SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

Translating Super Intelligence to Clinical Practice: Methodological Challenges and Safety Limits in Autonomous Healthcare

Parhlad Singh Ahluwalia[1]
[1] *Dr. Bhim Rao Ambedkar Law University, Rajasthan, India*
Correspondence: ahluwalia002@zohomail.in

Abstract

Background: *Public and professional discourse increasingly describes medical artificial intelligence (AI) in the vocabulary of "superintelligence", implying that clinical autonomy is a matter of engineering time rather than of evidence. This framing is rarely tested against the methodological quality of the studies on which it rests.*

Objective: *To appraise, in a deliberately sceptical manner, the methodological obstacles and the safety limits that separate benchmark-level machine performance from accountable, autonomous clinical care.*

Methods: *A structured narrative synthesis was conducted, reported in line with the principles of PRISMA 2020 and its scoping-review extension (PRISMA-ScR). Two evidence streams were combined: a supplied bibliographic corpus of 117 records, which was screened to 41 topically eligible items, and 47 purposively selected external sources comprising systematic reviews, trials, reporting standards, regulatory instruments and safety-science literature. Findings were organised into evidence quality, validity, human-factor, equity, governance and liability domains.*

Results: *Across the benchmark literature, headline accuracy was repeatedly accompanied by retrospective designs, limited external validation, high risk of bias, scarce prospective evaluation and weak reporting. Documented deployments revealed failures that laboratory metrics did not predict, including poor sensitivity of a widely implemented sepsis alert, proxy-driven racial bias in a population-health algorithm and workflow friction in a field-deployed retinal screening tool. Human oversight was found to be a fragile safeguard because of automation bias and alert fatigue. Only a narrow class of tightly bounded tasks currently has trial-grade evidence supporting autonomous operation.*

Conclusions: *Autonomy in healthcare should be earned incrementally, task by task, through prospective, externally validated, equity-audited and legally accountable evaluation. A five-level Clinical Autonomy Ladder and an evidence-gated promotion protocol are proposed. Claims of superintelligent medicine are, at present, aspirations about the future and not findings about the present.*

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

Keywords: *artificial intelligence; autonomous systems; clinical safety; patient safety; algorithmic bias; automation bias; health governance; evidence synthesis; PRISMA*

1. Introduction

Medicine has always borrowed its metaphors from the technology of its era. The clock-work body of the seventeenth century gave way to the body as a thermodynamic engine, and later to the body as an information-processing network. The metaphor currently in circulation is that of the machine mind, and with it arrives a vocabulary that is unusually grand. Commentators speak of systems that will out-diagnose, out-prescribe and eventually out-reason the clinician, and the word “superintelligence” is used, sometimes carefully and often carelessly, to describe the destination [1,2]. The rhetorical force of the term is considerable. If an intelligence exceeds human competence in essentially every cognitive domain, then it seems perverse to insist on human supervision of the bedside, and the argument for clinical autonomy acquires an air of inevitability.

This paper contends that the inevitability is manufactured. Between a laboratory benchmark and an accountable clinical decision there lies a long chain of inferential steps, and each step is vulnerable to a distinct class of error. Data must represent the population that will be treated. Labels must represent clinical truth rather than clinical habit. Performance estimates must survive transport to new hospitals, new scanners, new patient mixes and new documentation cultures. Outputs must be interpretable by the humans who remain responsible, and the surrounding organisation must be capable of detecting drift, investigating harm and assigning liability. Where the chain is intact, autonomy may be defensible. Where it is broken, performance figures that look impressive on paper can conceal real danger.

The literature on medical AI has produced striking demonstrations of narrow competence. Deep learning systems have matched specialists in grading retinal photographs, classifying skin lesions and reading screening mammograms [3–5]. Large language models have achieved passing-level scores on licensing-style examinations [6] and have prompted serious discussion of their benefits and limits in clinical conversation [7,8]. Yet the same literature contains a sobering counter-narrative. Systematic appraisals have found that many studies claiming clinician-level performance are retrospective, poorly reported and at high risk of bias [9,10]. Hundreds of prediction models developed during the COVID-19 pandemic proved unfit for clinical use because of methodological flaws [11,12]. Post-deployment audits have exposed systematic harm that no internal validation anticipated [13,14].

Scholarship from within the supplied corpus reinforces the need for caution from a different angle. Work on AI in diagnostic medicine, patient engagement in AI-driven systems and machine-learning-supported hospital operations treats these technologies as socio-technical interventions whose success depends on organisational readiness as much as on algorithmic skill [15–17]. Related analyses of digital transformation in hospital management and of telemedicine equity emphasise that access, infrastructure and governance shape who benefits from digital health [18,19]. A complementary body of sociological writing argues that algorithmic systems are not neutral instruments but reproduce and amplify existing

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

stratification [20,21]. These strands rarely converge with the safety-science and trial-methodology literatures, and the present paper attempts to bring them into one analytical frame.

The aim is therefore threefold. The first aim is conceptual: to separate three things that are habitually blurred, namely narrow clinical AI, general-purpose foundation models and the hypothetical artificial superintelligence of the safety literature. The second is methodological: to examine how the evidence for autonomous clinical performance is generated, reported and generalised, and where it systematically fails. The third is practical: to specify the safety limits that should govern any movement from decision support toward autonomous operation, and to propose a graduated, evidence-gated pathway for crossing them.

The paper makes four contributions. It provides a transparent, PRISMA-guided synthesis that integrates a supplied regional and sociological corpus with external benchmark evidence. It presents a comparative evidence table that juxtaposes headline claims with the methodological limits of the studies that produced them. It identifies a set of cross-cutting failure mechanisms, including specification gaming, distribution shift, label contamination, automation bias and accountability diffusion, and maps each to a concrete safeguard. Finally, it proposes the Clinical Autonomy Ladder, a five-level classification coupled with explicit promotion criteria, intended to give regulators, hospital leaders and developers a shared vocabulary.

2. Conceptual Foundations: From Narrow Competence to the Superintelligence Frame

2.1 What "superintelligence" does and does not denote

In the safety literature the term has a specific and demanding meaning. Bostrom defined superintelligence as an intellect that greatly exceeds the best human minds across virtually all domains of interest, and he distinguished speed, collective and quality forms [1]. Nothing in contemporary clinical practice meets that definition. The systems that currently enter hospitals are narrow in scope even when they are technically sophisticated, and the general-purpose language models that appear more flexible still fail unpredictably outside the distributions on which they have been trained and tuned [7,22]. When medical commentary invokes superintelligence, it generally means something much weaker, namely a system that performs one bounded task at or above the level of a typical specialist. Calling such a system superintelligent is a category error that carries practical consequences, because it licenses expectations of robustness, generality and self-correction that the system does not possess.

It is nevertheless legitimate to ask what would follow if capability continued to rise. The safety literature offers a structured answer. Russell argues that the central difficulty lies not in malevolence but in objective misspecification: a capable optimiser pursuing a subtly wrong objective can cause harm precisely because it is competent [23]. Amodei and colleagues set out a taxonomy of concrete problems that already arise in present-day systems, including negative side effects, reward hacking, limited oversight, unsafe exploration and fragility under distributional shift [24]. Hendrycks and colleagues organise catastrophic risk into malicious use, competitive pressure, organisational failure and loss of control [25], and an international group of senior researchers has argued that governance has not kept pace with capability growth

[26]. For the purposes of clinical translation, the value of this literature lies less in its long-term scenarios than in its diagnostic vocabulary. Each of the concrete problems has a recognisable clinical analogue, as the later sections demonstrate.

2.2 Three tiers of medical AI

A precise vocabulary is a precondition for honest evaluation. This paper distinguishes three tiers. Tier 1 comprises narrow, task-specific systems such as image classifiers, risk scores and alerting algorithms, whose input space and output space are tightly defined. Tier 2 comprises general-purpose foundation models, including large language models and multimodal systems, whose output space is open-ended and whose failure modes include fluent fabrication [8,22]. Tier 3 denotes hypothetical artificial superintelligence and is treated here strictly as a scenario that motivates precaution and not as an empirical object. The distinction matters because the evidentiary standards, the failure modes and the appropriate degree of oversight differ across tiers. A diabetic retinopathy classifier evaluated in a prospective pivotal trial occupies a different epistemic position from a conversational model that has scored well on multiple-choice examinations [6,27].

2.3 Autonomy is a property of a task and a context, not of a model

A second conceptual correction concerns the unit of analysis. Autonomy is often discussed as if it were an attribute of a model, but in clinical practice it is an attribute of a task embedded in a workflow, a legal regime and a staffing pattern. The same classifier can be advisory in one hospital, because a specialist reviews every output, and effectively autonomous in another, because an overstretched primary-care clinic acts on its output unreviewed. Work on clinical decision support has long recognised that benefits and risks depend heavily on local implementation [28,29]. Research on healthcare leadership and management similarly stresses that structures of authority, escalation and organisational learning determine whether technical tools improve or degrade care [30–32]. The consequence is that safety claims cannot be inherited from a model card. They must be re-established for each task, population and site.

2.4 The translational gap

The distance between benchmark performance and clinical benefit can be decomposed into at least five gaps. The first is the data gap, in which training data under-represent the patients, devices or settings of deployment. The second is the label gap, in which the recorded outcome is a proxy shaped by access, practice and measurement and not the clinical truth. The third is the validation gap, in which retrospective and internal testing substitutes for prospective, external and impact-oriented evaluation. The fourth is the integration gap, in which a correct output is delivered at the wrong time, to the wrong person, or in a form that cannot be acted upon. The fifth is the accountability gap, in which no identifiable party is responsible when harm occurs. Kelly and colleagues emphasised a closely related set of barriers to clinical impact, including evaluation, generalisability, usability and regulation [33]. The remainder of this paper uses the five-gap structure as an organising device.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

3. Methods

3.1 Design and reporting framework

This study is a structured narrative review with scoping-review features. It follows the reporting logic of the PRISMA 2020 statement for transparent description of identification, screening, eligibility and inclusion [34], and it borrows from the PRISMA-ScR extension the emphasis on mapping the breadth of a heterogeneous evidence base and not on pooling effect sizes [35]. The decision to adopt a scoping orientation rests on the heterogeneity of the material, which spans randomised trials, retrospective validations, regulatory instruments, conceptual treatises and sociological analyses. Practical guidance on conducting and reporting PRISMA-ScR reviews in emerging research settings informed the structuring of the process [36,37].

The review was not prospectively registered, and no formal protocol was published. It is therefore not a systematic review in the strict sense, and the findings should be interpreted as a critically organised synthesis and not as an exhaustive census. This limitation is stated at the outset because transparency about method is itself a safeguard against the overclaiming that the paper criticises.

3.2 Evidence streams

Stream A: the supplied corpus. A reference corpus of 117 records was supplied for use in this work. Each record was examined at the bibliographic level, that is, by title, author list, venue and year, because full texts were not available for appraisal. Records were excluded if publication details could not be verified (n = 2, both lacking a year or venue) or if their subject matter had no demonstrable bearing on clinical AI, digital health, health-system governance, algorithmic society, legal accountability or research integrity (n = 74). Forty-one records were retained. Topics removed at this stage included, for example, contemplative practice, higher-education accreditation, hydrological hazards unrelated to health systems and cultural studies. This exclusion was a deliberate choice to avoid citation padding, because including thematically unrelated sources would inflate the reference count without strengthening the argument.

Stream B: external benchmark sources. Because the supplied corpus contains almost no primary clinical-AI trials and no safety-science literature, it was complemented by 47 externally published sources selected purposively against four criteria: peer-reviewed publication or authoritative institutional status; direct relevance to autonomy, validity, safety, bias or governance of clinical AI; methodological weight, giving priority to systematic reviews, randomised or pivotal trials, external validations and formal reporting standards; and diversity of perspective across clinical, computational, regulatory and ethical viewpoints. No systematic database search strings were executed for Stream B, and its selection is therefore vulnerable to the familiar biases of expert curation. This is acknowledged in the limitations.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

3.3 Screening flow

The flow of records is summarised in Figure 1.

Figure 1. PRISMA-style flow of records.

Stream A (supplied corpus). Records identified: 117 → records removed for unverifiable publication details: 2 → records screened by title and venue: 115 → records excluded as topically irrelevant: 74 → records included in synthesis: 41.
Stream B (purposive external sources). Sources selected against pre-stated criteria: 47 → sources included in synthesis: 47.
Combined evidence base: 88 sources.

3.4 Data charting and appraisal

For every included source, the following were charted where applicable: design, population or domain, principal quantitative result, principal limitation and relevance to one of the five translational gaps. Methodological appraisal drew on established instruments. Risk of bias and applicability in prediction-model studies were considered with reference to the PROBAST framework [38], while the completeness of reporting in trials and prediction studies was judged against the CONSORT-AI, TRIPOD+AI and DECIDE-AI guidance [39–41]. No numerical risk-of-bias scores were generated for individual studies by the present author; instead, the appraisals reported by the original systematic reviews were relied on and are attributed to them. Records in Stream A were not appraised for internal validity, and they are used for conceptual framing, contextual evidence and sociolegal analysis and not as sources of effect estimates.

3.5 Synthesis and integrity safeguards

Findings were synthesised narratively and organised by translational gap. Where several sources reported on the same construct, the results were juxtaposed in a comparative table so that agreement and divergence could be inspected directly. Quantitative values are reported exactly as given by the original authors and are never pooled across non-comparable designs. Research-integrity safeguards followed accepted practice for ethical scholarship [42]. All prose was composed originally, all attributions are numbered, and standard guidance on plagiarism prevention and detection was observed [43–45]. Authors who submit the manuscript are advised to run a formal similarity check before submission, because originality cannot be certified by drafting alone.

3.6 A note on corpus composition

A reality check on the supplied corpus is warranted. Of the 115 verifiable records, 113 share a single lead or co-author, which indicates a highly concentrated citation network. Such concentration is not inherently problematic in a scholarly programme that develops over time, but it limits the diversity of viewpoints and may create an echo effect. The external stream was designed in part to counter this concentration, and the reader should weigh the two streams accordingly.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

4. Results

4.1 Characteristics of the evidence base

The combined evidence base comprised 88 sources. Stream B contributed the quantitative and methodological core: systematic and critical reviews of study quality, reports of deployment and prospective evaluation, reporting and appraisal standards, safety-science treatises, regulatory and ethics instruments, and landmark technical demonstrations and human-factors studies. Stream A contributed predominantly conceptual, contextual and sociolegal material, with only a small number of items self-described as systematic reviews or meta-analyses [18,46,47]. This asymmetry is itself a finding. The strongest quantitative evidence on autonomy resides in the external stream, while the supplied corpus is strongest on organisational, equity and legal context.

4.2 Headline performance versus methodological quality

The first consistent result is a gap between the confidence of performance claims and the quality of the designs that support them. Table 1 juxtaposes representative benchmark studies and systematic appraisals. Four patterns emerge from it.

Table 1. Related reviews and benchmark studies: design, principal results and implications for autonomy

Source	Design and scope	Principal result (as reported by the original authors)	Implication for autonomous use
Nagendran et al. [9]	Systematic review of deep-learning studies compared with clinicians	Among 81 non-randomised studies only 9 were prospective and 6 were tested in real-world settings; 58 were at high risk of bias; data and code were unavailable in about 95% and 93% of studies respectively	Claims of clinician-level performance rest on weak designs and cannot be independently reproduced
Liu et al. [10]	Systematic review and meta-analysis of deep learning versus health-care professionals in imaging	82 studies reviewed, 69 meta-analysed; only 14 compared algorithm and clinicians on the same sample; pooled sensitivity about 87.0% vs 86.4% and specificity about 92.5% vs 90.5%	Parity is plausible for narrow imaging tasks, but head-to-head evidence is scarce
Wynants et al. [11]	Living systematic review of COVID-19 prediction models	232 models from 169 studies; all judged at high or unclear risk of bias	Crisis-time urgency can degrade methodological discipline
Roberts et al. [12]	Systematic review of machine-learning COVID-19 imaging models	2,212 studies identified, 415 retained after initial screening, 62 after quality assessment; none judged of potential clinical use	Quality filters eliminate nearly all apparent evidence
Bedi et al. [48]	Systematic review of evaluations of large language	519 studies; about 5% used real patient-care data	Most language-model evidence is exam-style and not clinical

	models in health care		
Goh et al. [49]	Randomised trial of physicians with and without GPT-4 access	50 physicians; median diagnostic-reasoning score 76% with the model vs 74% without; the model alone scored about 92%	Machine accuracy did not translate into better human-machine performance
Abràmoff et al. [27]	Pivotal prospective trial of an autonomous diabetic retinopathy system in primary care	819 participants analysed; sensitivity 87.2% and specificity 90.7%	A bounded autonomous task with a prospective trial and a pre-specified threshold
McKinney et al. [5]	International retrospective evaluation of breast-screening AI	Absolute reductions in false positives of 5.7% (US) and 1.2% (UK) and in false negatives of 9.4% (US) and 2.7% (UK)	Strong signal, but retrospective and dependent on local reading practice
Wong et al. [14]	External validation of a widely implemented proprietary sepsis model	38,455 hospitalisations; discrimination of about 0.63; sensitivity about 33% at the alert threshold; alerts in about 18% of hospitalisations	Widely deployed does not mean validated
Obermeyer et al. [13]	Audit of an algorithm used to allocate care-management resources	Used for roughly 200 million people annually; correcting the bias would raise the proportion of Black patients flagged for extra help from 17.7% to 46.5%	A proxy label can embed structural inequity at national scale
Beede et al. [50]	Human-centred field evaluation of a deep-learning retinal screening tool in clinics	Qualitative findings: lighting, image quality, connectivity and workflow constraints reduced usability	Laboratory accuracy is not clinical utility
Komorowski et al. [51]	Retrospective reinforcement-learning policy for sepsis treatment	Policy derived from historical records and evaluated off-policy; no prospective test	Learning from observational data cannot by itself establish causal benefit

Note: Values are reproduced from the cited publications and are not pooled. Designs are not directly comparable.

Pattern 1: retrospective evidence dominates. The review of Nagendran and colleagues is the clearest warning. Of the 81 non-randomised studies, fewer than one in eight was prospective, and fewer than one in ten had been tested in a real clinical setting, while open access to data and code was exceptional [9]. Evidence of this kind can generate hypotheses but is poorly suited to licensing autonomy.

Pattern 2: apparent parity masks thin comparison. The meta-analysis by Liu and colleagues found that pooled diagnostic accuracy of deep learning was similar to that of health-care professionals, yet only a small fraction of the reviewed studies had compared both on the same

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
Vol. 4, Issue 5, November 2026
Available Online : https://shodhprakashan.in/

sample, and reporting was heterogeneous [10]. The result supports the proposition that narrow imaging tasks can be performed at specialist level. It does not support the more sweeping proposition that machines have equalled clinicians as clinicians.

Pattern 3: pandemic evidence shows how quickly standards erode. When the need was greatest, quality fell. Of 232 COVID-19 prediction models, all were at high or unclear risk of bias [11], and in a separate review of machine-learning models applied to chest imaging, a pool of thousands of studies was reduced by quality screening to a few dozen with none of potential clinical use [12]. The lesson is not that machine learning cannot help in a crisis. It is that crisis conditions demand more, not less, methodological discipline, a point echoed in analyses of organisational learning from pandemic failures in healthcare [31].

Pattern 4: performance in the laboratory does not predict performance with humans. In the randomised trial by Goh and colleagues, the language model alone outperformed physicians who used it, and physicians given access to the model did not significantly outperform those without it [49]. This result is consequential for the superintelligence narrative because it implies that a capability advantage does not automatically become a clinical advantage. Integration, interface design, trust calibration and training mediate the relationship.

4.3 Documented deployment failures

Three well-documented cases illustrate the translational gap in practice.

The first is the proprietary sepsis prediction model. External validation across tens of thousands of hospitalisations found modest discrimination, low sensitivity at the operating threshold and a high alert burden, which is the combination most likely to produce missed cases alongside alert fatigue [14]. The case demonstrates a data gap and a validation gap simultaneously, since vendor-reported performance had not been reproduced in independent evaluation.

The second is the population-health algorithm audited by Obermeyer and colleagues. The system predicted future healthcare cost as a proxy for health need. Because Black patients incurred lower costs at the same level of illness, owing to unequal access to care, the proxy systematically under-identified their needs [13]. The failure was not a coding error. It was a label gap in which a seemingly neutral target variable transported historical inequity into an automated decision. Broader surveys of bias and fairness in machine learning confirm that such mechanisms are general and not exceptional [52,53]. Sociological work in the supplied corpus on algorithmic bias, platform invisibility and spatial inequality in technology-mediated cities supports the same structural reading from outside medicine [20,21,54].

The third is the field deployment of a retinal screening tool in clinics, where nurses working under real constraints encountered image-quality rejections and connectivity problems that the development environment had not captured [50]. This is an integration gap. Its significance lies in showing that a system with acceptable laboratory accuracy can still impose delays and exclusions on patients when the surrounding workflow is not designed with it.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

4.4 Where autonomy has genuine support

A balanced appraisal must also record where evidence is stronger. The autonomous diabetic retinopathy system evaluated in primary care is instructive precisely because it satisfied conditions that most systems do not: a narrow, well-defined task; a prospective pivotal design; pre-specified performance targets; a defined referral pathway for positive cases; and an explicit rule that ungradable images are escalated to humans [27]. Regulatory trackers show that approvals of AI-enabled devices have increased, though pathways and transparency differ between jurisdictions [55]. The reality exposed here is that autonomy has been justified in a thin slice of medicine, characterised by bounded input, bounded output, cheap error detection and a safe failure mode. It has not been justified for open-ended, multi-step, high-consequence decisions.

4.5 Thematic results from the supplied corpus

The forty-one retained items from Stream A cluster into seven themes. Their contribution is contextual and is summarised below, with the caveat that appraisal was bibliographic.

Clinical and operational AI. Items on AI in diagnostic medicine, patient engagement in AI-driven systems, machine-learning-supported facility management and AI-driven prediction of antimicrobial resistance frame AI as a set of applications that span diagnosis, operations and stewardship [15–17,56]. The antimicrobial resistance work is illustrative because it links genomic epidemiology to prescribing decisions, a domain where an erroneous autonomous recommendation could propagate across a population through selection pressure, and where the cost of error is therefore externalised [56].

Health-system capacity and equity. Work on digital transformation in hospitals and on telemedicine equity highlights infrastructure, workforce and socioeconomic gradients as determinants of benefit [18,19]. Socio-legal analysis of health disparities and epidemiological analyses of the disease burden and changing disease patterns in India situate any autonomous system within a context of uneven access and a dual burden of communicable and non-communicable disease [57–59].

Safety culture and leadership. Items on hospital leadership, crisis leadership, patient-safety strategies for infection reduction, evidence-based environmental design and trauma-informed organisational culture converge on the claim that safety is produced by systems and cultures and not by devices alone [30,31,60–62]. This supports the later argument that oversight must be designed as an organisational function.

Resilience and infrastructure. Analyses of flood impacts on health and healthcare infrastructure and of vulnerability and adaptation emphasise that physical and digital infrastructure can fail in the same event, with direct consequences for any system dependent on connectivity or compute [63,64].

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

Law and accountability. Items on health legislation, negligence, the sociology of law and judicial interpretation provide the doctrinal and theoretical background for the liability analysis in Section 7 [65–68].

Algorithmic society and geopolitics. Work on techno-authoritarianism, cyber-conflict, domestic automation, decolonial epistemology and technology-driven geopolitical competition widens the lens to questions of data sovereignty, surveillance, unequal labour effects and the global distribution of AI capability [69–73].

Evidence, integrity and human factors. Items on scoping-review methodology, research ethics, plagiarism and AI-supported integrity, together with analyses of cognitive overload, workplace disengagement and stigma, bear on the quality of the evidence base and on the human conditions under which clinicians and patients interact with automated systems [36,37,42,45,74–76].

5. Related Research Reviews, Their Outputs and Analytical Perspectives

5.1 Why a review-of-reviews layer is necessary

A paper that claims to define authentic research on autonomous healthcare must show how its conclusions relate to what earlier reviews have already established. The reviews summarised in Table 1 are the most relevant anchors. Their outputs fall into four families: critiques of study quality, comparisons of machine and human performance, audits of deployed systems, and investigations of human-machine interaction. Complementary reviews in the supplied corpus address digital transformation of hospital management, trauma-related outcomes in conflict settings and cardiovascular outcomes of a behavioural intervention, which are not clinical-AI reviews but exemplify how systematic and meta-analytic designs are used in adjacent health fields [18,46,47].

5.2 Review family A: critiques of study quality

The reviews of Nagendran, Wynants and Roberts converge on one message. Studies were frequently retrospective, derived from convenience datasets, lightly validated and poorly reported [9,11,12]. The remedy proposed by the field has been to strengthen reporting and appraisal standards. TRIPOD+AI updates guidance for prediction models that use machine learning, CONSORT-AI extends trial reporting to AI interventions, DECIDE-AI addresses early live clinical evaluation, and PROBAST provides a structured tool for judging risk of bias and applicability [38–41]. The practical output of these reviews is therefore not a verdict against AI but an evidentiary floor.

An analytical observation is warranted. Reporting standards improve transparency but cannot by themselves guarantee validity. A study can adhere to every reporting item and still use a flawed label or an unrepresentative sample. Reporting guidelines are necessary conditions for scrutiny, not sufficient conditions for trust.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

5.3 Review family B: machine versus clinician comparisons

The review by Liu and colleagues and the primary studies in dermatology, ophthalmology and breast screening show that, for tightly defined perceptual tasks, algorithms can reach specialist-level discrimination [3–5,10]. The output of this family is a credible proof of capability on narrow tasks. Two caveats recur. The first is that comparison conditions often disadvantage clinicians, who are tested without clinical context or prior records. The second is that sensitivity and specificity at a laboratory threshold do not capture downstream consequences such as overdiagnosis, referral burden or the psychological impact of false positives.

5.4 Review family C: audits of deployed systems

Audits are the most epistemically valuable category because they test systems under real conditions. The sepsis-model validation and the population-health algorithm audit both showed that deployment can reveal failure modes invisible to developers [13,14]. Challen and colleagues catalogued sources of clinical safety risk in machine-learning systems, including distributional shift, insensitivity to impact, black-box decision-making, unsafe failure modes and the risk of automation complacency [77]. The output of this family is a taxonomy of hazards and, more importantly, a demonstration that external, independent, post-deployment evaluation must be treated as a standing obligation rather than an optional extra.

5.5 Review family D: human-machine interaction

Syntheses of automation bias show that clinicians and other professionals tend to over-rely on automated advice, accept incorrect recommendations and fail to notice failures when the automation is usually reliable [78,79]. Reviews of decision-support systems report that alert fatigue and poor integration blunt benefits and create new risks [28]. The randomised trial by Goh and colleagues demonstrates that the problem is not merely theoretical, since access to a strong model did not materially improve physician performance [49]. The output of this family is a challenge to the assumption that a human in the loop is a sufficient safeguard.

5.6 Analysis from different perspectives

An authentic appraisal requires triangulation across perspectives that do not always agree. Six are considered.

The clinical perspective. Clinicians value tools that reduce workload and error, but they prioritise reliability at the margins. A tool that is right 95% of the time and fails silently the rest of the time may be less useful than a weaker tool that signals its uncertainty. Clinical safety science therefore emphasises failure modes and escalation, not mean accuracy [33,77].

The computational perspective. Machine-learning researchers focus on generalisation, robustness and scaling. They emphasise adversarial vulnerability, in which imperceptible perturbations can alter outputs in medical imaging pipelines and in which economic incentives to manipulate billing or reimbursement outcomes exist [80]. They also draw attention to model

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

collapse, in which models trained recursively on synthetic outputs degrade, a risk that grows as machine-generated text enters clinical records and training corpora [81].

The safety-science perspective. This perspective supplies the concepts of specification gaming, side effects, scalable oversight and loss of control [23–25]. Its strength is anticipatory analysis. Its weakness, in clinical settings, is that its examples are often abstract. The present paper's contribution is to supply concrete clinical instantiations.

The regulatory perspective. Regulators attempt to convert abstract risk into enforceable duties. The European Union's AI Act classifies many medical AI systems as high-risk and attaches obligations concerning risk management, data governance, documentation, transparency and human oversight [82]. The World Health Organization's guidance articulates principles including protection of autonomy, safety, transparency, accountability, inclusiveness and sustainability, and has been extended to large multimodal models [83,84]. A comparative reading shows that regulatory instruments are more developed on pre-market conformity than on post-market learning and drift, which is where several of the documented failures arose.

The ethical and equity perspective. Ethicists point to risks of opacity, privacy erosion and unequal benefit [53,85,86]. This perspective is reinforced by sociological analyses of algorithmic stratification and by regional analyses of health and educational disparities in India that stress structural inequity and legislative gaps [20,57]. Its practical implication is that equity should be treated as a safety property and not as a separate moral add-on.

The health-systems and Global South perspective. Evidence generated in high-resource settings may not transport to settings with different disease profiles, device quality, connectivity and staffing. Analyses of disease transitions, telemedicine equity and infrastructure vulnerability in South Asia suggest that autonomous tools might be most attractive where specialists are scarce, which is also where validation data, regulatory capacity and recourse are weakest [19,58,63]. Scholarship on digital sovereignty and surveillance cautions that technology transfer can entrench dependency and monitoring [69,72]. This creates an ethical tension: the contexts with the greatest apparent need are those with the least capacity to detect harm.

5.7 Synthesis of the perspectives

Six perspectives, one conclusion. Each perspective, applied independently, arrives at a similar diagnosis: performance on benchmarks is an unreliable proxy for safe clinical autonomy, and the missing evidence concerns prospective impact, external validity, equitable performance and accountable oversight. What differs is the remedy. Clinicians ask for escalation rules, computer scientists for robustness testing, safety scientists for alignment and oversight, regulators for conformity and monitoring, ethicists for fairness and consent, and health-systems scholars for capacity and context. A credible pathway to autonomy must satisfy all six simultaneously, which is why single-discipline solutions repeatedly fail.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

6. Methodological Challenges in Evaluating Autonomous Clinical AI

6.1 Representativeness and the problem of the label

Every supervised system learns a mapping from recorded inputs to recorded outputs, and the recorded output is rarely the clinical truth. Diagnoses are entered for billing, reports are written in a defensive style, and outcomes are censored by who attends follow-up. The audit of the cost-based care-management algorithm showed how a proxy can embed inequity [13], and the survey literature on fairness catalogues the mechanisms through which sampling, measurement and aggregation bias enter models [52]. Representativeness problems are compounded in countries and regions where digital records are incomplete and where marginalised groups are less likely to be captured in datasets, a concern that regional analyses of health disparities and disease burden make plain [57–59]. A model trained on the data of well-resourced urban hospitals is evaluated implicitly against a population that does not include the patients for whom automation is most often proposed.

The methodological implication is that label audit should precede model development. Developers should be required to state what the target variable measures, who is missing, what clinical decisions shaped it and how it could differ across subgroups. This requirement is cheap compared with the cost of discovering label contamination after deployment.

6.2 Internal validation, external validation and temporal drift

The commonest design in the literature is internal validation on a random split of a single-site dataset. Such designs estimate performance on patients who resemble those used for training, collected in the same way, at the same time. They do not estimate performance at a new hospital, with a new scanner, after a change in coding practice or during a new epidemic. The pandemic literature offered a natural experiment in failure of generalisation [11,12], and the sepsis alert case showed that a model with respectable vendor-reported performance did not replicate externally [14]. Reviews of clinical decision support have long warned that local context and workflow shape outcomes [28].

Temporal drift deserves separate emphasis. Clinical practice changes, populations age, laboratory assays are replaced and documentation templates are redesigned. A static model therefore decays, sometimes silently. For autonomous systems, drift is not a maintenance problem; it is a safety problem, because the system continues to act while its calibration erodes. Prospective surveillance of performance in each site, with pre-specified thresholds that trigger de-escalation of autonomy, is the logical response and has few analogues in the current regulatory pathways for static products [55,82].

6.3 Prospective and impact-oriented evaluation

Diagnostic accuracy is an intermediate outcome. Autonomous deployment should be justified by patient-important outcomes such as mortality, morbidity, time to treatment, harms of overtreatment and equity of access. The distance between accuracy and outcome can be large. In the randomised trial of physician diagnostic reasoning, the model's own accuracy exceeded

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

that of the physicians but did not translate into better human performance [49]. Reporting standards for early live evaluation and for trials of AI interventions were created precisely to address the absence of such designs [39,41]. Nevertheless, the review evidence indicates that prospective evaluation remains exceptional [9].

There is also an ethical dimension. Prospective evaluation of an autonomous system exposes patients to its errors, so early-stage evaluation should be staged, with shadow-mode operation preceding supervised operation and supervised operation preceding independent operation. The ladder proposed in Section 8 formalises this progression.

6.4 Causal inference and sequential decision-making

Many proposals for autonomous care concern not one-off classification but sequences of interventions, such as dosing or fluid management. These are problems of causal inference under confounding, not prediction. The sepsis treatment policy learned from historical data illustrates the promise and the peril. It was estimated off-policy, on retrospective records, and its apparent benefit rests on assumptions about unmeasured confounding and overlap between the learned policy and clinician behaviour [51]. Guidelines for reinforcement learning in healthcare urge restraint, emphasising careful problem formulation, rigorous off-policy evaluation, uncertainty quantification and the avoidance of deployment on the strength of retrospective results alone [87]. In the language of safety science, such systems are subject to reward hacking and unsafe exploration [24], and in clinical language they risk learning to treat the measurement and not the patient.

6.5 Evaluating language models and general-purpose systems

Evaluation of foundation models raises distinctive problems. Benchmarks derived from licensing-style questions are convenient but unrepresentative of the open-ended, context-laden reasoning that clinical practice demands. A model that achieves a passing score on such questions [6] has demonstrated recall and pattern use in a constrained format, not safe management of an ambiguous patient. The systematic review of health-care evaluations of language models reported that only a small proportion used real patient-care data [48]. Potential contamination of test items in training data, prompt sensitivity, version drift and the tendency of models to produce fluent but fabricated content compound the difficulty [7,22]. Reviews of the medical applications of such models identify clinical documentation, summarisation and information retrieval as plausible near-term uses, while cautioning about reliability and accountability in diagnosis and treatment decisions [8].

A further and underappreciated problem is the integrity of the data ecosystem. As machine-generated text enters clinical notes and public corpora, later models risk training on the outputs of earlier ones, and experimental work has shown that recursive training can cause irreversible degradation of model distributions [81]. A health system that automates documentation without provenance tracking may therefore degrade the very data on which future evaluation depends.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

6.6 The human-machine team as the unit of evaluation

Evaluating a model in isolation, then deploying it within a team, is a category mistake. Performance depends on how humans interpret, trust and override outputs. The review literature on automation bias documents over-reliance and omission errors, with effect sizes that vary by task difficulty, workload and the presentation of advice [78]. The attentional account of complacency suggests that over-reliance increases when automation is reliable most of the time, which paradoxically makes high-accuracy systems more dangerous when they fail [79]. Studies of cognitive overload under information saturation, although developed outside medicine, converge on the claim that sustained load degrades judgement, which is relevant to clinicians operating under alert burden [74]. Analyses of burnout-related disengagement among employees under technologically intensified work add a workforce dimension [75]. Evaluation therefore needs team-level endpoints, including override appropriateness, time-to-detection of model failure and calibration of trust.

6.7 Reporting, reproducibility and incentives

The review by Nagendran and colleagues found that code and data were rarely accessible and that adherence to reporting standards was suboptimal [9]. Reproducibility is a safety prerequisite because an unreproducible claim cannot be independently checked before a patient is exposed to risk. Commercial confidentiality complicates the picture, since some of the most widely deployed systems are proprietary, as the sepsis case shows [14]. Incentives matter as well. Publication and funding systems reward positive and novel results, and a dynamic of competitive pressure, identified in the safety literature as a driver of risk, applies in health technology as it does elsewhere [25]. Ethical guidance for research, together with work on plagiarism, detection and AI-assisted integrity checking, reminds us that the credibility of the evidence base depends on the integrity of the people who produce and report it [42–45]. A related question concerns authorship and creative agency in generative systems, which has been examined in the context of art and is relevant to attribution of clinical documentation produced with machine assistance [88].

7. Safety Limits of Autonomous Healthcare

The term “safety limit” is used here to denote a boundary beyond which current science cannot give adequate assurance that autonomy is safe. Seven limits are identified.

7.1 Limit 1: Objective misspecification

A clinical objective is multidimensional and context-dependent. It includes survival, quality of life, patient preference, proportionality of treatment and fairness. A system optimising a single measurable proxy will predictably neglect the rest. The cost-based algorithm is a canonical case [13], and the theoretical literature explains why greater capability amplifies the risk: a more competent optimiser exploits the gap between proxy and intent more efficiently [23,24]. Medicine has partial remedies, including multi-objective evaluation, patient-reported outcomes and shared decision-making. The limit remains binding for any system that acts without such

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

inputs, and it is the reason why the prescribing and triage decisions that involve value trade-offs are poor candidates for early autonomy [85].

7.2 Limit 2: Distribution shift and silent failure

Models fail outside their training distribution, and often without signalling that they have done so. Silent failure is the most dangerous form because it defeats the oversight that depends on noticing errors. The pandemic models and the sepsis alert illustrate this, and the clinical-safety taxonomy flags it explicitly [11,14,77]. Techniques such as uncertainty estimation and out-of-distribution detection reduce but do not eliminate the problem. Until they do, autonomy should be restricted to tasks where inputs can be checked against a validated envelope and where out-of-envelope cases are escalated by design, as was done in the autonomous retinal screening system [27].

7.3 Limit 3: Adversarial and security vulnerability

Medical machine-learning systems are open to adversarial manipulation, whether by perturbing images, altering metadata or exploiting reimbursement incentives [80]. Surveillance and data-governance concerns widen the surface further. Privacy erosion in the age of medical big data, including the re-identification of ostensibly anonymised records, is a recognised risk [86]. At the level of states, analyses of digital surveillance, sovereignty and cyber-conflict suggest that health data and clinical infrastructure are plausible targets and levers of influence [69,70]. An autonomous system that can act on patients is, by that very fact, a high-value target.

7.4 Limit 4: Fabrication and unverifiable reasoning

Language-based systems can produce fluent, confident, incorrect statements, a phenomenon documented across natural-language-generation research [22]. Clinical consequences range from fabricated citations to incorrect dosing advice. Explanations generated by the model are not guaranteed to reflect the computation that produced the output, so they cannot be treated as audit trails. Existing assessments of language models in medicine emphasise assisted, human-verified use [7,8]. At present there is no empirical basis for allowing a generative system to make unreviewed clinical decisions.

7.5 Limit 5: The fragility of human oversight

Human-in-the-loop is frequently presented as the universal safeguard. The evidence suggests otherwise. Automation bias, complacency, alert fatigue and workload undermine the capacity of a human to detect machine error [28,78,79]. A rubber-stamp review satisfies a regulatory checkbox without delivering protection. Oversight is effective only when the reviewer has time, competence, authority, information and a realistic expectation of being right to disagree. Organisational analyses of leadership and safety culture indicate that such conditions are products of management decisions about staffing, training and psychological safety [30,32,60,62]. The EU AI Act's requirement of effective human oversight for high-risk systems recognises the principle, but its operationalisation in understaffed settings will be the real test [82].

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

7.6 Limit 6: Inequity as a safety hazard

If performance differs across subgroups, the system is unsafe for the groups for whom it performs worse, regardless of its average accuracy. Equity should therefore be a release criterion, with subgroup performance reported and thresholds set for the worst-performing group and not for the mean [52,53]. The sociological literature on algorithmic stratification and on digital health equity cautions that access gaps, language, device quality and trust shape both exposure to and benefit from automated care [19,20,76]. In settings in which health legislation is still maturing, the absence of statutory duties of equity audit compounds the problem [57,65].

7.7 Limit 7: Accountability diffusion and the liability vacuum

Autonomy creates a responsibility problem. When an autonomous system causes harm, candidates for liability include the developer, the deploying institution, the supervising clinician, the data provider and the regulator. Each can plausibly attribute the harm to another. The law of negligence supplies a conceptual starting point, requiring a duty of care, a breach of the standard of care, causation and damage [66]. Applying this framework to autonomous systems raises unresolved questions. What is the standard of care when a machine's performance exceeds that of the average clinician but fails unpredictably? Is a clinician negligent for following the machine, or for departing from it? Socio-legal theory reminds us that law operates as an instrument of social control whose doctrines adapt slowly and unevenly to technological change [67], and jurisprudential work on interpretation illustrates how courts reason by analogy when statutes are silent [68]. Health legislation scholarship argues that explicit statutory design, not judicial improvisation, is needed to align incentives within health systems [65].

Regulation adds a further layer. The European framework allocates obligations across providers and deployers of high-risk systems [82], and global guidance stresses accountability and the availability of redress [83]. Yet no jurisdiction has resolved the question of compensation for harm caused by a system whose failure mode was statistically foreseeable but individually unpredictable. A workable principle, drawn from the foregoing, is that autonomy should not be permitted without a pre-identified accountable party, mandatory incident reporting, an audit trail and a mechanism for patient redress.

7.8 Cross-cutting concerns: infrastructure, geopolitics and frontier risk

Two further concerns cut across the limits. The first is infrastructural dependence. Analyses of flood impacts on healthcare systems show that physical hazards can disable connectivity, power and facilities simultaneously [63,64]. An autonomous system that depends on cloud inference becomes a single point of failure during exactly those events in which clinical need is greatest, so graceful degradation to human-only operation must be engineered and rehearsed.

The second is the geopolitical and frontier dimension. Competition over compute and semiconductors, analysed in the context of strategic decoupling, shapes who builds and governs powerful models [73]. Safety scholars warn that competitive pressure can erode safety margins and that governance of increasingly capable systems has lagged behind capability [25,26]. For

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

clinical translation, the implication is modest but important: decisions about autonomy in healthcare should not be taken on the assumption that safety assurance developed for narrow systems will extend automatically to more general ones.

7.9 Summary matrix

Table 2 maps each translational gap and safety limit to its principal failure mechanism, a documented or analogous example and the corresponding safeguard.

Table 2. Translational gaps, safety limits and corresponding safeguards

Gap or limit	Failure mechanism	Documented or analogous evidence	Minimum safeguard before autonomy
Data gap	Unrepresentative training population	COVID-19 models at high risk of bias [11]; regional disparities in access and burden [57,58]	Dataset documentation; subgroup coverage report; external test sites
Label gap / objective misspecification	Proxy target diverges from clinical need	Cost-based allocation algorithm [13]	Label audit; multi-objective evaluation; patient-centred outcomes
Validation gap	Retrospective, single-site testing	Systematic review of deep-learning claims [9]; sepsis model [14]	Independent external and prospective validation, shadow mode
Integration gap	Correct output, unusable workflow	Field retinal screening evaluation [50]	Human-factors testing; workflow redesign; usability thresholds
Silent failure under shift	Out-of-distribution inputs undetected	Safety taxonomy [77]	Input envelope checks; uncertainty-triggered escalation
Adversarial and privacy risk	Manipulation; re-identification	Adversarial attacks on medical ML [80]; privacy analysis [86]	Security testing; data minimisation; incident response
Fabrication	Fluent but false outputs	Hallucination survey [22]; clinical LLM evaluations [48]	Verified-use only; provenance logging; no unreviewed decisions
Human oversight failure	Automation bias; alert fatigue	Systematic review of automation bias [78]; randomised trial [49]	Team-level evaluation; override monitoring; training and staffing
Inequity	Differential subgroup performance	Fairness frameworks [53]; digital health equity [19]	Worst-group performance thresholds; equity audit
Accountability gap	Diffuse liability	Negligence doctrine [66]; health legislation [65]	Named accountable party; audit trail; redress mechanism
Infrastructure fragility	Connectivity and facility loss	Flood-related healthcare vulnerability [63]	Offline fallback; downtime drills

Note: Evidence entries illustrate rather than exhaust the support for each row.

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

8. A Proposed Pathway: The Clinical Autonomy Ladder and Evidence-Gated Promotion

8.1 Rationale

Existing discussions tend to frame autonomy as binary: either a human decides or a machine does. Reality is graded. Drawing on the five-gap structure and the seven limits above, this paper proposes a Clinical Autonomy Ladder in which a system is classified by the degree of independence it exercises for a specific task, in a specific population and at a specific site. The ladder is an original conceptual proposal by the present author and is offered for critical testing, not as an established standard. It is intended to complement, not replace, regulatory classification and reporting standards [41,82].

8.2 The five levels

Level 1: Informational. The system displays data or retrieves information, and the clinician makes all interpretations. Examples include summarisation of a record under verification. The principal risks are fabrication and omission [22].

Level 2: Advisory. The system generates a recommendation, such as a risk score or a differential diagnosis, and the clinician must actively decide. Most deployed decision-support tools and current language-model applications belong here. The principal risks are automation bias and alert fatigue [28,78].

Level 3: Supervised autonomy. The system initiates an action unless a clinician vetoes it within a defined window, with systematic sampled audit. The risk is that the veto becomes illusory under workload [79].

Level 4: Bounded autonomy. The system acts independently within a narrowly specified, validated envelope, with automated escalation of out-of-envelope cases and retrospective audit. The autonomous retinal screening system approximates this level [27].

Level 5: Open-ended autonomy. The system manages multi-step, high-consequence decisions across varied contexts, with minimal human involvement. On the evidence reviewed, no current system justifies this level, and the safety limits in Section 7 apply with greatest force.

8.3 Promotion gates

Movement up the ladder should be contingent on passing gates that are cumulative and reversible. Seven gates are proposed.

Gate 1, specification. The clinical objective, the target variable, the input envelope and the failure modes are documented, and labels are audited for proxy contamination [13,52].

Gate 2, external validation. Performance is confirmed at independent sites and in temporal holdouts, and risk of bias is assessed with a structured instrument [38,40].

SHODH PRAKASHAN
An International Open Access, Peer Reviewed, Refereed Journal
ISSN : 2583-9861
VOL. 4, ISSUE 5, NOVEMBER 2026
Available Online : https://shodhprakashan.in/

Gate 3, prospective evaluation. Shadow-mode and then supervised prospective evaluation with pre-specified endpoints, reported to the relevant trial or early-evaluation standard [39,41].

Gate 4, equity audit. Subgroup performance meets a minimum threshold for every group, and the audit is published [53].

Gate 5, human-factors validation. Team-level evaluation shows that clinicians can detect and override errors in realistic workload conditions [49,78].

Gate 6, accountability readiness. A named accountable party exists; incident reporting, audit trails and patient redress are operational; and the legal basis for deployment is clear [65,66,82].

Gate 7, post-deployment surveillance. Continuous performance monitoring with automatic demotion to a lower level when drift thresholds are breached, and rehearsed fallback to human-only operation during infrastructure failure [63,77].

8.4 Mapping current systems

A realistic mapping is sobering. Most clinical language-model applications, most risk scores and most image-based decision-support tools sit at Levels 1 and 2 and should stay there until Gates 3 to 5 are met. A small number of bounded screening applications may justify Level 4 for narrow tasks. Reinforcement-learning treatment policies have not, on retrospective evidence alone, passed Gate 3 [51,87]. Population-level allocation algorithms must pass Gate 4 before use at any level above informational [13].

8.5 Implications for stakeholders

For developers, the ladder means documenting intended autonomy level and designing for demotion as well as promotion. For hospital leaders, it means treating oversight as a resourced function, with staffing, training and governance that reflect the autonomy level adopted, consistent with scholarship on leadership, patient safety and organisational learning [30,32,60]. For regulators, it means linking authorisation to level and requiring post-market evidence. For health-system planners in lower-resource settings, it means resisting premature leaps to high autonomy as a substitute for workforce investment, and insisting on local validation before import [19,58]. For researchers, it means shifting effort from benchmark gains to prospective, equity-audited, team-level trials.

9. Limitations

Several limitations qualify this work. First, the review was not prospectively registered, and it did not use a systematic multi-database search; Stream B was selected purposively and therefore carries curation bias. Second, items from the supplied corpus were assessed at the bibliographic level, because full texts were unavailable, so they support framing and context but not effect estimates. Third, the supplied corpus is concentrated around a small author network, which restricts diversity of viewpoint. Fourth, quantitative figures are drawn from heterogeneous designs and are presented descriptively without pooling, so cross-study

comparison is indicative only. Fifth, the field changes quickly, and systems, regulations and evidence published after the preparation of this manuscript are not reflected. Sixth, the Clinical Autonomy Ladder is a conceptual proposal that has not yet been empirically validated or tested for inter-rater reliability. Finally, the paper focuses on safety and methodological limits and does not provide a cost-effectiveness analysis, which would be needed before large-scale adoption.

Future research should test the ladder in Delphi consensus exercises, develop standardised metrics for oversight quality, evaluate subgroup-robust performance in under-represented regions, and design prospective trials that randomise at the level of workflow and not of patient alone.

10. Conclusion

The proposition that superintelligence is arriving in the clinic is, on the evidence, a statement about rhetoric and not about reality. What the evidence supports is that narrow systems can perform specific, well-bounded perceptual tasks at a level comparable with specialists, that general-purpose models offer promising but unreliable assistance, and that between these demonstrations and accountable autonomous care lie methodological gaps that have not yet been closed. The studies that make the boldest claims are, on balance, retrospective, single-site, poorly reported and at high risk of bias. The deployments that have been audited have revealed failures of sensitivity, fairness and usability that laboratory metrics did not forecast. The human oversight on which many arguments rely is itself fragile.

None of this is a counsel of despair. It is a counsel of sequence. Autonomy can be earned for tasks that are bounded, verifiable, safe in failure and supported by prospective evidence, and the autonomous retinal screening example shows what that looks like. For everything else, autonomy should be withheld until it is demonstrated through external validation, prospective impact evaluation, equity audit, human-factors testing and operational accountability. The Clinical Autonomy Ladder proposed here offers a vocabulary and a set of gates for doing so.

The deeper lesson concerns the nature of translation. Translating a capability into a clinical practice means translating it into a set of institutions, duties, incentives and relationships, which are slower to change than software. A responsible movement toward autonomous healthcare will therefore be measured less by the intelligence of the machine than by the maturity of the system that surrounds it.

References

1. Bostrom N. Superintelligence: paths, dangers, strategies. Oxford: Oxford University Press; 2014.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
3. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402-10.

4. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-8.
5. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.
6. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
7. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-9.
8. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40.
9. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689.
10. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271-97.
11. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020;369:m1328.
12. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3(3):199-217.
13. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53.
14. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065-70.
15. Ahluwalia PS. Artificial intelligence in healthcare: the role of artificial intelligence in diagnostic medicine. Siddhanta's Int J Adv Res Arts Humanit. 2025:164-75.
16. Ahluwalia PS. Redesigning patient engagement in AI-driven healthcare systems: an analysis. Siddhanta's Int J Adv Res Arts Humanit. 2025:203-17.
17. Singh P. Integrating machine learning for smart facility management in healthcare hospitality environments. Shodh Patra Int J Multidiscip Stud. 2025:82-92.
18. Ahluwalia PS. Assessing the impact of digital transformation on hospital management: a systematic review. Int J Healthc Hosp Manag Stud. 2025:1-16.
19. Ahluwalia PS. Digital health equity: evaluating access and outcomes in telemedicine across socioeconomic groups. Siddhanta's Int J Adv Res Arts Humanit. 2025:84-92.
20. Ahluwalia PS. Algorithmic bias and social stratification: how AI shapes inequality in digital societies. Siddhanta's Int J Adv Res Arts Humanit. 2025:28-37.
21. Ahluwalia PS, Kumar A. Platform poverty: the social consequences of algorithmic invisibility on gig workers. Siddhanta's Int J Adv Res Arts Humanit. 2025:240-9.
22. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):248.
23. Russell S. Human compatible: artificial intelligence and the problem of control. New York: Viking; 2019.
24. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. arXiv:1606.06565 [Preprint]. 2016.
25. Hendrycks D, Mazeika M, Woodside T. An overview of catastrophic AI risks. arXiv:2306.12001 [Preprint]. 2023.

26. Bengio Y, Hinton G, Yao A, Song D, Abbeel P, Darrell T, et al. Managing extreme AI risks amid rapid progress. Science. 2024;384(6698):842-5.
27. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018;1:39.
28. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17.
29. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA. 2018;320(21):2199-200.
30. Ahluwalia PS. Healthcare leadership and management: effective leadership styles in hospital administration. Siddhanta's Int J Adv Res Arts Humanit. 2025:69-82.
31. Ahluwalia PS. Crisis leadership in healthcare: organizational learning from pandemic failures. Siddhanta's Int J Adv Res Arts Humanit. 2025:136-46.
32. Ahluwalia PS, John S, Husain I, Elliason E. Organization learning & development: a comprehensive analysis of strategic human capital development in Indian corporate ecosystem. Interdiscip J Afr Alliance Res Advocacy Innov. 2025;1(3).
33. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195.
34. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
35. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467-73.
36. Ahluwalia PS. Enhancing systematic transparency through PRISMA-ScR: a scoping review approach for emerging researchers. Siddhanta's Int J Adv Res Arts Humanit. 2025:1-27.
37. Ahluwalia PS. Structured exploration through PRISMA-ScR: a transparent method for mapping research landscapes. Siddhanta's Int J Adv Res Arts Humanit. 2025:137-67.
38. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-8.
39. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health. 2020;2(10):e537-48.
40. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
41. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-33.
42. Ahluwalia PS. Ethical issues in social science research. Shodh Patra Int J Multidiscip Stud. 2024:1-8.
43. Ahluwalia PS. How we make plagiarism-free content: theory and practices. Shodh Prakashan J Humanit Soc Sci. 2025:46-53.
44. Ahluwalia PS. Plagiarism: detection techniques and tools. Siddhanta's Int J Adv Res Arts Humanit. 2024:1-11.
45. Ahluwalia PS. The role of artificial intelligence in combating plagiarism in scholarly publishing. Shodh Prakashan J Humanit Soc Sci. 2025:26-30.
46. Elliason EK, Ahluwalia PS. Prevalence and predictors of trauma-related mental health outcomes among refugees and internally displaced persons in Africa during conflict and crisis: a systematic review and meta-analysis. Interdiscip J Afr Alliance Res Advocacy Innov. 2025;1(2).

47. Mumtaz M, Ahluwalia PS. Yoga and cardiovascular health: meta-analysis of clinical studies. Sanatanodaya. 2025:302-12.
48. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. 2025;333(4):319-28.
49. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969.
50. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York: ACM; 2020. p. 1-12.
51. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24(11):1716-20.
52. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):115.
53. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866-72.
54. Ahluwalia PS. Smart cities and social divide: spatial justice in the age of urban tech. Siddhanta's Int J Adv Res Arts Humanit. 2025:185-202.
55. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. Lancet Digit Health. 2021;3(3):e195-203.
56. Ahluwalia PS. Artificial intelligence-driven prediction of antimicrobial resistance patterns: integrating machine learning with genomic epidemiology for precision antibiotic stewardship. J Integr Res. 2025:97-122.
57. Singh D, Ahluwalia PS. A socio-legal study on health disparities in Punjab: structural inequities, legislative gaps, and pathways to equitable healthcare. Shodh Prakashan. 2026:1-14.
58. Ahluwalia PS, John S. Disease pattern/trends in India: a comprehensive analysis of epidemiological transitions and public health challenges (2020-2025). Siddhanta's Int J Adv Res Arts Humanit. 2025:20-51.
59. Ahluwalia PS. Disease burden in India: a comprehensive analysis. Siddhanta's Int J Adv Res Arts Humanit. 2025:179-235.
60. Ahluwalia PS. Improving patient safety: strategies for reducing hospital-acquired infections. Siddhanta's Int J Adv Res Arts Humanit. 2024:126-45.
61. Ahluwalia PS. Transforming healthcare outcomes through evidence-based environmental design and educational interventions. Siddhanta's Int J Adv Res Arts Humanit. 2025:279-318.
62. Ahluwalia PS, John S, Elliason EK. Embedding trauma-informed care into healthcare leadership: a paradigm shift in organizational culture and patient outcomes. Interdiscip J Afr Alliance Res Advocacy Innov. 2025;1(2).
63. Ahluwalia PS. Impacts of floods on health and healthcare systems in Asian countries: a comprehensive analysis of direct and indirect health consequences, economic burden, and healthcare infrastructure vulnerability. Shodh Prakashan. 2025:72-102.
64. Ahluwalia PS. Unveiling the hidden crisis: comprehensive analysis of flood disasters and their socioeconomic impacts in Asian countries - a multi-dimensional assessment of risk, vulnerability, and adaptation strategies. Siddhanta's Int J Adv Res Arts Humanit. 2025:127-225.
65. Ahluwalia PS, John S. The role of health legislation in strengthening national health systems: a policy-legal perspective. Interdiscip J Anal Educ Sci Commer Manag Law. 2025:1-20.
66. Ahluwalia PS. Negligence under law of torts. Shodh Prakashan. 2025:1-18.
67. Ahluwalia PS, Singh D. The sociology of law: understanding the relationship between law, society, and social control. Siddhanta's Int J Adv Res Arts Humanit. 2026:142-8.

68. Ahluwalia PS. Judicial philosophy: foundations, schools of thought, and the problem of constitutional interpretation. Int J Humanit Leg Res. 2026:1-13.
69. Ahluwalia PS. The rise of techno-authoritarianism: digital surveillance and sovereignty in the Global South. Siddhanta's Int J Adv Res Arts Humanit. 2025:34-54.
70. Ahluwalia PS. Cyber-diplomacy and conflict: how states weaponize the internet in 21st century geopolitics. Siddhanta's Int J Adv Res Arts Humanit. 2025:157-71.
71. Ahluwalia PS. Invisible labor in the smart home: a sociological look at gender, automation, and domesticity. Siddhanta's Int J Adv Res Arts Humanit. 2025:143-56.
72. Ahluwalia PS, Elliason EK. Decolonizing sociology: indigenous epistemologies and the rewriting of sociological theory. Siddhanta's Int J Adv Res Arts Humanit. 2025:125-36.
73. Ahluwalia PS. Decoupling and the new cold war: US-China strategic competition in the global supply chain. Siddhanta's Int J Adv Res Arts Humanit. 2025:55-70.
74. Ahluwalia PS. Doomscrolling and mental fatigue: cognitive overload in the era of crisis media. Siddhanta's Int J Adv Res Arts Humanit. 2025:250-70.
75. Ahluwalia PS. The rise of quiet quitting: work ethic, alienation, and resistance in modern employment. Siddhanta's Int J Adv Res Arts Humanit. 2025:92-106.
76. Elliason EK, Singh K, Singh AP. Breaking the silence: understanding cancer-related stigma among tertiary students in Punjab. Siddhanta's Int J Adv Res Arts Humanit. 2025:1-24.
77. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf. 2019;28(3):231-7.
78. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19(1):121-7.
79. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. Hum Factors. 2010;52(3):381-410.
80. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science. 2019;363(6433):1287-9.
81. Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. Nature. 2024;631(8022):755-9.
82. 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 series.
83. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021.
84. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization; 2024.
85. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med. 2018;378(11):981-3.
86. Price WN 2nd, Cohen IG. Privacy in the age of medical big data. Nat Med. 2019;25(1):37-43.
87. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. Nat Med. 2019;25(1):16-8.
88. Ahluwalia PS, Khandey PD. AI as muse: redefining artistic authorship in the age of generative art. Siddhanta's Int J Adv Res Arts Humanit. 2025:15-33.