A global log for medical AI

Ayush Noori[1,2,3], Adam Rodman[4], Alan Karthikesalingam[5], Bilal A. Mateen[6,7],
Christopher A. Longhurst[8,9], Daniel Yang[10], Dave deBronkart[11], Gauden Galea[12,13],
Harold F. Wolf III[14], Jacob Waxman[15], Joshua C. Mandel[1,16], Juliana Rotich[17],
Kenneth D. Mandl[1,18,19], Maryam Mustafa[20,21], Melissa Miles[19], Nigam H. Shah[22,23,24],
Peter Lee[16], Robert Korom[25], Scott Mahoney[17], Seth Hain[26], Tien Yin Wong[27,28,29],
Trevor Mundel[19], Vivek Natarajan[5], Noa Dagan[3,15,30], David A. Clifton[2,31], Ran D. Balicer[3,15,32],
Isaac S. Kohane[1,3,18,†], Marinka Zitnik[1,3,33,34,35,†]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Department of Engineering Science, University of Oxford, Oxford, UK
[3]The Ivan and Francesca Berkowitz Family Living Laboratory Collaboration at
Harvard Medical School and Clalit Research Institute, Boston, MA, USA
[4]Division of General Internal Medicine, Department of Medicine,
Beth Israel Deaconess Medical Center, Boston, MA, USA
[5]Google DeepMind, London, UK
[6]University of Birmingham, Birmingham, UK
[7]PATH, Seattle, WA, USA
[8]Department of Medicine, University of California, San Diego, La Jolla, CA, USA
[9]Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA
[10]Kaiser Foundation Health and Hospitals, Oakland, CA, USA
[11]e-Patient Dave, LLC, Nashua, NH, USA
[12]Regional Office for Europe, World Health Organization, Copenhagen, Denmark
[13]University of Malta, Msida, Malta
[14]Healthcare Information and Management Systems Society, Chicago, IL, USA
[15]Clalit Research Institute, Innovation Division, Clalit Health Services, Ramat Gan, Israel
[16]Microsoft Research, Redmond, WA, USA
[17]Gates Foundation, Seattle, WA, USA
[18]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA
[19]Department of Pediatrics, Harvard Medical School, Boston, MA, USA
[20]Department of Computer Science, Lahore University of Management Sciences, Lahore, Pakistan
[21]Awaaz-e-Sehat, Lahore, Pakistan
[22]Technology and Digital Solutions, Stanford Healthcare, Palo Alto, CA, USA
[23]Department of Medicine, Stanford University, Stanford, CA, USA
[24]Clinical Excellence Research Center, Stanford University, Stanford, CA, USA
[25]Penda Health, Nairobi, Kenya
[26]Epic Systems Corporation, Verona, WI, USA
[27]Tsinghua Medicine, Tsinghua University, Beijing, China
[28]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
[29]Beijing Visual Science and Translational Eye Research Institute (BERI),
School of Clinical Medicine, Beijing Tsinghua Changgung Hospital, Beijing, China
[30]Software and Information Systems Engineering, Ben Gurion University of the Negev, Be'er Sheva, Israel
[31]Oxford Suzhou Centre for Advanced Research, University of Oxford, Suzhou, China
[32]Faculty of Health Sciences, School of Public Health, Ben Gurion University of the Negev, Be'er Sheva, Israel
[33]Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University, Allston, MA, USA
[34]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[35]Harvard Data Science Initiative, Cambridge, MA, USA

[†]Correspondence: isaac_kohane@hms.harvard.edu, marinka@hms.harvard.edu

Abstract

Modern computer systems often rely on syslog, a simple, universal protocol that records every critical event across heterogeneous infrastructure. However, healthcare's rapidly growing clinical AI stack has no equivalent. As hospitals rush to pilot large language models and other AI-based clinical decision support tools, we still lack a standard way to record how, when, by whom, and for whom these AI models are used. Without that transparency and visibility, it is challenging to measure real-world performance and outcomes, detect adverse events, or correct bias or dataset drift. In the spirit of syslog, we introduce MedLog, a protocol for event-level logging of clinical AI. Any time an AI model is invoked to interact with a human, interface with another algorithm, or act independently, a MedLog record is created. This record consists of nine core fields: header, model, user, target, inputs, artifacts, outputs, outcomes, and feedback, providing a structured and consistent record of model activity. To encourage early adoption, especially in low-resource settings, and minimize the data footprint, MedLog supports risk-based sampling, lifecycle-aware retention policies, and write-behind caching; detailed traces for complex, agentic, or multi-stage workflows can also be captured under MedLog. MedLog can catalyze the development of new databases and software to store and analyze MedLog records. Realizing this vision would enable continuous surveillance, auditing, and iterative improvement of medical AI, laying the foundation for a new form of digital epidemiology.

Introduction

Artificial intelligence (AI), including predictive and generative foundation models, is being implemented in clinical settings globally at an unprecedented rate and in a fragmented and largely unregulated manner. As of January 2025, at least 377 healthcare systems and providers in the U.S. alone have piloted or adopted 70 generative AI tools developed by 49 different companies for clinical decision support, patient communication, documentation, claims processing, and healthcare administration [1–3]; the majority of American physicians now report using AI technologies in clinical care [4]. These trends translate globally; for example, 48% of clinicians surveyed across 109 countries report using AI for work [5], and more than 300 hospitals in China have attempted to integrate local DeepSeek deployments into hospital systems [6]. This rapid adoption is driven by the strong technical performance of new AI models across several largely synthetic medical benchmarks. Large language models (LLMs), for example, have matched or outperformed physicians in diagnostic accuracy [7–9], clinical text summarization [10], medical question-answering tasks including licensing examinations [11–14], patient dialogue evaluated for quality and empathy [15, 16], and multi-step medical reasoning [17]. However, concerns about coherence, accuracy, hallucinations [18, 19], and bias [20, 21] persist, and the real-world clinical performance of these models at the healthcare system-level remains poorly understood [22]. Although more realistic benchmarks have begun to emerge [23, 24], systematic, production-grade evaluations that measure

clinical impact in deployed settings are rare [25–27].

Systematic evaluation in real-world healthcare settings critically depends on standardized data collection. Without consistent records, there can be no reliable analysis. Although some AI applications log limited usage information, and guidelines exist for clinical trial reporting of AI models – including TRIPOD+AI [28, 29], STARD-AI [30], DECIDE-AI [31], SPIRIT-AI, and CONSORT-AI [32] – these frameworks largely reflect the "classical" machine learning paradigm of feature engineering and task-specific training. They are, with few emerging exceptions [33], not designed to accommodate modern generative AI or AI agents that rely on pre-training, fine-tuning, prompting, and tool use. Moreover, once AI models pass clinical testing and enter deployment, no consensus scheme exists for event-level monitoring and auditing of real-world AI usage. Several frameworks – including OPTICA [34], FURM [35], FUTURE-AI [36], POLARIS-GM [37], Epic's open source `seismometer` package [38], and the Coalition for Health AI's Assurance Standards Guide [39] – offer guidance on model evaluation or governance before and during deployment, and individual health systems have developed custom workflows [2, 40]. Some have advocated for federated registries of clinical AI systems akin to ClinicalTrials.gov [41] or a national network of health AI assurance laboratories [42]. However, there is still no broadly adopted standard for logging each instance of AI use in clinical care or administration. The existing lack of standardized logging is a fundamental barrier to understanding and improving the safety and effectiveness of medical AI systems.

Beyond real-world evaluation of clinical AI performance, standardization of logging will likely be an essential part of any future regulatory regime for generative AI decision support tools. The FDA has previously argued that clinicians cannot possibly be expected to provide oversight for all outputs of generative AI, even with current implementations such as AI scribes [43]. This suggests using either "off-the-shelf" commercially available oversight tools or the development of new AI or LLM-specific monitoring devices. Either situation would require standards for data capture and analysis. In computer science, centralized logging protocols such as `syslog` have long enabled unified monitoring, troubleshooting, and auditing across complex, distributed systems [44, 45]. For example, `syslog` allows diverse network devices and applications to send standardized messages to a centralized logging server, recording the source system, message severity, and application-specific structured or free-text data (Table 1). By aggregating consistently formatted `syslog` messages in a single location, system administrators and security analysts can monitor live dashboards, rapidly diagnose root causes of errors, and maintain an auditable trail for compliance [46]. In fact, many enterprises, including healthcare organizations, operate

3

security information and event management pipelines that ingest terabytes of daily organization-wide system- and application-level log data to perform AI-assisted cybersecurity analytics [47]. It is clear that AI in healthcare needs an equivalent solution.

To address this gap, we introduce MedLog, a protocol for event-level logging of clinical AI. MedLog specifies a logging schema; we refer to single log entries of clinical AI interactions that conform to this schema as "MedLog records," and software that emits, transports, stores, or analyzes MedLog records as "MedLog systems." 'MedLog accommodates both single-shot prompts and agentic, multi-stage workflows by assembling a record from immutable messages over time. By standardizing capture and linkage across systems, MedLog enables continuous surveillance, comparative evaluation, and iterative improvement of medical AI. By recording key attributes of every AI interaction, MedLog systems will generate a critical data resource for understanding both immediate performance and longer-term effects of AI on healthcare delivery, and for comparative analysis of AI models across different healthcare institutions, systems, and countries. As AI systems increasingly mediate clinical care, a unified log of AI interactions in healthcare will become essential for monitoring healthcare outcomes, safeguarding patient interests, ensuring accountability, and guiding future development.

Existing clinical AI monitoring focuses on human-AI interactions. The scope of MedLog is broader and covers all AI processes that touch health data and can influence patient outcomes. This includes interactions between models and patients, clinicians, administrators, and other stakeholders; background services such as batch inference, autonomous triage, claim routing, and continuous monitoring; and AI-AI exchanges within agentic workflows and orchestration frameworks. As hospitals deploy LLM-based decision support, there is still no standard to record how, when, by whom, and for whom models are used, or to link use to what happened next. Without that visibility, health systems cannot measure real-world performance, detect adverse events or bias, or manage dataset shift. Shared expectations and lightweight conformance profiles will speed adoption by guiding health systems toward uniform capture that enables safety monitoring and comparative evaluation across sites.

We envision MedLog as a catalyst for a new form of data science and epidemiology – one centered on human behaviors and biology, and how they are influenced by AI – as well as the basis of safety systems for scalable oversight of AI outputs. Just as aviation relies on black box data to investigate incidents and drive safety improvements, healthcare must establish similar infrastructure for medical AI. The bottom line is clear: to realize the promise of AI in medicine, we must systematically monitor how AI interacts with patients, clinicians, and other AI models at

every instance of use. MEDLOG provides the standard and structure needed to make this possible.

MEDLOG systems capture key metadata for medical AI

MEDLOG complements two existing pillars of clinical AI development: model cards and data sheets. Model cards provide structured summaries of key facts about AI models, including architecture, training objectives, performance metrics, potential biases, and ethical considerations [48–50]. Data sheets describe attributes of the training datasets, such as data collection processes, pre-processing methods, demographic distributions, and known limitations or biases [51]. However, a critical layer is still missing: a systematic record of how AI models are actually used in clinical and operational contexts. MEDLOG systems would address this gap by monitoring model behavior and outputs in practice and, where feasible, linking those events to downstream clinical and operational outcomes.

We define the MEDLOG protocol to collect key aspects of clinical AI usage at each inference call to any deployed model (Figure 1a). While many current deployments follow a simple "prompt to model to response" pattern, clinical AI is rapidly moving toward agentic and multi-stage workflows that perform iterative retrieval, tool use, inline rubric evaluations, and multi-agent orchestration. MEDLOG is designed to accommodate these richer workflows by treating each model invocation as an event while allowing optional linkage across events that belong to the same run or episode (Figure 1b). This enables consistent logging for both simple and compositional systems. Each MEDLOG record, corresponding to a single model invocation, should include the following elements.

1. **Header.** The MEDLOG record header consists of provenance information, execution context, and system metadata available at inference time, including server identifiers, timestamps of model invocation and input retrieval, and a stable event identifier for the newly-created MED-LOG record. Akin to `PROCID` in `syslog`, this field can also optionally include identifiers for run- or episode-level linkage (*e.g.*, `run_id`, `parent_event_id`) [45]. These identifiers allow grouping events across multi-stage or agentic workflows without imposing a specific orchestration architecture. Lastly, to ensure interoperability, the header must include the version of the MEDLOG protocol specification that the record complies with.

2. **Model instance.** Stable identifiers of the AI model and version, with references to its model card and data sheet. In the absence of a data sheet, the entry should record the version of the training data and any databases queried for retrieval-augmented generation or other knowledge injection methods. Any test-time edits to the model should also be recorded.

3. **User identity.** The technical process, service, or workflow that invokes the model call. At a minimum, the identifier of the immediate calling process should be logged. When possible, the upstream users who initiated the call should also be recorded as a provenance chain. Human users, such as clinicians or patients, should be identified by their electronic health record (EHR) identifiers, such as National Provider Identifier (NPI) or medical record number (MRN). Users can also be algorithms or automated systems, such as AI agents or scheduled jobs that automatically trigger models for tasks such as risk calculation or triage [52–54]. The level of user detail may vary across clinical settings, and it may be easier to attribute model use to a single clinician in outpatient care compared to inpatient teams.

4. **Target identity.** When applicable, a reference to the entity about which the model produces output. For example, a patient ID number for clinical predictions or a claim ID for administrative tasks. Some models may not produce outputs about discrete targets, making this field optional.

5. **Inputs.** The input data provided to the model. For structured predictive models, this includes feature vectors or structured fields. For generative models such as LLMs, it should include prompts, instructions, and any relevant environmental variables. When inputs are too large to log directly, such as imaging or genomic data, stable identifiers sufficient to retrieve the input data retrospectively should be recorded.

6. **Internal artifacts.** Computational artifacts generated during inference, intended for technical audiences such as researchers and MLOps teams. This field can include reasoning traces, such as chain-of-thought [55], tree-of-thought [56], or graph-of-thought prompting paths [57]; external context retrieved during retrieval-augmented generation [58]; agent interaction traces, such as iterative hypothesis testing or round-table discussions among AI agents [54, 59]; uncertainty estimates, such as confidence scores, prediction intervals, generation quality metrics, or entropy-based measures; interpretability artifacts, including attribution maps, feature-importance scores, or saliency maps [60, 61]. To accommodate adaptive models that update continuously or episodically – including self-evolving models or agents [62] with Bayesian updating, persistent memory [63], lifelong learning [64], or dynamic routing or reconfiguration [65] – relevant model states or memory snapshots can also be logged in this field. These records allow retrospective reconstruction of the model's configuration at the moment of use with more granularity than the version identifier recorded in the Model instance field.

7. **Patient- or clinician-facing outputs.** The outputs intended for human users, including: predictive outputs such as labels, risk scores, or forecasts, with associated confidence measures; generative outputs such as text, images, or videos; explanations or rationales distilled from internal artifacts; and recommendations generated by single or multi-agent systems. Any explainability or reasoning components presented to users should also be recorded in this field, as well as any triage levels or risk scores that determine if a MEDLOG record is flagged for human review.

8. **Outcomes.** When feasible, records of downstream clinical actions or patient outcomes linked to the model's recommendation. For instance, whether a suggested therapy was administered and the observed clinical result. Capturing outcomes faces three constraints: the causal link between recommendation and action is often indirect; outcomes may only become observable after significant delays; and relevant data may reside outside the immediate AI workflow. Although outcome data may be incomplete, even partial linkage is valuable for post-deployment surveillance, epidemiological analysis, and iterative model improvement. Outcomes may also include traces of how patients or clinicians interact with the EHR system after viewing AI outputs, as recorded in EHR audit logs [66]. Retrospective outcome data can be linked to the appropriate MEDLOG record identifier using provider attestations, temporal proximity, trial emulations, or other automated queries. This field can also record the strength and basis of each linkage to enable tiered evidence standards for outcome attribution.

9. **User feedback.** Any feedback provided by users, whether structured ratings or free-text comments, should be recorded to support model refinement and user experience improvements.

MEDLOG balances two goals: providing lightweight, high-level indicators of AI use that enable system-wide querying, and supporting detailed, reproducible traces for complex workflows when needed. By default, MEDLOG emphasizes compact, standardized fields that are easy to integrate across systems. At the same time, the Internal artifacts field can include optional, detailed execution traces, with institutions choosing capture policies such as continuous logging, random sampling, or targeted collection during periods of elevated risk (*e.g.*, after major updates or during phased deployments).

Importantly, a complete MEDLOG record is assembled *incrementally* from a sequence of immutable, event-level messages that the AI system emits as it operates (Figure 1b). Each event is written to a collector that exposes a dedicated write-only endpoint for that event type. When
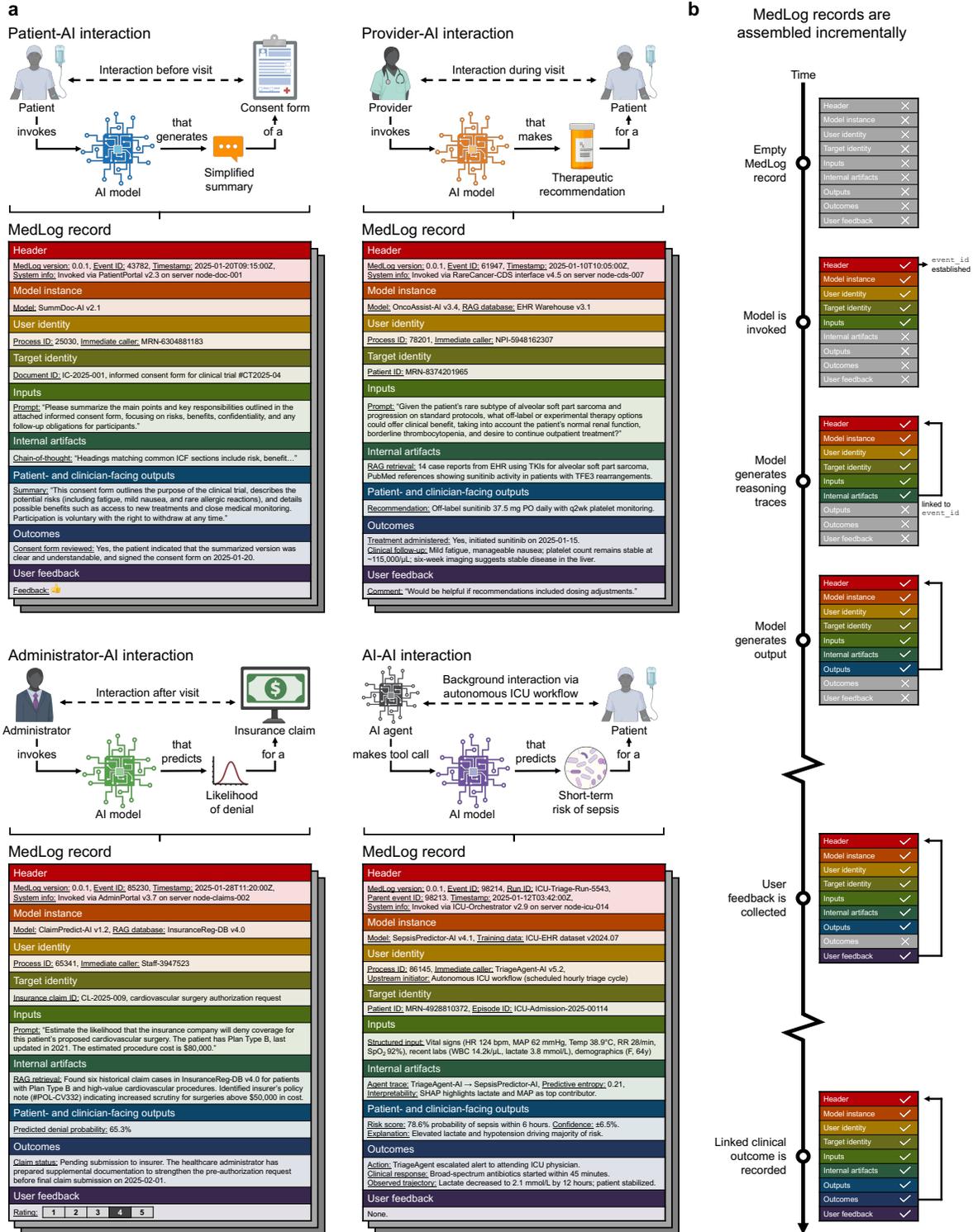
a

Patient–AI interaction

Interaction before visit

Patient invokes → AI model that generates Simplified summary → Consent form of a

MedLog record

Header
MedLog version: 0.0.1, Event ID: 43782, Timestamp: 2025-01-20T09:15:00Z, System info: Invoked via PatientPortal v2.3 on server node-doc-001

Model instance
Model: SummDoc-AI v2.1

User identity
Process ID: 25030, Immediate caller: MRN-6304881183

Target identity
Document ID: IC-2025-001, informed consent form for clinical trial #CT2025-04

Inputs
Prompt: "Please summarize the main points and key responsibilities outlined in the attached informed consent form, focusing on risks, benefits, confidentiality, and any follow-up obligations for participants."

Internal artifacts
Chain-of-thought: "Headings matching common ICF sections include risk, benefit…"

Patient- and clinician-facing outputs
Summary: "This consent form outlines the purpose of the clinical trial, describes the potential risks (including fatigue, mild nausea, and rare allergic reactions), and details possible benefits such as access to new treatments and close medical monitoring. Participation is voluntary with the right to withdraw at any time."

Outcomes
Consent form reviewed: Yes, the patient indicated that the summarized version was clear and understandable, and signed the consent form on 2025-01-20.

User feedback
Feedback: 👍

Provider–AI interaction

Interaction during visit

Provider invokes → AI model that makes Therapeutic recommendation → Patient for a

MedLog record

Header
MedLog version: 0.0.1, Event ID: 61947, Timestamp: 2025-01-10T10:05:00Z, System info: Invoked via RareCancer-CDS interface v4.5 on server node-cds-007

Model instance
Model: OncoAssist-AI v3.4, RAG database: EHR Warehouse v3.1

User identity
Process ID: 78201, Immediate caller: NPI-5948162307

Target identity
Patient ID: MRN-8374201965

Inputs
Prompt: "Given the patient's rare subtype of alveolar soft part sarcoma and progression on standard protocols, what off-label or experimental therapy options could offer clinical benefit, taking into account the patient's normal renal function, borderline thrombocytopenia, and desire to continue outpatient treatment?"

Internal artifacts
RAG retrieval: 14 case reports from EHR using TKIs for alveolar soft part sarcoma, PubMed references showing sunitinib activity in patients with TFE3 rearrangements.

Patient- and clinician-facing outputs
Recommendation: Off-label sunitinib 37.5 mg PO daily with q2wk platelet monitoring.

Outcomes
Treatment administered: Yes, initiated sunitinib on 2025-01-15.
Clinical follow-up: Mild fatigue, manageable nausea; platelet count remains stable at ~115,000/µL; six-week imaging suggests stable disease in the liver.

User feedback
Comment: "Would be helpful if recommendations included dosing adjustments."

Administrator–AI interaction

Interaction after visit

Administrator invokes → AI model that predicts Likelihood of denial → Insurance claim for a

MedLog record

Header
MedLog version: 0.0.1, Event ID: 85230, Timestamp: 2025-01-28T11:20:00Z, System info: Invoked via AdminPortal v3.7 on server node-claims-002

Model instance
Model: ClaimPredict-AI v1.2, RAG database: InsuranceReg-DB v4.0

User identity
Process ID: 65341, Immediate caller: Staff-3947523

Target identity
Insurance claim ID: CL-2025-009, cardiovascular surgery authorization request

Inputs
Prompt: "Estimate the likelihood that the insurance company will deny coverage for this patient's proposed cardiovascular surgery. The patient has Plan Type B, last updated in 2021. The estimated procedure cost is $80,000."

Internal artifacts
RAG retrieval: Found six historical claim cases in InsuranceReg-DB v4.0 for patients with Plan Type B and high-value cardiovascular procedures. Identified insurer's policy note (#POL-CV332) indicating increased scrutiny for surgeries above $50,000 in cost.

Patient- and clinician-facing outputs
Predicted denial probability: 65.3%

Outcomes
Claim status: Pending submission to insurer. The healthcare administrator has prepared supplemental documentation to strengthen the pre-authorization request before final claim submission on 2025-02-01.

User feedback
Rating: 1 2 3 **4** 5

AI–AI interaction

Background interaction via autonomous ICU workflow

AI agent makes tool call → AI model that predicts Short-term risk of sepsis → Patient for a

MedLog record

Header
MedLog version: 0.0.1, Event ID: 98214, Run ID: ICU-Triage-Run-5543, Parent event ID: 98213, Timestamp: 2025-01-12T03:42:00Z, System info: Invoked via ICU-Orchestrator v2.9 on server node-icu-014

Model instance
Model: SepsisPredictor-AI v4.1, Training data: ICU-EHR dataset v2024.07

User identity
Process ID: 86145, Immediate caller: TriageAgent-AI v5.2, Upstream initiator: Autonomous ICU workflow (scheduled hourly triage cycle)

Target identity
Patient ID: MRN-4928810372, Episode ID: ICU-Admission-2025-00114

Inputs
Structured input: Vital signs (HR 124 bpm, MAP 62 mmHg, Temp 38.9°C, RR 28/min, SpO₂ 92%), recent labs (WBC 14.2k/µL, lactate 3.8 mmol/L), demographics (F, 64y)

Internal artifacts
Agent trace: TriageAgent-AI → SepsisPredictor-AI, Predictive entropy: 0.21, Interpretability: SHAP highlights lactate and MAP as top contributor.

Patient- and clinician-facing outputs
Risk score: 78.6% probability of sepsis within 6 hours. Confidence: ±6.5%. Explanation: Elevated lactate and hypotension driving majority of risk.

Outcomes
Action: TriageAgent escalated alert to attending ICU physician.
Clinical response: Broad-spectrum antibiotics started within 45 minutes.
Observed trajectory: Lactate decreased to 2.1 mmol/L by 12 hours; patient stabilized.

User feedback
None.

b

MedLog records are assembled incrementally

Time

Empty MedLog record

Model is invoked — event_id established

Model generates reasoning traces — linked to event_id

Model generates output

User feedback is collected

Linked clinical outcome is recorded

(Record sections: Header, Model instance, User identity, Target identity, Inputs, Internal artifacts, Outputs, Outcomes, User feedback)

Figure 1: (a) Examples of clinical AI interactions that will be logged under the MEDLOG protocol, as well as the MEDLOG records they would create. **(b)** Timeline demonstrating that MEDLOG records are progressively built from a stream of messages.

inference begins, an initial message is emitted containing the immediately available fields: Header, Model instance, User identity, Target identity, and Inputs. This message establishes the primary `event_id` (and, optionally, a `run_id` for multi-step workflows). All subsequent messages containing Internal artifacts, Patient- or clinician-facing outputs, Outcomes, and User feedback reference this identifier to append new, schema-compliant fragments until the inference episode concludes. Because records are constructed incrementally, a MEDLOG record can be created even if the model generation ultimately fails, and Outcomes or Feedback – which require time to observe or collect – can be appended at any time after the inference has completed.

Building a MEDLOG system

Patient privacy and data security. Protecting patient privacy is paramount when implementing MEDLOG across healthcare systems. Like EHR databases, MEDLOG systems will record identifiable protected health information and sensitive operational data. Access to MEDLOG records must therefore be tightly controlled within secure computing environments, and the same regulatory standards and security systems used to protect current EHR databases – like Health Insurance Portability and Accountability Act (HIPAA), Health Information Technology for Economic and Clinical Health (HITECH) Act, General Data Protection Regulation (GDPR), or ISO/IEC 27001 [67, 68] compliance; role-based access control; audit logging; the use of pseudonymous identifiers; and storing content-addressed references rather than raw media – should be adopted for MEDLOG records. For example, AI outputs that influence clinical decision-making can reside in the EHR, while pointers to these outputs can be included in MEDLOG records, excluding MEDLOG from the HIPAA-designated record set or legal medical record. Similar to how LLMs and other clinical foundation models are now fine-tuned and adapted within institutional firewalls, healthcare systems can deploy MEDLOG systems locally to prevent unauthorized access. Where privacy regulations permit, secondary analysis tools or federated algorithms can operate within these environments to de-identify MEDLOG records and compute aggregated performance metrics or summary statistics, allowing meta-analyses while preserving patient confidentiality (Figure 2a). Data sharing agreements could then support inter-institutional comparisons of anonymized MEDLOG records [42], especially to evaluate models deployed outside their training settings [69]. For example, much as pharmacovigilance programs like the FDA Adverse Event Reporting System and the WHO Programme for International Drug Monitoring pool deidentified case reports to assess drug safety, deidentified or aggregated MEDLOG records could be shared with regulatory bodies tasked with post-market surveillance of clinical AI systems [70]. However, these cross-institutional analy-

ses will require privacy-preserving mechanisms for exchanging MEDLOG entries, such as secure multi-party computation, homomorphic encryption, federated learning, or blockchain [71, 72].

Data storage and management. Capturing each AI interaction is necessary to ensure real-world safety and accountability, but it will generate substantial data volumes. Implementing MED-LOG will require investment in large-scale data storage, management, and networking infrastructure, similar to the investments made in developing EHR systems in previous decades. However, data storage and networking demands are not unique to MEDLOG systems. Healthcare systems around the world are projected to generate more than 10,800 exabytes of data annually by 2025 [73, 74], and a single hospitalization already produces approximately 150,000 discrete data elements [73, 75]. The widespread adoption of clinical AI only increases the urgency of building a robust and secure healthcare data infrastructure capable of transmitting and storing exabytes or zettabytes of data. To maximize the impact of MEDLOG, access to such infrastructure must be democratized so that all healthcare systems can participate in AI monitoring and improvement efforts.

In practice, organizations may default to retaining all MEDLOG events indefinitely to maximize observability and support retrospective analysis. However, as in other high-compliance domains, tailored strategies can balance safety, privacy, and cost: institutions can adopt lifecycle-aware capture and retention policies that use full tracing during pilots and post-update periods, sampled or risk-triggered tracing in steady state, and tiered retention (*e.g.*, long-term summaries with shorter-lived detailed artifacts). This approach preserves forensic and regulatory value while containing operational overhead.

Pathways to real-world deployment. MEDLOG can be adopted unilaterally within a health system to improve safety monitoring and evaluation with no external mandate. However, shared expectations and harmonized interfaces across EHR vendors, AI vendors, and health systems will reduce integration costs and enable consistent, multi-site analyses. We anticipate a mixed adoption pathway: early adopters implement MEDLOG locally, while emerging guidance from professional bodies and regulators fosters convergence toward uniform capture and exchange. To support legacy and MEDLOG-naive systems, organizations can implement MEDLOG at high-leverage points in the technology stack, *e.g.*, as LLM proxies or API gateways that intercept AI calls, extract or augment metadata, and emit MEDLOG-compliant entries. Sidecar services can similarly wrap agent frameworks and tool calls to record inputs, retrieved context, outputs, and uncertainty estimates. These patterns accelerate adoption without waiting for deep vendor changes.

MEDLOG can be implemented using existing open standards and tooling. MEDLOG software should adopt the W3C PROV conceptual model for computational provenance; for example, the

MEDLOG record and its fragments are `prov:Entity` instances; the model invocation itself is a `prov:Activity`; and the model and user are instances of `prov:Agent` or its subclasses, such as `prov:SoftwareAgent` or `prov:Person` [76]. For operational telemetry, OpenTelemetry provides consistent schemas, collectors, and backends to transport and store event data across languages and platforms. For clinical semantics and linkage, Fast Healthcare Interoperability Resources (FHIR) data formats and elements (*e.g.*, AuditEvent, Patient, Condition, Observation, Practitioner, PractitionerRole) can anchor MEDLOG entries to standardized clinical entities [77]. These interoperability layers enable scalable, vendor-agnostic deployments and lower the barrier to multi-institutional analyses.

Global implementation of MEDLOG. To support deployments in low- and lower-middle-income countries (LMICs), MEDLOG allows partial or incremental compliance. A minimal conformance profile can capture Header, Model instance, and Outputs, with other fields added as capacity grows. Local write-behind caching enables offline operation with delayed synchronization – for example, between lightweight smartphone applications and a centralized MEDLOG server – when connectivity becomes available. Where EHR systems or unique identifiers are limited, MEDLOG records can anchor to encounter-level metadata such as visit, time, location, and department, with optional FHIR linkage when feasible. MEDLOG can also map to widely used platforms such as OpenMRS [78] and DHIS2 [79]. In settings with limited infrastructure, lifecycle-aware retention and risk-triggered sampling will be essential. In countries with emerging regulatory frameworks, MEDLOG records can be overseen by health system administrators, health ministries, or international partners such as the World Health Organization. Funders should back pilot implementations across health systems to prevent widening disparities and to establish systematic monitoring of LLM outputs in patient care.

Aligning incentives and governance. The promise of MEDLOG depends on policy as much as technology. We must learn from past efforts to build distributed digital public health infrastructure [80]. For example, although the HITECH Act of 2009 was successful in ubiquitizing EHRs, information exchange encountered roadblocks: in 2015, 96% of hospitals either claimed exclusion from or did not report to specialized public health registries [81]. To encourage the adoption of MEDLOG, the business case must be explicit: even without cross-institutional data sharing, MEDLOG delivers safety and quality improvement, liability management, and operational efficiency. Institutions can preserve competitive data advantages by retaining raw logs locally while participating in benchmarking consortia that return site-level insights as a reciprocal benefit. MEDLOG could even unlock new pathways to financial sustainability; for example, providers could leverage AI

performance data to establish value-based contracts for AI services. Furthermore, to encourage clinician participation, deployments must build trust by presenting MEDLOG as a collaborative tool for enhancing clinical learning, addressing potential concerns about professional autonomy. To that end, establishing governance bodies with strong clinician representation is essential.

Incentives and risks for model developers and vendors must also be considered. For example, large numbers of input-output examples, uncertainty estimates, or reasoning traces from MEDLOG records could enable membership inference attacks that expose a model's training set [82, 83] or extraction attacks to distill proprietary models into imitations [84] or reconstruct proprietary prompting techniques or tool calls [85, 86]. To mitigate such adversarial misuse, which could discourage vendor participation, governance frameworks for MEDLOG should require technical safeguards and intellectual property protection agreements that prevent reverse engineering [87, 88]. Certain MEDLOG fields, such as Model instance, Inputs, and Outputs, can also include data ownership tags to ensure clear provenance. Ultimately, success will hinge on mechanisms such as these to align incentives, governance structures, and funding models.

How MEDLOG can transform medical AI

Large-scale logging of human-AI interactions has shown clear value for safety monitoring, jailbreak detection, usage analysis, benchmark construction, and instruction fine-tuning to align model outputs with human preferences [89, 90]. The same applies in healthcare and medicine: monitoring medical AI models through MEDLOG systems can advance model development, evaluation, safety, reliability, and transparency. Below, we outline the opportunities opened by MEDLOG and the research, policy, and organizational changes needed to achieve them (Figure 2b).

From human epidemiology to human-AI epidemiology. Continuous monitoring of AI-human interactions establishes a data layer that has not previously existed in medicine: systematic records of how AI systems participate in healthcare. Traditional epidemiology studies the distribution and determinants of health and disease in populations. Digital epidemiology expanded this scope by drawing on electronic health records, claims, and patient-generated data. MEDLOG extends it further by treating AI itself as a measurable agent in clinical environments. With MEDLOG, for the first time, epidemiology can also study machine behavior alongside human behavior. By capturing inputs, reasoning traces, outputs, and linked outcomes, MEDLOG records document how AI influences decision-making, clinical actions, and patient trajectories. This transforms epidemiology from studying how humans and environments shape health to also studying how algorithms mediate those processes. Using implementation science approaches [91], epidemiologists can an-

Figure 2: (a) Example patterns of model invocation and corresponding record creation in a MEDLOG implementation. De-identified MEDLOG records can be aggregated across healthcare systems to support downstream applications. **(b)** MEDLOG will transform medicine by enabling evaluation, auditing, and improvement of medical AI.

alyze MEDLOG records to detect both positive [92] and negative [93] performance changes with clinician-AI collaboration; variation in recommendations across demographic groups, specialties, or regions; clinician over- or under-reliance on AI; workflow changes introduced by automation; and the long-term effects of AI-assisted care on quality and safety. These analyses will generate quantitative evidence for health policy, support population-tailored AI interventions, and guide precision medicine. MEDLOG systems will also provide hospital quality-improvement teams with operational intelligence across outpatient clinics, inpatient wards, operating rooms, and ancillary

services. Analysts can link MedLog records to clinical and financial outcomes, identify patterns of risk, and refine AI workflows to improve safety and efficiency.

Real-time surveillance of medical AI safety. MedLog records give regulators the means to detect adverse events from near misses, errors, or model failures in real time. Event-level logs support auditing and compliance by supplying the "artifact collection" required by frameworks for medical algorithmic audits [94, 95]. Algorithmic auditing can be automated: as in modern cybersecurity monitoring, AI systems can process, summarize, and triage MedLog records, triggering alerts for additional human review when necessary and making continuous organization-wide oversight operationally feasible [96–98]. Real-time surveillance will depend on regulatory mandates for post-market reporting, such as those proposed by the U.S. Food and Drug Administration (FDA) [99] and the European Commission, and on public-private partnerships such as a network of health AI assurance laboratories [42]. Recent FDA guidance stresses credibility assessment plans, life cycle management, and automated oversight as part of regulatory approval [100, 101]. Without MedLog, these regulatory requirements cannot be met in practice; with it, they become operationally feasible.

Detecting dataset shifts in medical AI. By logging AI-human interaction events, MedLog records allow developers to detect when model behavior deviates from expectations because of shifts in deployment data. Dataset shift arises when models encounter changes in patient demographics, clinical practices, medical technologies, or care settings compared to their training environments [102, 103]. Event-level logging preserves the full context of each model invocation, including inputs, outputs, and outcomes, making it possible to track how shifts manifest across subgroups, workflows, or care settings rather than only at an aggregate level. The structure of MedLog records, which captures retrievals, generations, outcomes, and feedback, also allows us to distinguish between shifts in the underlying data distribution and shifts in how models are applied in practice. Comparing data distributions between training and deployment stages, using methods such as deep learning-based hypothesis testing [104, 105], enables early detection of global and subgroup-level shifts that standard outlier detection may miss [106]. MedLog records can also reveal performance degradation when LLMs are retrained on new data, helping to identify risks such as data poisoning attacks [107].

Monitoring for bias in medical AI. MedLog records enable systematic assessment of bias by tracking model performance across attributes such as age, sex, race, ethnicity, socioeconomic status, and insurance coverage [108–111]. Automated slice discovery methods applied to MedLog records can detect differential performance [112]. Panels of clinicians and ethicists can then

14

determine whether observed differences reflect clinically justified variation or inequities that require remediation [110, 113]. Continuous bias monitoring in this way strengthens accountability and supports equitable use of AI across diverse populations. Standardized logging can also provide regulators with the evidence needed to evaluate whether clinical AI systems meet fairness and safety requirements.

Using MEDLOG data to improve AI models.   In MEDLOG systems, real-world error cases, near misses, and uncertainty signals provide a rich diagnostic layer for model refinement. For example, uncertainty estimates recorded for each prediction can support active learning strategies, where the least confident predictions are flagged for expert review and used for model post-training [114]. Beyond traditional supervised or reinforcement fine-tuning, these diagnostics can also power meta-learning strategies like curriculum learning [115–117], lifelong learning [64], or self-evolving agents [62]. Based on feedback signals in MEDLOG traces, pre-trained models or agents can adaptively sequence prompt [118] or data [119, 120] examples, critique or revise their own actions [121, 122], generate new tasks from failures [123, 124], update long-term memory modules [63], or even autonomously modify their tools or design [125, 126]. De-identified MEDLOG records can also be mined to construct challenging, realistic benchmarks that outperform today's artificial evaluations at estimating real-world performance on clinical tasks [127]. Machine learning teams embedded in healthcare systems can use MEDLOG records or MEDLOG-derived benchmarks to determine when to deploy, retire, or retrain AI models and compare the performance of different models (*e.g.*, GPT-5 versus Claude Sonnet 4 versus Gemini 2.5 Pro) in real-time or retrospectively. Beyond model-level decisions, MEDLOG records can capture user interaction patterns, informing improvements in user interfaces, workflow integration, system management, and the development of new clinical AI applications.

International evaluation of AI models.   As MEDLOG records capture the same fields across sites, public health agencies and consortia can aggregate de-identified logs to benchmark AI models worldwide across geographic and economic strata. Medical AI research shows marked geoeconomic disparities: as of August 2024, only 2.3% of studies were conducted in LMICs and only 6.3% spanned more than one nation [128, 129]. Yet, early real-world LMIC deployments [130] have already reduced diagnostic and treatment errors [131]. The opportunity to measure the global impact of clinical AI is therefore urgent and largely untapped. Routine, standardized logging can show whether AI tools narrow or widen performance gaps between resource-rich and resource-constrained hospitals, identify deployments that need additional context-specific training [69], and inform developers, regulators, and funders seeking to support systems that reduce rather

than entrench global health inequities.

Advancing transparency for patients and clinicians. MEDLOG records create traceable documentation of AI-generated content in EHRs and administrative claims. Clinicians can review the rationales behind AI outputs to guide patient care. In the United States, health data downloads are identified by the "Blue Button" logo [132, 133]; with wider MEDLOG adoption, exported health data could also include AI interaction traces. Patients using personal health LLM [134] may then rely on MEDLOG exports to audit the information available to their models or to transfer their digital medical assistants across platforms.

Case study: AI monitoring detects real-life data drift

To illustrate how continuous AI monitoring can safeguard model performance and patient outcomes, we present the following case study. Clalit Health Services deployed an AI tool designed to predict hospitalization risk and prioritize chronic patients more effectively for proactive periodic nursing follow-up. Specifically, a gradient-boosting machine with 48 features – including demographics, laboratory values, diagnoses, medications, and medical procedures – was trained to predict non-ambulatory hospitalization over one year of follow-up. Based on their predicted hospitalization risk, patients were prioritized for nursing follow-up every six months to two years. This tool was embedded into Clalit's proactive-preventative interventions platform, C-Pi, which combines an advanced decision-support system with an AI-based prioritization engine. C-Pi enables thousands of primary care physicians and community nurses to identify which patients require proactive outreach, while also providing detailed management recommendations to reduce care gaps across the population.

The hospitalization risk model was trained in 2020 using data from January 2018 to January 2019, and tested on a temporal hold-out set from January 2019 to January 2020. After demonstrating strong performance, it was deployed in practice. Importantly, Clalit implemented monitoring of the model's input features. In mid-2023, the monitoring system detected a distribution shift in the "Lactate Dehydrogenase Last Value (LDH)" feature. A subsequent investigation revealed that this distribution shift was caused by a centralized switch to a new LDH testing kit in March 2023. After March 2023, LDH testing was conducted with the new kit, gradually altering the distribution of LDH values. This real-world instance of data drift was detected by the AI monitoring system.

Figure 3 depicts the shift in LDH distributions across time: starting from the training period (January 2018), remaining stable until the kit change (March 2023), and then gradually diverging

Figure 3: Density plots show the distribution of the "Lactate Dehydrogenase Last Value (LDH)" feature during the training period (January 2018), immediately after the test kit change (March 2023), and in subsequent quarterly snapshots through September 2024. The introduction of the new test kit caused a gradual shift in the distribution of LDH values, which was automatically detected by the AI monitoring system.

over the next 18 months. We further simulated the impact of the drift on hospitalization risk predictions by running the model retrospectively at 3-monthly intervals from June 2023 through September 2024 both with and without correction for the shifted LDH values. Even small feature drifts could propagate into clinically meaningful prediction errors: by 18 months, nearly 10% of patients would have had the absolute risk scores shifted by $> 0.1\%$, and about 1% by $> 1\%$. If the change had occurred in a feature with increased feature importance, this data drift could have resulted in a dramatic effect on the final predictive score. This case study demonstrates how continuous monitoring of AI can detect subtle, system-level changes that would otherwise degrade model accuracy and quality.

A call to action for monitoring medical AI

Systematic monitoring of interactions between medical AI models and patients, providers, administrators, and other algorithms is urgently needed. These interactions occur across all encounters between individuals and the healthcare system, from routine visits to large-scale deployments and trials spanning multiple institutions. Without event-level records, it is impossible to measure real-world performance and outcomes, detect patient and enterprise safety risks, identify and correct biases, enable continuous model improvement, or compare models and systems across settings.

The need is analogous to pharmacovigilance: just as adverse drug events cannot be detected or mitigated without standardized reporting, the safety of medical AI cannot be assured without logging. Harmonized monitoring will also allow global benchmarking, making it possible to assess whether AI tools generalize across populations and health systems, or whether they reinforce existing inequities. For health systems, routine logging will provide practical benefits beyond safety, including liability protection, quality improvement, and operational intelligence. For regulators, MEDLOG offers the infrastructure required to implement forthcoming oversight regimes, which emphasize post-market surveillance, accountability, and lifecycle management. The MEDLOG protocol provides the framework to record and analyze AI use. We call on policymakers, regulators, healthcare leaders, and the AI community to adopt standards for monitoring AI in practice.

The `syslog` protocol became a global standard under the Internet Engineering Task Force's principle of "rough consensus and running code." MEDLOG should follow the same path. Interoperability will require community consensus on a consistent nomenclature for the nine fields of MEDLOG records. This process must engage stakeholders across large-scale compute and storage, logging and tracing infrastructure, AI development, and clinical operations. At the same time, we plan to build robust, production-ready MEDLOG systems and pilot them in real-world healthcare settings. We invite partners from research, industry, and health systems to join this effort.

Building the MEDLOG foundation is essential to realizing the benefits of medical AI while protecting patient outcomes and public trust. Without systematic monitoring, medical AI cannot be safe, fair, or effective. With it, medicine can be reshaped on a foundation of transparency and accountability. Just as `syslog` became indispensable for modern computing, MEDLOG must become indispensable for medicine.

Ethics approval.   Parts of this work that relate to the Clalit Health Services prediction model use case were approved by the Clalit Health Services Institutional Review Board (Helsinki) committee.

Code availability.   To illustrate how the design ideas of the MEDLOG protocol may translate into practice, we have developed a proof-of-concept demonstration:

- Source code: https://github.com/mims-harvard/medlog
- Documentation: https://zitniklab.hms.harvard.edu/medlog

This minimal prototype employs an OpenAPI-described HTTP REST interface, though the MEDLOG specification itself is transport-agnostic and may be implemented using telemetry-specific alternatives such as the OpenTelemetry Protocol (OTLP). Future work and community consensus

are critically needed to develop interoperability standards and test MEDLOG systems in real-world healthcare settings.

Competing interests. A.K. and V.N. are currently employed by Google DeepMind. D.D. is currently employed by e-Patient Dave, LLC. H.F.W. is currently employed by Healthcare Information and Management Systems Society, Inc. J.C.M. and P.L. are currently employed by Microsoft Research. J.R. is currently employed by E-Citizen Solutions Africa. S.H. is currently employed by Epic Systems Corporation. The other authors declare no competing interests.

Attribute	`syslog`	MEDLOG
Purpose	A protocol to send event messages from network devices and applications to a centralized logging server	A protocol for event-level logging of clinical AI
Users	IT and security teams	Clinicians, AI/ML engineers and researchers, safety regulators
Structure	1. `HEADER`, which includes `PRI` (facility and severity), `VERSION` (version of the `syslog` protocol), `TIMESTAMP`, `HOSTNAME` (hostname and the domain name of the originator), `APP-NAME` (device or application that originated the message), `PROCID` (used to detect discontinuities or group messages), and `MSGID` (identifies the type of message) 2. `STRUCTURED-DATA`, which can contain multiple structured data elements (`SD-ELEMENT`), each recorded as a name (`SD-ID`) and a parameter name-value pair (`SD-PARAM`) 3. `MSG`, a free-text message	1. Header 2. Model instance 3. User identity 4. Target identity 5. Inputs 6. Internal artifacts 7. Patient- and clinician-facing outputs 8. Outcome 9. User feedback
Granularity	One record per event	One record per model invocation
Privacy	Clear-text protocol with no default encryption; not designed for sensitive data	Contains protected health information; encryption required
Storage	Moderate ($\gtrsim$ KB-GB day^{-1} per hospital)	Large ($\gtrsim$ GB-TB day^{-1} per hospital)

Table 1: Design and feature comparison of `syslog` (RFC 5424 [45]; for the legacy BSD `syslog` protocol specification, see RFC 3164 [44]) and MEDLOG.

References

1. Palmer, K., Ross, C. & Parker, J. E. *A guide to the health systems and companies driving adoption* 2023.

2. Umeton, R. *et al.* GPT-4 in a Cancer Center — Institute-Wide Deployment Challenges and Lessons Learned. *NEJM AI* **1,** AIcs2300191. doi:10.1056/AIcs2300191 (2024).

3. Landi, H. Epic introduces Launchpad to spur generative AI adoption. *Fierce Healthcare* (2025).

4. American Medical Association. *Physician sentiments around the use of AI in heath care: motivations, opportunities, risks, and use cases* tech. rep. (American Medical Association, Chicago, Illinois, 2025).

5. Elsevier. *Clinician of the Future 2025* tech. rep. (Elsevier, Amsterdam, Netherlands, 2025).

6. Zeng, D., Qin, Y., Sheng, B. & Wong, T. Y. DeepSeek's "Low-Cost" Adoption Across China's Hospital Systems: Too Fast, Too Soon? *JAMA* **333,** 1866–1869. doi:10.1001/jama.2025.6571 (2025).

7. Kanjee, Z., Crowe, B. & Rodman, A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA* **330,** 78–80. doi:10.1001/jama.2023.8288 (2023).

8. Goh, E. *et al.* Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open* **7,** e2440969. doi:10.1001/jamanetworkopen.2024.40969 (2024).

9. Cabral, S. *et al.* Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. *JAMA Internal Medicine* **184,** 581–583. doi:10.1001/jamainternmed.2024.0295 (2024).

10. Van Veen, D. *et al.* Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine* **30,** 1134–1142. doi:10.1038/s41591-024-02855-5 (2024).

11. Xie, Y. *et al. A Preliminary Study of o1 in Medicine: Are We Closer to an AI Doctor?* 2024. doi:10.48550/arXiv.2409.15277.

12. Katz, U. *et al.* GPT versus Resident Physicians — A Benchmark Based on Official Board Scores. *NEJM AI* **1,** AIdbp2300192. doi:10.1056/AIdbp2300192 (2024).

13. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620,** 172–180. doi:10.1038/s41586-023-06291-2 (2023).

14. Tu, T. *et al.* Towards Generalist Biomedical AI. *NEJM AI* **1,** AIoa2300138. doi:10.1056/AIoa2300138 (2024).

15. Ayers, J. W. *et al.* Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine* **183,** 589–596. doi:10.1001/jamainternmed.2023.1838 (2023).

16. Tu, T. *et al. Towards Conversational Diagnostic AI* 2024. doi:`10.48550/arXiv.2401.05654`.

17. Brodeur, P. G. *et al. Superhuman performance of a large language model on the reasoning tasks of a physician* 2024. doi:`10.48550/arXiv.2412.10849`.

18. Ahmad, M. A., Yaramis, I. & Roy, T. D. *Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI* 2023. doi:`10.48550/arXiv.2311.01463`.

19. Kim, Y. *et al. Medical Hallucinations in Foundation Models and Their Impact on Healthcare* 2025. doi:`10.48550/arXiv.2503.05777`.

20. Mittermaier, M., Raza, M. M. & Kvedar, J. C. Bias in AI-based models for medical applications: challenges and mitigation strategies. *npj Digital Medicine* **6,** 113. doi:`10.1038/s41746-023-00858-z` (2023).

21. Cross, J. L., Choma, M. A. & Onofrey, J. A. Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health* **3,** e0000651. doi:`10.1371/journal.pdig.0000651` (2024).

22. Wornow, M. *et al.* The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine* **6,** 1–10. doi:`10.1038/s41746-023-00879-8` (2023).

23. Arora, R. K. *et al. HealthBench: Evaluating Large Language Models Towards Improved Human Health* 2025. doi:`10.48550/arXiv.2505.08775`.

24. Bedi, S. *et al. MedHELM: Holistic Evaluation of Large Language Models for Medical Tasks* 2025. doi:`10.48550/arXiv.2505.23802`.

25. Shah, N. H., Milstein, A. & Bagley Steven C., P. Making Machine Learning Models Clinically Useful. *JAMA* **322,** 1351–1352. doi:`10.1001/jama.2019.10306` (2019).

26. Youssef, A., Pencina, M., Thakur, A., Zhu, T., Clifton, D. & Shah, N. H. External validation of AI models in health should be replaced with recurring local validation. *Nature Medicine* **29,** 2686–2687. doi:`10.1038/s41591-023-02540-z` (2023).

27. Agweyu, A. *et al. Retrospective Evaluation of a Generative AI-Enabled Electronic Medical Record System in Primary Health Care Facilities in Kenya* 2025. doi:`10.1101/2025.09.05.25335163`.

28. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *The Lancet* **393,** 1577–1579. doi:`10.1016/S0140-6736(19)30037-6` (2019).

29. Collins, G. S. *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385,** e078378. doi:`10.1136/bmj-2023-078378` (2024).

30. Sounderajah, V. *et al.* Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nature Medicine* **26,** 807–808. doi:`10.1038/s41591-020-0941-1` (2020).

31. Vasey, B. *et al.* DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nature Medicine* **27,** 186–187. doi:`10.1038/s41591-021-01229-5` (2021).

32. Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K. & Calvert, M. J. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine* **26,** 1351–1363. doi:10.1038/s41591-020-1037-7 (2020).

33. The CHART Collaborative. Reporting Guideline for Chatbot Health Advice Studies: The CHART Statement. *JAMA Network Open* **8,** e2530220. doi:10.1001/jamanetworkopen.2025.30220 (2025).

34. Dagan, N. *et al.* Evaluation of AI Solutions in Health Care Organizations — The OPTICA Tool. *NEJM AI* **1,** AIcs2300269. doi:10.1056/AIcs2300269 (2024).

35. Callahan, A. *et al.* Standing on FURM Ground: A Framework for Evaluating Fair, Useful, and Reliable AI Models in Health Care Systems. *NEJM Catalyst* **5,** CAT.24.0131. doi:10.1056/CAT.24.0131 (2024).

36. Lekadir, K. *et al.* FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **388,** e081554. doi:10.1136/bmj-2024-081554 (2025).

37. Ong, J. C. L. *et al.* International partnership for governing generative artificial intelligence models in medicine. *Nature Medicine,* 1–4. doi:10.1038/s41591-025-03787-4 (2025).

38. Epic Open Source. *epic-open-source/seismometer* Verona, WI, 2025.

39. Coalition for Health AI. *Assurance Standards Guide* 2024.

40. Bedoya, A. D. *et al.* A framework for the oversight and local deployment of safe and high-quality prediction models. *Journal of the American Medical Informatics Association* **29,** 1631–1636. doi:10.1093/jamia/ocac078 (2022).

41. Pencina, M. J., McCall, J. & Economou-Zavlanos, N. J. A Federated Registration System for Artificial Intelligence in Health. *JAMA* **332,** 789–790. doi:10.1001/jama.2024.14026 (2024).

42. Shah, N. H. *et al.* A Nationwide Network of Health AI Assurance Laboratories. *JAMA* **331,** 245–249. doi:10.1001/jama.2023.26930 (2024).

43. Warraich, H. J., Tazbaz, T. & Califf, R. M. FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine. *JAMA* **333,** 241–247 (2025).

44. Lonvick, C. M. *The BSD Syslog Protocol* RFC 3164 (Internet Engineering Task Force, 2001). doi:10.17487/RFC3164.

45. Gerhards, R. *The Syslog Protocol* RFC 5424 (Internet Engineering Task Force, 2009). doi:10.17487/RFC5424.

46. Zhang, T., Qiu, H., Castellano, G., Rifai, M., Chen, C. S. & Pianese, F. System Log Parsing: A Survey. *IEEE Transactions on Knowledge and Data Engineering* **35,** 8596–8614. doi:10.1109/TKDE.2022.3222417 (2023).

47. Forrester Research, Inc. *The Total Economic Impact™ Of Microsoft Sentinel* Forrester Total Economic Impact™ Study (Forrester Research, Inc., Cambridge, Massachusetts, 2024).

48. Mitchell, M. *et al. Model Cards for Model Reporting* in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), 220–229. doi:`10.1145/3287560.3287596`.

49. OpenAI. *GPT-4o System Card* 2024.

50. Meta. *Llama 3.2 Model Card* 2024.

51. Gebru, T. *et al. Datasheets for Datasets* 2021. doi:`10.48550/arXiv.1803.09010`.

52. Williams, C. Y. K. *et al.* Use of a Large Language Model to Assess Clinical Acuity of Adults in the Emergency Department. *JAMA Network Open* **7,** e248895. doi:`10.1001/jamanetworkopen.2024.8895` (2024).

53. Hinson, J. S. *et al.* Multisite implementation of a workflow-integrated machine learning system to optimize COVID-19 hospital admission decisions. *npj Digital Medicine* **5,** 1–10. doi:`10.1038/s41746-022-00646-1` (2022).

54. Gao, S. *et al.* Empowering biomedical discovery with AI agents. *Cell* **187,** 6125–6151. doi:`10.1016/j.cell.2024.09.022` (2024).

55. Wei, J. *et al.* Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* **35,** 24824–24837 (2022).

56. Yao, S. *et al.* Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in Neural Information Processing Systems* **36,** 11809–11822 (2023).

57. Besta, M. *et al.* Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence* **38,** 17682–17690. doi:`10.1609/aaai.v38i16.29720` (2024).

58. Ng, K. K. Y., Matsuba, I. & Zhang, P. C. RAG in Health Care: A Novel Framework for Improving Communication and Decision-Making by Addressing LLM Limitations. *NEJM AI* **0,** AIra2400380. doi:`10.1056/AIra2400380` (2024).

59. Chen, J. C.-Y., Saha, S. & Bansal, M. *ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs* 2024. doi:`10.48550/arXiv.2309.13007`.

60. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* **3,** 610–619. doi:`10.1038/s42256-021-00338-7` (2021).

61. Saraswat, D. *et al.* Explainable AI for Healthcare 5.0: Opportunities and Challenges. *IEEE Access* **10,** 84486–84517. doi:`10.1109/ACCESS.2022.3197671` (2022).

62. Gao, H.-a. *et al. A Survey of Self-Evolving Agents: On Path to Artificial Super Intelligence* 2025. doi:`10.48550/arXiv.2507.21046`.

63. Li, Z. *et al. MemOS: An Operating System for Memory-Augmented Generation (MAG) in Large Language Models* 2025. doi:`10.48550/arXiv.2505.22101`.

64. Zheng, J. *et al. Lifelong Learning of Large Language Model based Agents: A Roadmap* 2025. doi:`10.48550/arXiv.2501.07278`.

65. Guo, Y., Cheng, Z., Tang, X., Tu, Z. & Lin, T. *Dynamic Mixture of Experts: An Auto-Tuning Approach for Efficient Transformer Models* 2025. doi:`10.48550/arXiv.2405.14297`.

66. Kannampallil, T. & Adler-Milstein, J. Using electronic health record audit log data for research: insights from early efforts. *Journal of the American Medical Informatics Association : JAMIA* **30,** 167–171. doi:`10.1093/jamia/ocac173` (2022).

67. ISO/TC 215. *ISO 27799:2016: Health informatics — Information security management in health using ISO/IEC 27002* 2016.

68. ISO/IEC JTC 1/SC 27. *ISO/IEC 27001:2022: Information security, cybersecurity and privacy protection — Information security management systems — Requirements* 2022.

69. Yang, J. *et al.* Generalizability assessment of AI models across hospitals in a low-middle and high income country. *Nature Communications* **15,** 8270. doi:`10.1038/s41467-024-52618-6` (2024).

70. Embi, P. J. Algorithmovigilance—Advancing Methods to Analyze and Monitor Artificial Intelligence–Driven Health Care for Effectiveness and Equity. *JAMA Network Open* **4,** e214622. doi:`10.1001/jamanetworkopen.2021.4622` (2021).

71. Rieke, N. *et al.* The future of digital health with federated learning. *npj Digital Medicine* **3,** 1–7. doi:`10.1038/s41746-020-00323-1` (2020).

72. Xia, Q., Sifah, E. B., Asamoah, K. O., Gao, J., Du, X. & Guizani, M. MeDShare: Trust-Less Medical Data Sharing Among Cloud Service Providers via Blockchain. *IEEE Access* **5,** 14757–14767. doi:`10.1109/ACCESS.2017.2730843` (2017).

73. Banks, M. A. Sizing up big data. *Nature Medicine* **26,** 5–6. doi:`10.1038/s41591-019-0703-0` (2020).

74. Telenti, A. & Jiang, X. Treating medical data as a durable asset. *Nature Genetics* **52,** 1005–1010. doi:`10.1038/s41588-020-0698-y` (2020).

75. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nature Medicine* **25,** 24–29. doi:`10.1038/s41591-018-0316-z` (2019).

76. Khalid Belhajjame *et al. PROV-O: The PROV Ontology* W3C Recommendation (World Wide Web Consortium (W3C), Cambridge, MA, 2013).

77. HL7. *FHIR* 2023.

78. Mamlin, B. W. *et al.* Cooking Up An Open Source EMR For Developing Countries: OpenMRS – A Recipe For Successful Collaboration. *AMIA Annual Symposium Proceedings* **2006,** 529–533 (2006).

79. Dehnavieh, R. *et al.* The District Health Information System (DHIS2): A literature review and meta-synthesis of its strengths and operational challenges based on the experiences of 11 countries. *Health Information Management Journal* **48,** 62–75. doi:`10.1177/1833358318777713` (2019).

80. Kadakia, K. T., Howell, M. D. & DeSalvo, K. B. Modernizing Public Health Data Systems: Lessons From the Health Information Technology for Economic and Clinical Health (HITECH) Act. *JAMA* **326,** 385–386. doi:`10.1001/jama.2021.12000` (2021).

81. Office of the National Coordinator for Health Information Technology. *Hospital Selection of Public Health Measures in Medicare EHR Incentive Program* tech. rep. 16 (Washington D.C., 2016).

82. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. *Membership Inference Attacks Against Machine Learning Models* in *2017 IEEE Symposium on Security and Privacy (SP)* (2017), 3–18. doi:`10.1109/SP.2017.41`.

83. Mattern, J., Mireshghallah, F., Jin, Z., Schoelkopf, B., Sachan, M. & Berg-Kirkpatrick, T. *Membership Inference Attacks against Language Models via Neighbourhood Comparison* in *Findings of the Association for Computational Linguistics: ACL 2023* (eds Rogers, A., Boyd-Graber, J. & Okazaki, N.) (Association for Computational Linguistics, Toronto, Canada, 2023), 11330–11343. doi:`10.18653/v1/2023.findings-acl.719`.

84. Zhao, K., Li, L., Ding, K., Gong, N. Z., Zhao, Y. & Dong, Y. *A Systematic Survey of Model Extraction Attacks and Defenses: State-of-the-Art and Perspectives* 2025. doi:`10.48550/arXiv.2508.15031`.

85. Sha, Z. & Zhang, Y. *Prompt Stealing Attacks Against Large Language Models* 2024. doi:`10.48550/arXiv.2402.12959`.

86. Zhang, C., Morris, J. X. & Shmatikov, V. *Extracting Prompts by Inverting LLM Outputs* in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (eds Al-Onaizan, Y., Bansal, M. & Chen, Y.-N.) (Association for Computational Linguistics, Miami, Florida, USA, 2024), 14753–14777. doi:`10.18653/v1/2024.emnlp-main.819`.

87. He, X. *et al.* CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks. *Advances in Neural Information Processing Systems* **35,** 5431–5445 (2022).

88. Li, Q. *et al.* LLM-PBE: Assessing Data Privacy in Large Language Models. *Proc. VLDB Endow.* **17,** 3201–3214. doi:`10.14778/3681954.3681994` (2024).

89. Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y. & Deng, Y. *WildChat: 1M ChatGPT Interaction Logs in the Wild* 2024. doi:`10.48550/arXiv.2405.01470`.

90. Zheng, L. *et al. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset* 2024. doi:`10.48550/arXiv.2309.11998`.

91. Longhurst, C. A., Singh, K., Chopra, A., Atreja, A. & Brownstein, J. S. A Call for Artificial Intelligence Implementation Science Centers to Evaluate Clinical Effectiveness. *NEJM AI* **1,** AIp2400223. doi:`10.1056/AIp2400223` (2024).

92. McDuff, D. *et al.* Towards accurate differential diagnosis with large language models. *Nature* **642,** 451–457. doi:`10.1038/s41586-025-08869-4` (2025).

93. Budzyń, K. *et al.* Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology* **10,** 896–903. doi:`10.1016/S2468-1253(25)00133-5` (2025).

94. Raji, I. D. *et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing* in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, New York, NY, USA, 2020), 33–44. doi:`10.1145/3351095.3372873`.

95. Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K. & Oakden-Rayner, L. The medical algorithmic audit. *The Lancet Digital Health* **4,** e384–e397. doi:`10.1016/S2589-7500(22)00003-6` (2022).

96. Zhu, J., He, S., He, P., Liu, J. & Lyu, M. R. *Loghub: A Large Collection of System Log Datasets for AI-driven Log Analytics* in *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)* (2023), 355–366. doi:`10.1109/ISSRE59848.2023.00071`.

97. Qi, J. *et al. LogGPT: Exploring ChatGPT for Log-Based Anomaly Detection* in *2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)* (2023), 273–280. doi:`10.1109/HPCC-DSS-SmartCity-DependSys60770.2023.00045`.

98. Karlsen, E., Luo, X., Zincir-Heywood, N. & Heywood, M. Benchmarking Large Language Models for Log Analysis, Security, and Interpretation. *Journal of Network and Systems Management* **32,** 59. doi:`10.1007/s10922-024-09831-x` (2024).

99. Warraich, H. J., Tazbaz, T. & Califf, R. M. FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine. *JAMA.* doi:`10.1001/jama.2024.21451` (2024).

100. Center for Veterinary Medicine *et al. Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products* Draft Guidance (Food and Drug Administration, Silver Spring, MD, 2025).

101. Center for Devices and Radiological Health, Center for Biologics Evaluation and Research & Center for Drug Evaluation and Research. *Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations* Draft Guidance (Food and Drug Administration, Silver Spring, MD, 2025).

102. Subbaswamy, A. & Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **21,** 345–352. doi:`10.1093/biostatistics/kxz041` (2020).

103. Finlayson, S. G. *et al.* The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine* **385,** 283–286. doi:`10.1056/NEJMc2104626` (2021).

104. Koch, L. M., Baumgartner, C. F. & Berens, P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. *npj Digital Medicine* **7,** 1–11. doi:`10.1038/s41746-024-01085-w` (2024).

105. Schrouff, J. *et al.* Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems* **35,** 19304–19318 (2022).

106. Koch, L. M., Schürch, C. M., Gretton, A. & Berens, P. *Hidden in Plain Sight: Subgroup Shifts Escape OOD Detection* in *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning* (PMLR, 2022), 726–740.

107. Alber, D. A. *et al.* Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine* **31,** 618–626. doi:`10.1038/s41591-024-03445-1` (2025).

108. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. *Nature Medicine* **27,** 2176–2182. doi:10.1038/s41591-021-01595-0 (2021).

109. Daneshjou, R. *et al.* Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances* **8,** eabq6147. doi:10.1126/sciadv.abq6147 (2022).

110. Liu, M. *et al.* A translational perspective towards clinical AI fairness. *npj Digital Medicine* **6,** 1–6. doi:10.1038/s41746-023-00918-4 (2023).

111. Xiong, Z. *et al.* How Generalizable Are Foundation Models When Applied to Different Demographic Groups and Settings? *NEJM AI* **2,** AIcs2400497. doi:10.1056/AIcs2400497 (2025).

112. Eyuboglu, S. *et al. Domino: Discovering Systematic Errors with Cross-Modal Embeddings* 2022. doi:10.48550/arXiv.2203.14960.

113. De Kanter, A.-F. J., van Daal, M., de Graeff, N. & Jongsma, K. R. Preventing Bias in Medical Devices: Identifying Morally Significant Differences. *The American Journal of Bioethics* **23,** 35–37. doi:10.1080/15265161.2023.2186516 (2023).

114. Budd, S., Robinson, E. C. & Kainz, B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis* **71,** 102062. doi:10.1016/j.media.2021.102062 (2021).

115. Bengio, Y., Louradour, J., Collobert, R. & Weston, J. *Curriculum learning* in *Proceedings of the 26th Annual International Conference on Machine Learning* (Association for Computing Machinery, New York, NY, USA, 2009), 41–48. doi:10.1145/1553374.1553380.

116. Liu, F., Ge, S., Zou, Y. & Wu, X. *Competence-based Multimodal Curriculum Learning for Medical Report Generation* 2023. doi:10.48550/arXiv.2206.14579.

117. Rui, S., Chen, K., Ma, W. & Wang, X. *Improving Medical Reasoning with Curriculum-Aware Reinforcement Learning* 2025. doi:10.48550/arXiv.2505.19213.

118. Liu, Y., Liu, J., Shi, X., Cheng, Q., Huang, Y. & Lu, W. *Let's Learn Step by Step: Enhancing In-Context Learning Ability with Curriculum Learning* 2024. doi:10.48550/arXiv.2402.10738.

119. Bae, S., Hong, J., Lee, M. Y., Kim, H., Nam, J. & Kwak, D. *Online Difficulty Filtering for Reasoning Oriented Reinforcement Learning* 2025. doi:10.48550/arXiv.2504.03380.

120. Chen, X. *et al. Self-Evolving Curriculum for LLM Reasoning* 2025. doi:10.48550/arXiv.2505.14970.

121. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. & Yao, S. Reflexion: language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* **36,** 8634–8652 (2023).

122. Madaan, A. *et al.* Self-Refine: Iterative Refinement with Self-Feedback. *Advances in Neural Information Processing Systems* **36,** 46534–46594 (2023).

123. Qi, Z. *et al. WebRL: Training LLM Web Agents via Self-Evolving Online Curriculum Reinforcement Learning* 2025. doi:`10.48550/arXiv.2411.02337`.

124. Zhou, Y., Levine, S., Weston, J., Li, X. & Sukhbaatar, S. *Self-Challenging Language Model Agents* 2025. doi:`10.48550/arXiv.2506.01716`.

125. Robeyns, M., Szummer, M. & Aitchison, L. *A Self-Improving Coding Agent* 2025. doi:`10.48550/arXiv.2504.15228`.

126. Hu, S., Lu, C. & Clune, J. *Automated Design of Agentic Systems* 2025. doi:`10.48550/arXiv.2408.08435`.

127. Lin, B. Y. *et al. WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild* 2024. doi:`10.48550/arXiv.2406.04770`.

128. Yang, R. *et al.* Disparities in clinical studies of AI enabled applications from a global perspective. *npj Digital Medicine* **7,** 209. doi:`10.1038/s41746-024-01212-7` (2024).

129. Han, R., Acosta, J. N., Shakeri, Z., Ioannidis, J. P. A., Topol, E. J. & Rajpurkar, P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *The Lancet Digital Health* **6,** e367–e373. doi:`10.1016/S2589-7500(24)00047-5` (2024).

130. Mateen, B. A. *et al.* Trials for LLM-supported clinical decisions in African primary healthcare. *Nature Medicine,* 1–3. doi:`10.1038/s41591-025-03815-3` (2025).

131. Korom, R. *et al. AI-Based Clinical Decision Support for Primary Care: A Real-World Study* 2025. doi:`10.48550/arXiv.2507.16947`.

132. Assistant Secretary for Technology Policy. *Blue Button* 2022.

133. Mandl, K. D. *et al.* Push Button Population Health: The SMART/HL7 FHIR Bulk Data Access Application Programming Interface. *npj Digital Medicine* **3,** 151. doi:`10.1038/s41746-020-00358-4` (2020).

134. Khasentino, J. *et al.* A personal health large language model for sleep and fitness coaching. *Nature Medicine,* 1–10 (2025).