [\[PubMed\]](#)

69. Feuerriegel, S.; Frauen, D.; Melnychuk, V.; Schweisthal, J.; Hess, K.; Curth, A.; Bauer, S.; Kilbertus, N.; Kohane, I.S.; van der Schaar, M. Causal machine learning for predicting treatment outcomes. *Nat. Med.* **2024**, *30*, 958–968. [[CrossRef](#)] [[PubMed](#)]
70. McKinney, S.M.; Sieniek, M.T.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G.C.; Darzi, A.; et al. International evaluation of an AI system for breast cancer screening. *Nature* **2020**, *577*, 89–94. [[CrossRef](#)] [[PubMed](#)]
71. Lång, K.; Josefsson, V.; Larsson, A.-M.; Larsson, S.; Högberg, C.; Sartor, H.; Hofvind, S.; Andersson, I.; Rosso, A. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): A clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol.* **2023**, *24*, 936–944. [[CrossRef](#)] [[PubMed](#)]
72. Gommers, J.; Hernström, V.; Josefsson, V.; Sartor, H.; Schmidt, D.; Hjelmgren, A.; Larsson, A.-M.; Hofvind, S.; Andersson, I.; Rosso, A.; et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: A randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. *Lancet* **2026**, *407*, 505–514. [[CrossRef](#)] [[PubMed](#)]
73. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **2016**, *316*, 2402–2410. [[CrossRef](#)] [[PubMed](#)]
74. De Fauw, J.; Ledsam, J.R.; Romera-Paredes, B.; Nikolov, S.; Tomasev, N.; Blackwell, S.; Askham, H.; Glorot, X.; O'Donoghue, B.; Visentin, D.; et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **2018**, *24*, 1342–1350. [[CrossRef](#)] [[PubMed](#)]
75. Abràmoff, M.D.; Lavin, P.T.; Birch, M.; Shah, N.; Folk, J.C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digit. Med.* **2018**, *1*, 39. [[CrossRef](#)] [[PubMed](#)]
76. Campanella, G.; Hanna, M.G.; Geneslaw, L.; Mirafior, A.; Werneck Krauss Silva, V.; Busam, K.J.; Brogi, E.; Reuter, V.E.; Klimstra, D.S.; Fuchs, T.J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **2019**, *25*, 1301–1309. [[CrossRef](#)] [[PubMed](#)]
77. Echle, A.; Rindtorff, N.T.; Brinker, T.J.; Luedde, T.; Pearson, A.T.; Kather, J.N. Deep learning in cancer pathology: A new generation of clinical biomarkers. *Br. J. Cancer* **2021**, *124*, 686–696. [[CrossRef](#)] [[PubMed](#)]
78. Bera, K.; Schalper, K.A.; Rimm, D.L.; Velcheti, V.; Madabhushi, A. Artificial intelligence in digital pathology—New tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 703–715. [[CrossRef](#)] [[PubMed](#)]
79. Goldstein, B.A.; Navar, A.M.; Pencina, M.J.; Ioannidis, J.P.A. Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *J. Am. Med. Inform. Assoc.* **2017**, *24*, 198–208. [[CrossRef](#)] [[PubMed](#)]
80. Hannun, A.Y.; Rajpurkar, P.; Haghpanahi, M.; Tison, G.H.; Bourn, C.; Turakhia, M.P.; Ng, A.Y. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **2019**, *25*, 65–69. [[CrossRef](#)] [[PubMed](#)]
81. Wong, A.; Otles, E.; Donnelly, J.P.; Krumm, A.; McCullough, J.; DeTroyer-Cooley, O.; Pesttrue, J.; Phillips, M.; Konye, J.; Penzo, C.; et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.* **2021**, *181*, 1065–1070. [[CrossRef](#)] [[PubMed](#)]
82. Lyons, P.G.; Hofferford, M.R.; Yu, S.C.; Michelson, A.P.; Payne, P.R.O.; Hough, C.L.; Singh, K. Factors associated with variability in the performance of a proprietary sepsis prediction model across 9 networked hospitals in the US. *JAMA Intern. Med.* **2023**, *183*, 611–612. [[CrossRef](#)] [[PubMed](#)]
83. Rossetti, S.C.; Dykes, P.C.; Knaplund, C.; Cho, S.; Withall, J.; Lowenthal, G.; Albers, D.; Lee, R.Y.; Jia, H.; Bakken, S.; et al. Real-time surveillance system for patient deterioration: A pragmatic cluster-randomized controlled trial. *Nat. Med.* **2025**, *31*, 1895–1902. [[CrossRef](#)] [[PubMed](#)]
84. Hassan, C.; Spadaccini, M.; Mori, Y.; Foroutan, F.; Facciorusso, A.; Gkolfakis, P.; Tziatzios, G.; Triantafyllou, K.; Antonelli, G.; Khalaf, K.; et al. Real-time computer-aided detection of colorectal neoplasia during colonoscopy: A systematic review and meta-analysis. *Ann. Intern. Med.* **2023**, *176*, 1209–1220. [[CrossRef](#)] [[PubMed](#)]
85. Zimmermann-Fraedrich, K.; Sehner, S.; Rösch, T.; Aschenbeck, J.; Schubert, S.; Liceni, T.; Moog, G.; Neumann, H.; Berndt, R.; Weigt, J.; et al. No effect of computer-aided diagnosis on colonoscopic adenoma detection in a large pragmatic multicenter randomized study. *Am. J. Gastroenterol.* **2026**, *121*, 881–889. [[CrossRef](#)] [[PubMed](#)]
86. Muehlematter, U.J.; Daniore, P.; Vokinger, K.N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–2020): A comparative analysis. *Lancet Digit. Health* **2021**, *3*, e195–e203. [[CrossRef](#)] [[PubMed](#)]
87. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Benetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [[CrossRef](#)]
88. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [[CrossRef](#)] [[PubMed](#)]

89. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774. [\[CrossRef\]](#)
90. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should I trust you?” Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144. [\[CrossRef\]](#)
91. Selvaraju, R.R.; 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, Venice, Italy, 22–29 October 2017; pp. 618–626. [\[CrossRef\]](#)
92. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [\[CrossRef\]](#) [\[PubMed\]](#)
93. Ghassemi, M.; Oakden-Rayner, L.; Beam, A.L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **2021**, *3*, e745–e750. [\[CrossRef\]](#) [\[PubMed\]](#)
94. Habli, I.; Lawton, T.; Porter, Z. Artificial intelligence in health care: Accountability and safety. *Bull. World Health Organ.* **2020**, *98*, 251–256. [\[CrossRef\]](#) [\[PubMed\]](#)
95. National Research Council. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*; National Academies Press: Washington, DC, USA, 2011. [\[CrossRef\]](#) [\[PubMed\]](#)
96. Ziegelstein, R.C. Personomics: The missing link in the evolution from precision medicine to personalized medicine. *J. Pers. Med.* **2017**, *7*, 11. [\[CrossRef\]](#) [\[PubMed\]](#)
97. Schork, N.J. Artificial intelligence and personalized medicine. *Cancer Treat. Res.* **2019**, *178*, 265–283. [\[CrossRef\]](#) [\[PubMed\]](#)
98. Johnson, K.B.; Wei, W.Q.; Weeraratne, D.; Frisse, M.E.; Misulis, K.; Rhee, K.; Zhao, J.; Snowdon, J.L. Precision medicine, AI, and the future of personalized health care. *Clin. Transl. Sci.* **2021**, *14*, 86–93. [\[CrossRef\]](#) [\[PubMed\]](#)
99. Stenzinger, A.; Moltzen, E.K.; Winkler, E.; Molnar-Gabor, F.; Malek, N.; Costescu, A.; Jensen, B.N.; Nowak, F.; Pinto, C.; Ottersen, O.P.; et al. Implementation of precision medicine in healthcare—A European perspective. *J. Intern. Med.* **2023**, *294*, 437–454. [\[CrossRef\]](#) [\[PubMed\]](#)
100. Liu, M.; Vnencak-Jones, C.L.; Roland, B.P.; Gatto, C.L.; Mathe, J.L.; Just, S.L.; Peterson, J.F.; Van Driest, S.L.; Weitkamp, A.O. A tutorial for pharmacogenomics implementation through end-to-end clinical decision support based on ten years of experience from PREDICT. *Clin. Pharmacol. Ther.* **2021**, *109*, 101–115. [\[CrossRef\]](#) [\[PubMed\]](#)
101. Khoury, M.J.; Bowen, S.; Dotson, W.D.; Drzymalla, E.; Green, R.F.; Goldstein, R.; Kolor, K.; Liburd, L.C.; Sperling, L.S.; Bunnell, R. Health equity in the implementation of genomics and precision medicine: A public health imperative. *Genet. Med.* **2022**, *24*, 1630–1639. [\[CrossRef\]](#) [\[PubMed\]](#)
102. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 121–127. [\[CrossRef\]](#) [\[PubMed\]](#)
103. Park, S.H.; Han, K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* **2018**, *286*, 800–809. [\[CrossRef\]](#) [\[PubMed\]](#)
104. Hogg, H.D.J.; Gunatilleke, J.; Maniopoulos, G. Editorial: Place-based evidence for clinical artificial intelligence implementation. *Front. Health Serv.* **2025**, *5*, 1736656. [\[CrossRef\]](#) [\[PubMed\]](#)
105. Reddy, S.; Rogers, W.; Makinen, V.-P.; Coiera, E.; Brown, P.; Wenzel, M.; Weicken, E.; Ansari, S.; Mathur, P.; Casey, A.; et al. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform.* **2021**, *28*, e100444. [\[CrossRef\]](#) [\[PubMed\]](#)
106. Callahan, A.; McElfresh, D.C.; Banda, J.M.; Bunney, G.; Char, D.; Chen, J.H.; Corbin, C.K.; Dash, D.; Downing, N.L.; Jain, S.S.; et al. Standing on FURM Ground: A Framework for Evaluating Fair, Useful, and Reliable AI Models in Health Care Systems. *NEJM Catal. Innov. Care Deliv.* **2024**, *5*. [\[CrossRef\]](#)
107. Lekadir, K.; Frangi, A.F.; Porras, A.R.; Glocker, B.; Cintas, C.; Langlotz, C.P.; Weicken, E.; Asselbergs, F.W.; Prior, F.; Collins, G.S.; et al. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **2025**, *388*, e081554. [\[CrossRef\]](#) [\[PubMed\]](#)
108. Wells, B.J.; Nguyen, H.M.; McWilliams, A.; Pallini, M.; Bovi, A.; Kuzma, A.; Kramer, J.; Chou, S.-H.; Hetherington, T.; Corn, P.; et al. A practical framework for appropriate implementation and review of artificial intelligence (FAIR-AI) in healthcare. *npj Digit. Med.* **2025**, *8*, 514. [\[CrossRef\]](#) [\[PubMed\]](#)
109. Abramoff, M.D.; Roehrenbeck, C.; Trujillo, S.; Goldstein, J.; Graves, A.S.; Repka, M.X.; Silva, E.Z., III. A reimbursement framework for artificial intelligence in healthcare. *npj Digit. Med.* **2022**, *5*, 72. [\[CrossRef\]](#) [\[PubMed\]](#)
110. Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **2022**, *28*, 924–933. [\[CrossRef\]](#) [\[PubMed\]](#)
111. Friedman, C.P.; Wong, A.K.; Blumenthal, D. Achieving a nationwide learning health system. *Sci. Transl. Med.* **2010**, *2*, 57cm29. [\[CrossRef\]](#) [\[PubMed\]](#)

112. Nair, M.; Nygren, J.; Nilsen, P.; Gama, F.; Neher, M.; Larsson, I.; Svedberg, P. Critical activities for successful implementation and adoption of AI in healthcare: Towards a process framework for healthcare organizations. *Front. Digit. Health* **2025**, *7*, 1550459. [[CrossRef](#)] [[PubMed](#)]
113. European Parliament and Council of the European Union. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices. *Off. J. Eur. Union* **2017**, *L 117*, 1–175.
114. van Kolfschooten, H.; van Oirschot, J. The EU Artificial Intelligence Act (2024): Implications for healthcare. *Health Policy* **2024**, *149*, 105152. [[CrossRef](#)] [[PubMed](#)]
115. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*; World Health Organization: Geneva, Switzerland, 2021.
116. Wolff, J.; Pauling, J.; Keck, A.; Baumbach, J. The economic impact of artificial intelligence in health care: Systematic review. *J. Med. Internet Res.* **2020**, *22*, e16866. [[CrossRef](#)] [[PubMed](#)]
117. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, *L 2024/1689*. Available online: <https://www.boe.es/buscar/doc.php?id=DOUE-L-2024-81079#1> (accessed on 23 September 2026).
118. Wahl, B.; Cossy-Gantner, A.; Germann, S.; Schwalbe, N.R. Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings? *BMJ Glob. Health* **2018**, *3*, e000798. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.