From Guardrails to Compilers: A Constraint-First Runtime Substrate for Clinical AI Agents

Kaiping Zheng
National University of Singapore
Singapore
dcszkai@nus.edu.sg

Horng-Ruey Chua
National University Hospital
Singapore
mdcchr@nus.edu.sg

Si Wei Kheok
Singapore General Hospital
Singapore
kheok.si.wei@singhealth.com.sg

Kee Yuan Ngiam
National University Hospital
Singapore
kee_yuan_ngiam@nuhs.edu.sg

Marcus Chun Jin Tan
National University Hospital
Singapore
ophcjmt@nus.edu.sg

Robert John Walsh
National University Cancer Institute
Singapore
robert_walsh@nuhs.edu.sg

ABSTRACT

LLM-powered clinical agents now read electronic health records, retrieve evidence, and draft recommendations inside molecular tumor boards, ICUs, and rare-disease clinics. Recent evaluations are sobering: agents hallucinate dosages and omit standard-of-care steps, induce automation bias even among AI-trained physicians, fall to medical-advice prompt injection at above 90% success rates, and often cannot produce a defensible audit trail when a decision is later contested. We argue this is a *runtime* failure, not a model failure: the substrate on which medical agents execute lacks first-class data primitives for constraint enforcement, provenance tracking, and policy-aware federation. We therefore sketch a constraint-first runtime that compiles every agent action, whether a tool call, a retrieval, a semantic operator, or a final recommendation, into a constrained, provenance-tracked query plan. Semantic integrity constraints, computable subsets of oncology and sepsis guidelines, ontological axioms over standard clinical vocabularies, and federation-aware access policies are expressed in one specification, from which the runtime compiles a plan whose safety, security, and auditability hold by construction. Under this view, the data-management primitives long refined by the community (semantic constraints and operators, semiring-annotated provenance, secure federated query processing, access control, and incremental view maintenance) become the trusted computing base for clinical agents, and contextual intelligence becomes a matter of query planning rather than model scale. We illustrate the design with a molecular-tumor-board example, sketch a four-layer architecture, and lay out the open questions it raises for the data-management community.

VLDB Workshop Reference Format:
Kaiping Zheng, Horng-Ruey Chua, Si Wei Kheok, Kee Yuan Ngiam, Marcus Chun Jin Tan, and Robert John Walsh. From Guardrails to Compilers: A Constraint-First Runtime Substrate for Clinical AI Agents. VLDB 2026 Workshop: Biomedical Data Management Systems (BioDMS).

Proceedings of the VLDB Endowment. ISSN 2150-8097.

1 INTRODUCTION

A molecular tumor board (MTB) at a major cancer center may review many complex cases in a single session under tight time constraints [76]; preparing each patient's dossier requires painstakingly integrating multiple clinical systems, external genomic reports, pathology slides, prior imaging, and treatment histories, together with a rapidly evolving research literature that may not surface the single most relevant trial in time [38, 72]. Elsewhere in the same hospital, an ICU intensivist operates under an even tighter decision cycle: the New York SEP-1 mandate operationalized a three-hour sepsis bundle whose timely completion has been associated with lower mortality [23, 66] , while any decision support that misses a contraindication or triggers a stale alert risks both patient harm and regulatory scrutiny [3, 24]. Both settings sit on the same edge: too much data and too many rules to weigh by hand, too little time to reason from scratch, and the same later question that an auditor or attorney may pose: *which evidence supported this decision, and was the standard of care satisfied?*

LLM-powered clinical agents promise to compress these workflows [8, 17, 48, 85], yet recent evidence is sobering. A 2025 multimodel study reported 91.8% of surveyed clinicians encountering medical hallucinations and 84.7% believing these could harm patients [36]. Omission rates reach 97% on common diagnostic prompts [79], and clinically relevant omissions outweigh hallucinations head-to-head (3.45% vs. 1.47%) [2]. A randomized trial showed that physicians with formal AI-literacy training still suffered a 14-point drop in diagnostic accuracy under plausible but incorrect LLM suggestions [27, 57]. Adversarial evaluations show degraded performance across leading models [52], and medical LLMs are vulnerable to misinformation injected into training or retrieval [32]; one recent study elicited harmful medical advice through prompt injection with a 94.4% success rate, including 91.7% in severe scenarios involving FDA Category-X drugs [42]. Vision-language oncology models exhibit similar vulnerabilities [16], and clinical-LLM-agent benchmark suites have emerged [34, 48]. The dominant AI-community response is *guardrails*: post-hoc filters, retrieval-augmented citations, and self-checks that treat the agent as a black-box text generator. Figure 1 (left) shows the result: a leaky post-hoc filter in front of a black-box LLM, an overflowing alert-fatigue inbox, a stale guideline binder, and a masked prompt-injection adversary slipping through it.

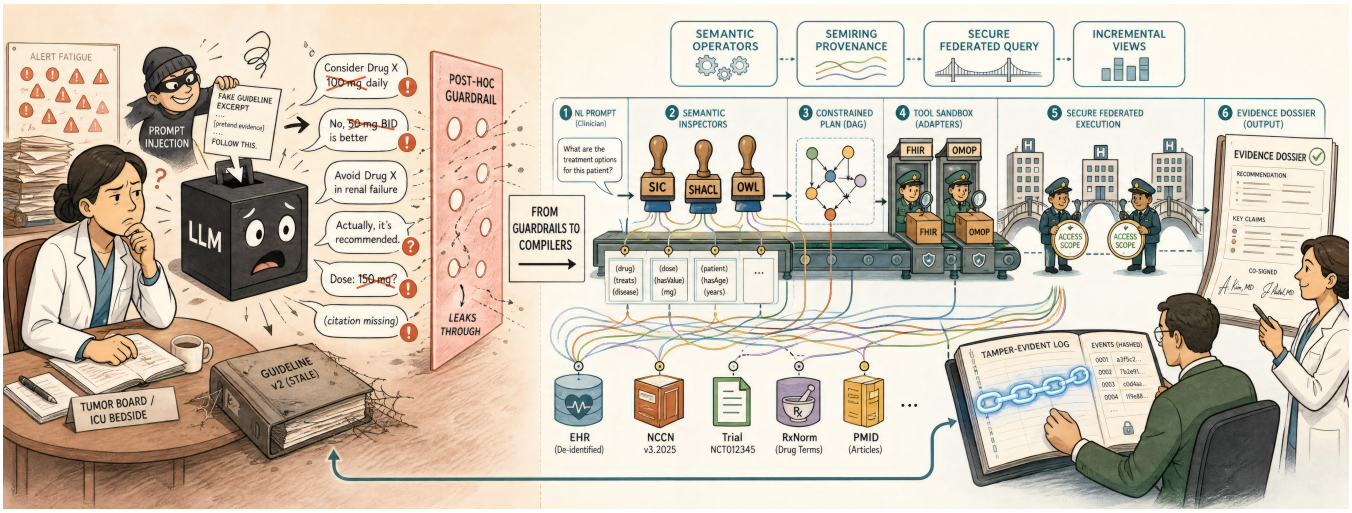

Figure 1: From guardrails to compilers: the runtime shift this paper argues for. *Left (today):* a black-box LLM sits behind a leaky post-hoc guardrail; alert-fatigue, a stale guideline binder, and a prompt-injection adversary all slip through. *Right (this paper):* the clinician's natural-language prompt is compiled into a constrained, provenance-tracked query plan: semantic inspectors (SIC, SHACL, OWL) check every step, a sandbox validates each tool output, secure federated execution honors per-institution access scopes over standard sources (EHR, NCCN, trials, RxNorm, PubMed), and a tamper-evident log yields an evidence dossier the clinician co-signs.

The data management community views the same problem differently. *What an agent does is execute a query plan over heterogeneous, governed, and uncertain data.* Under this view, safety and security enforcement belong in the planner and runtime, not after the text. The natural-language prompt is a *compiler frontend* for a constraint-aware physical plan [45, 77]; the LLM is merely one operator among many [25, 43, 47, 51, 53, 75]; and the relevant primitives already exist: semantic integrity constraints [40], fine-grained provenance [13, 14, 29, 56, 62], secure federated query processing [4, 55, 81], access control for data-backed systems [1, 49], and incremental view maintenance over evolving rules [9, 37]. What remains missing is their integration as the trusted computing base of a clinical agent. This reframes *agent security* as a query-optimization and constraint-enforcement problem rather than a text-filtering one. Concretely, it maps clinical-agent safety onto three problems the data-management community has studied for decades: *query optimization* (choosing and ordering operators, now including LLM calls, under a cost model), *integrity-constraint enforcement* (rejecting states that violate declared invariants), and *data integration* across governed, heterogeneous sources (schema mapping, ontology alignment, and access-controlled federation). An agent's sensitivity to the patient context, the active guideline version, and institutional access policy then becomes a compiled, auditable runtime artifact. We call this property *contextual intelligence*: an attribute of the runtime substrate, not of the language model that runs within it.

Running example. A tumor-board agent is asked, *"what systemic options does this patient have?"* for a case of metastatic colorectal cancer whose genomic report carries an activating KRAS mutation. A guardrail-wrapped LLM may propose the anti-EGFR antibody cetuximab [35], which is contraindicated for exactly this variant. In our runtime, the same request compiles into a plan: an access-scoped federated join fetches the patient's variant call and medication list; a semantic operator drafts candidate regimens; a semantic integrity constraint over `(variant, drug class, recommendation)` fires on the forbidden ($KRAS^{+}$, anti-EGFR) tuple and rewrites the plan to *warn-and-log* rather than recommend; and every step is stamped into a provenance-tracked dossier. We thread this example through the paper: the constraint appears as class C1 (§2), the primitives that enforce it in §3, and its compilation in §4.

A threat model for clinical agents. Four attack and failure surfaces, already observed in deployed prototypes, are structurally database problems. (i) *Prompt injection through retrieved evidence*: a poisoned trial abstract or a manipulated guideline excerpt steers the agent toward an unsafe recommendation [16, 30, 42, 94]. From the planner's perspective, this is untrusted input that must be schema-validated and confined before downstream operators consume it. (ii) *Tool-output corruption*: an upstream electronic health record (EHR) view, a faulty variant call, or a stale knowledge-graph entry introduces an incorrect tuple into the plan. Semantic constraints [40] and constraint-driven cleaning [59] detect the violation before it propagates. (iii) *Unauthorized composition*: an agent's sequence of otherwise permissible tool calls produces a join prohibited by institutional access policies [1, 49], even though no individual call violates policy in isolation. Capability-scoped planning [20] rules out such plans at compile time. (iv) *After-the-fact tampering*: an audit log is modified to retroactively justify a prior decision. Semiring-annotated, append-only provenance [19, 29, 50] renders such edits detectable. Each is what database research has long called an *integrity* problem; medical agents are its most consequential current setting. Figure 1 (right) places each defense on the runtime: stamped inspectors mitigate (i) and (ii), policy-customs at federation bridges defuse (iii), and tamper-evident chained logs address (iv).

Contributions. This is a vision paper whose goal is to convince the data-management community that clinical-agent safety is *its* problem to solve. Concretely, we (1) reframe the observed failures of clinical LLM agents as a *runtime* rather than a model deficiency, with a threat model whose four surfaces are classical integrity problems; (2) catalog six classes of constraint that clinical practice already imposes and map each onto a data-management primitive the community has built and validated (§2–§3); (3) sketch a four-layer constraint-first runtime that compiles a unified specification into a constrained, provenance-tracked query plan (§4); and (4) distill the resulting agenda into five open research questions (§5). We do not claim an implemented system; our aim is to fix the abstractions and encourage collaboration.

2 CONSTRAINTS FROM CLINICAL PRACTICE

Clinical decisions are made against a large but structured set of rules that today's agents rarely enforce. We organize these rules into six classes of constraint; each is documented in the clinical literature and has a corresponding harm or attack mode. What unifies them is that each is routinely observed in clinical practice yet is enforced today only by human vigilance, not by any runtime invariant on agent action.

C1: Guideline-derived contraindications and indications. NCCN, ASCO, ESMO, and Surviving Sepsis guidelines [23, 54] encode subsets of clinical decisions as near-deterministic rules. Our running example is canonical: cetuximab is contraindicated in metastatic colorectal cancer with activating KRAS mutations [35], a forbidden tuple over (variant, drug class, recommendation). Analogous deterministic structure governs HER2-targeted therapy [68], immune-checkpoint blockade in MSI-H/dMMR tumors [39], and KRAS G12C-targeted therapy [67]. ICU sepsis management has its own rule set: the SEP-1 three-hour bundle (lactate, blood culture before antibiotics, broad-spectrum antibiotics, fluid resuscitation), with measurable mortality increase per hour of delay [23, 66]. The computer-interpretable-guideline community has produced multiple formalisms (GLIF [6], PROforma [71], Asbru, GELLO, FHIR Clinical Reasoning + CQL [54, 70]); none integrates with semantic-operator or LLM-call planning at runtime.

C2: Ontological coding integrity. Clinical claims must reduce to coding systems the stack can interpret: SNOMED CT for conditions, LOINC for measurements, RxNorm for drugs, and UMLS as the binding glue [5, 11, 83]. Leading LLMs are poor medical coders, mis-mapping concepts at rates unacceptable for structured clinical pipelines [69]. Two failure modes follow. First, cross-domain coding errors (e.g., a measurement coded as a condition) and contradictions across sources [86] pass silently today but should become runtime errors; detecting such data deviations and biases in EHRs is itself an active data-management problem [92, 93]. Second, coding integrity is also a security boundary: free-text labels that are never grounded in an ontology are precisely the opening that prompt injection exploits to reach downstream operators.

C3: Provenance, corroboration, and tamper-evident audit. A multi-disciplinary tumor board chair will not co-sign a recommendation whose claims cannot be traced to specific records, guideline sections, and timestamps. Semiring-based provenance [13, 14, 29, 56, 62] provides the natural language, while VeriFact [15] and DECIDE-AI [80] show concrete clinical demand. The same machinery makes audit-trail tampering detectable via cryptographic hash chaining [19] and provenance-driven APT detection [50].

C4: Patient-specific longitudinal constraints. A patient on warfarin must not be prescribed a major-interaction drug; one with eGFR < 30 mL/min/1.73 m^2 cannot receive a renally-cleared agent at standard dosing; and one who has declined transfusion must not receive a transfusion-dependent plan. These constraints arise from joins between the agent output and the patient's longitudinal record, not generic facts. A 2024 stepped-wedge cluster RCT showed that ICU CDS alerts *tailored* to patient-specific longitudinal context, rather than generic drug-drug interaction (DDI) rules, reduced administration of high-risk drug combinations, whereas untailored alerts suffered approximately 90% override rates [3, 24, 84].

C5: Governance and federation policy. Cross-institutional cohort matching to identify "patients like this" [10, 91] is clinically valuable but constrained by HIPAA Safe Harbor, GDPR, jurisdiction-specific re-identification limits, and institutional data use agreements, forming a non-trivial policy lattice. Multi-institutional networks on the OMOP common data model [11, 58, 83] and secure federated query processing [4, 55, 81] demonstrate legal and technical feasibility, but their integration with semantic-operator pipelines remains ad hoc. An agent crossing institutions in a single trajectory must navigate the corresponding stack of institutional policies [1, 49]. Recent regulation now treats this lattice as legally binding for clinical AI systems, including the EU AI Act for high-risk medical AI [18], FDA Predetermined Change Control Plans [78], and other emerging guidance for non-deterministic clinical software [26, 73], so that policy compliance becomes an auditable obligation rather than a matter of best practice.

C6: Uncertainty and contestability. Epidemiologic priors come with confidence intervals, biomarker thresholds evolve with the literature, and oncologists may legitimately recommend treatment outside guideline parameters for refractory patients with no standard of care remaining, patients with contraindications to standard therapy, or other nuanced clinical situations requiring judgment. A runtime, therefore, cannot enforce a hard refusal uniformly: it must distinguish hard contraindications from soft norms while treating overrides as first-class auditable events. We draw the line explicitly. A *justified override* is a clinician-authored deviation from a *soft* norm (C6), attached to a recorded rationale and a responsible signer, that the specification permits and the audit layer preserves; it is a legitimate exercise of clinical judgment, not an error. An *agent failure*, by contrast, is a violation of a *hard* constraint (a C1 contraindication, a C2 coding error, a C4 interaction) that the runtime should have blocked, or an override lacking an authorized signer and rationale. The runtime's job is thus not to prevent all deviation but to guarantee that every deviation is classified, attributed, and logged, so that a justified override and a suppressed failure can never look alike after the fact. Recent clinical work on reasoning-oriented LLMs in medicine, RCTs of LLM influence on diagnosis, and analyses of consumer-LLM sycophancy under direct patient queries [21, 27, 82] all reinforce this; probabilistic and uncertain databases [22, 33] formalize the soft, judgment-bearing side of this lattice, and calibrating LLM agent confidences against such a formalism remains an open problem.

Clinical practice already operates within this lattice; what is missing is a runtime that enforces it on every agent action. Each constraint class above maps onto data-management primitives the community has already built and validated outside the clinical setting, and composing them, rather than scaling models, is what safety now depends on.

3 PRIMITIVES AS THE TRUSTED BASE

Semantic integrity constraints (SICs). Lee et al. [40] proposed a declarative framework for enforcing value, type, dependency, cardinality, and temporal constraints in AI-augmented pipelines. SICs cover most of C1, C2, and C4, while constraint-driven cleaning systems [7, 59, 64] extend the same principles to noisy upstream sources. For clinical use, what remains missing is integration with description-logic axioms over SNOMED class hierarchies, compilation strategies that determine whether a check belongs to pre-condition pruning vs. post-condition validation, and cost-aware ordering when checks have asymmetric latency and refusal costs. The longitudinal constraints of C4 push the framework further still: a rule such as "no renally-cleared drug at standard dose once eGFR falls below threshold" must compile into a dependency between the agent's proposed order and the most recent laboratory tuple, rather than a static value check.

Semantic operators with accuracy guarantees. LOTUS [53] and Palimpzest [45] show that LLM-call-based operators can be optimized with quality and cost guarantees, and CAESURA [77] extends this to multi-modal sources; data agents have since become a first-class community topic, reframing databases as a substrate for LLMs [25, 31, 43, 47, 51, 75]. Clinical workloads demand heterogeneous cost models, criticality-dependent guarantees, and constraint-driven operator reordering, for which constraint-driven NL2SQL [44, 60, 87] and clinical NL2SQL benchmarks [41] offer aligned starting points to build on. The distinguishing demand is that operator selection be criticality-aware, so that the planner never trades accuracy for latency on a contraindication check the way it might on a background literature summary.

Provenance and lineage. The semiring framework by Green, Karvounarakis, and Tannen [14, 29] provides a compositional algebra for tracking why each tuple in a query result was produced. Systems such as URSPRUNG [62], DPDS [13], and Smoke [56] extend the idea to data-science pipelines through automatic fine-grained lineage capture, while mlinspect and ArgusEyes [28, 63] screen ML pipelines using lineage as the substrate. These jointly provide a natural foundation for C3. The clinical setting, however, imposes stricter requirements: provenance must persist across federation boundaries, remain interpretable to non-technical reviewers in human-readable form, and be authenticated, since audit-log tampering is among the simplest attacks against deployed clinical agents [19]. A subtler requirement is that provenance survive summarization: when an LLM operator condenses several records into a single claim, the lineage of that claim must still resolve to the individual source tuples an auditor can re-examine.

Federated query under access control. SMCQL, Conclave, and Senate [4, 55, 81] demonstrate secure SQL execution over federated relational data, with motivating applications drawn explicitly from healthcare. Hippocratic Databases [1] and Qapla [49] formalized policy compliance for database-backed systems before the agent era, while the OHDSI ecosystem [58] and FHIR-Ontop-OMOP [83] demonstrate clinical-vocabulary and legal harmonization across institutions at scale. The open question is composition: enforcing the join of the policy lattice must rest with the planner, not the individual tools whose paths an agent might traverse.

Probabilistic and uncertain databases. ActivePDB and its predecessors, including MCDB [22, 33], provide a formal foundation for C6, including probabilistic constraints, Bayesian priors over biomarker thresholds, and confidence-bearing claims, whose connection to LLM agent outputs remains largely unexplored. Closing this gap means treating an operator's confidence as a probabilistic annotation that composes with the hard constraints, so that a soft threshold from C6 is evaluated explicitly and, once breached, surfaced as a calibrated warning rather than passing silently.

Materialized views over evolving guidelines. Clinical guidelines evolve irregularly: NCCN topics undergo multiple revisions per year, and ESMO has introduced "Living guidelines" for selected tumor types. A constraint-first runtime that compiles guideline excerpts into views must therefore support incremental view maintenance; modern IVM substrates such as DBSP [9] and DBToaster [37] apply directly to guideline-as-views compilation.

Authoring and validating the specification. A fair objection is that guidelines are not code: they contain ambiguity, exceptions, evolving recommendations, and passages that call for expert judgment, so encoding them as rigid rules risks a spurious sense of precision. Our design confronts this rather than assuming it away. First, only the *computable subset* is compiled: the near-deterministic contraindications and bundle steps of C1, with genuinely judgment-bearing passages left to the soft-norm and uncertainty machinery of C6 rather than forced into hard rules. Second, we do not expect each institution to formalize guidelines from scratch; the computer-interpretable-guideline community has spent two decades building formalisms and curation processes (GLIF, PROforma, FHIR Clinical Reasoning + CQL [6, 54, 70, 71]), which we reuse as *community-maintained, version-pinned "guideline-as-code" packages*. Third, validation becomes a first-class data-management task: a candidate package is a query one can test against retrospective cohorts, differentially compare across versions, and hold accountable through the same provenance and audit trail (C3) that governs live decisions, so that a mis-encoding surfaces as a measurable false-refusal or false-permit rate rather than a silent error. Curation, coverage, and drift of these packages at scale remain open (§5).

4 A CONSTRAINT-FIRST RUNTIME

The runtime has four layers (specification, planner, execution, audit), each exposing clinical concerns to clinicians and data concerns to data engineers behind deliberately narrow inter-layer interfaces.

Specification combines several layers: a SIC-style relational fragment [40], which LLM-assisted schema auditing [65] can help bootstrap; SHACL/OWL fragments over relevant ontologies; a Datalog-like rule layer for guideline excerpts curated as community-maintained guideline-as-code packages (version-pinned per institution, following HL7 FHIR Clinical Reasoning + CQL [6, 54, 70, 71]);

and a policy DSL naming the access scope of each agent capability [1, 49], a scope the planner may narrow but never widen.

Planner compiles the prompt with the active specification into a constrained DAG of tool calls (EHR queries, PrimeKG [12] and DDInter [84] lookups, literature retrieval, semantic-operator LLM calls), interleaved pre/mid/post-condition checks, and enforcement actions (refuse, warn-and-log, request-clinician-confirm). The cost model integrates classical query costs, LLM latency, and pricing [45, 53], per-tool capability scope [20], and a clinical-criticality dimension rendering contraindication checks non-skippable.

Execution wraps every tool call and LLM operator with automatic provenance capture, extending URSPRUNG-style techniques to agentic settings [13, 56, 62]. It coordinates federated access through OHDSI/FHIR endpoints [58, 83] with access-controlled query processing [4, 55, 81], and schema-validates tool outputs before downstream operators consume them [46, 61, 94]. Constraint violations interrupt execution and either suppress the failing branch or trigger a clinician override, recorded as a first-class event, extending human-in-the-loop task decomposition [90] to the agentic setting.

Audit emits, per agent decision, a replayable evidence dossier in the spirit of interpretable analytics for high-stakes clinical applications [89]: the active constraint set, the records and queries supporting each claim, LLM-operator outputs, constraint-trigger points, and clinician overrides, keyed by semiring-annotated provenance [29] and stored append-only with signed checkpoints in the spirit of in-toto and history-tree tamper-evident logging [19, 74]. Aligned with DECIDE-AI [80] and regulatory guidance [18, 73, 78], a tumor-board chair signs the dossier that an auditor replays.

Overhead and trade-offs. These layers are not free, and the design is deliberately shaped by their cost. Most constraint checks are cheap relational or ontology lookups that the planner pushes down and evaluates as preconditions, so they prune rather than add work; the expensive operators are the LLM calls the agent already makes. The cost model treats safety as a first-class dimension: contraindication checks (C1, C4) are marked non-skippable and paid regardless of latency, whereas soft checks are ordered by the same cost/quality optimization used for semantic operators [45, 53]. Provenance capture and append-only logging add storage and a bounded per-step write, the accepted price of auditability. The real trade-off is therefore not raw speed but *expressiveness versus enforceability*: constraints must be rich enough to catch real harms yet restricted enough to compile and check within the clinical decision cycle, and quantifying this frontier on logged MTB and ICU workloads is an explicit open question (§5).

Position. Most current clinical-agent prototypes [8, 17, 34, 85, 88] layer general-purpose agent frameworks atop an EHR query layer; the runtime we sketch sits beneath them, compiling tool calls and LLM operators into the constrained DAG, much as semantic-operator systems [45, 53] relate to ad hoc LLM pipelines.

5 OPEN RESEARCH QUESTIONS

Q1. Heterogeneous constraint compilation. How can a mixed specification (SIC [40] + SHACL/OWL + Datalog + probabilistic [22, 33]) compile into a single executable plan with provable enforcement? When do OWL axioms admit precondition pushdown rather than a dedicated reasoner call, and when do Datalog guideline rules such as the SEP-1 three-hour bundle [23] admit forbidden-tuple compilation versus runtime instantiation against the longitudinal patient record?

Q2. Tamper-evident provenance and override semantics. Define a provenance algebra in which clinician overrides are first-class events, queryable by auditors, statistically analyzable, and tamper-evident under an honest-but-curious storage adversary, building on the semiring framework [29], in-toto [74], history trees [19], and provenance-driven security analysis [50]. The conceptual challenge is defining a *justified* override: one that preserves clinical judgment while remaining automatable against the specification and aligned with emerging regulatory expectations [18, 26, 73, 78] for non-deterministic clinical software.

Q3. Cost-aware planning with clinical criticality and capability scope. Extend semantic-operator optimization [45, 53] with a criticality dimension, asymmetric refusal costs, and a capability-scope term penalizing tools whose access privileges exceed the active prompt's needs. Ground the cost model in logged MTB and ICU workloads, where clinical outcomes and time-to-treatment are already instrumented as operational signals [38, 66]. A subtlety is that criticality is not a fixed per-constraint weight: the same check may be non-skippable inside an ICU sepsis bundle, yet advisory at an outpatient follow-up, so criticality must itself be a function of the clinical context that the planner reads from the specification.

Q4. Capability-scoped federated execution. Compose institutional access policies across an agent's execution trajectory spanning OMOP/FHIR federations and secure federated query backends [4, 55, 58, 81]; plans whose composed access scope exceeds the permissions granted by any participating institution should be rejected at planning time rather than at execution time. This calls for a compositional semantics of access scope, in which the scope of a plan is derived from the scopes of its operators and checked against each institution's grant before any tuple crosses a federation boundary, rather than being audited only after the fact.

Q5. Incremental compliance views under guideline updates. Treat versioned guideline packages as base relations and active prompts as queries, maintaining compliance under version changes through incremental view maintenance [9, 37]. Pinning the compliance view active at the time of a past decision is itself a security property that auditors must reconstruct. The open challenge is to bound the recomputation a revision triggers, so that compliance can be re-established online for the affected plans without recompiling every active plan from scratch.

6 CONCLUSION

A clinical agent's safety, security, and contextual intelligence come from the plan it compiles, not from the model that proposes it. The building blocks already exist in data management and clinical informatics: a runtime can compile each agent action into a constrained, provenance-tracked, and capability-scoped query plan, then emit a dossier that the attending clinician signs and an auditor can replay. This extends the database community's long-standing role as a trusted computing base to language-model agents, and lets

us inspect an agent's behavior in its dossier instead of inferring it from a model card. The six constraint classes we catalog and the primitives we map them onto are not a wish list but an inventory of components the community has already built, tested, and deployed outside medicine; what is new is the demand that they compose into a single compiled artifact whose guarantees a clinician can sign and an auditor can replay. The bedside merely forces the question first; the same substrate fits any setting where agents act over governed, uncertain data under consequential rules. Realizing it will require the clinical and data-management communities to co-design the specification language, the cost model, and the audit format together, which is exactly the kind of collaboration this workshop exists to foster. A natural first step is not a larger clinical model but a minimal end-to-end slice of this runtime (one constraint class, one federation boundary, one signed dossier) exercised on a real molecular-tumor-board or ICU workload and measured for what it costs and what it catches. The next advance at the bedside is thus not a better model but a runtime that compiles whatever model we use into plans we can audit.

7 AUTHORS

Kaiping Zheng (data management). Senior Research Fellow at the School of Computing, National University of Singapore (NUS), working at the intersection of data management, machine learning, and clinical medicine. Her research develops robust medical data-management methods that turn noisy electronic health records into trustworthy clinical signals, with responsible-AI techniques aligned with clinical reasoning.

Horng-Ruey Chua (biomedical). Head and Senior Consultant in the Division of Nephrology, National University Hospital (NUH), and Adjunct Associate Professor at NUS Yong Loo Lin School of Medicine (YLLSoM).

Si Wei Kheok (biomedical). Senior Consultant in Diagnostic Radiology at Singapore General Hospital, Director of Head and Neck Radiology and of MRI Patient Care Services, and Deputy Head of Neuroradiology, with appointments at Duke-NUS, NTU, and NUS YLLSoM.

Kee Yuan Ngiam (biomedical). Head and Senior Consultant in the Division of General Surgery (Endocrine and Thyroid Surgery), Department of Surgery, NUH; Senior Consultant in Surgical Oncology at the National University Cancer Institute, Singapore (NCIS); and Adjunct Professor of Surgery at NUS YLLSoM.

Marcus Chun Jin Tan (biomedical). Consultant in Ophthalmology at NUH and Deputy Group Chief Technology Officer of the National University Health System.

Robert John Walsh (biomedical). Consultant in Haematology-Oncology at NCIS.

ACKNOWLEDGMENT OF GENERATIVE-AI USE

The illustration in Figure 1 was produced with a text-to-image model from an author-written prompt and then curated by the authors. Large language models were also used to assist with copy-editing and to smooth prose. All technical claims, the framework, the six constraint classes, and the overall research agenda are entirely the authors' own original work.

REFERENCES

[1] Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu. 2002. Hippocratic databases. In *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*. Elsevier, 143–154.

[2] Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ digital medicine* 8, 1 (2025), 274.

[3] Tinka Bakker, Joanna E Klopotowska, Dave A Dongelmans, Saeid Eslami, Wytze J Vermeijden, Stefaan Hendriks, Julia Ten Cate, Attila Karakus, Ilse M Purmer, Sjoerd HW van Bree, et al. 2024. The effect of computerised decision support alerts tailored to intensive care on the administration of high-risk drug combinations, and their monitoring: a cluster randomised stepped-wedge trial. *The Lancet* 403, 10425 (2024), 439–449.

[4] Johes Bater, Gregory Elliott, Craig Eggen, Satyender Goel, Abel Kho, and Jennie Rogers. 2017. SMCQL: secure querying for federated databases. *Proceedings of the VLDB Endowment* 10, 6 (2017), 673–684.

[5] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl_1 (2004), D267–D270.

[6] Aziz A Boxwala, Mor Peleg, Samson Tu, Omolola Ogunyemi, Qing T Zeng, Dongwen Wang, Vimla L Patel, Robert A Greenes, and Edward H Shortliffe. 2004. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. *Journal of biomedical informatics* 37, 3 (2004), 147–161.

[7] Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2019. Data validation for machine learning. *Proceedings of machine learning and systems* 1 (2019), 334–347.

[8] Dechao Bu, Jingbo Sun, Kun Li, Zihao He, Wei Huang, Jinlin Hu, Shanshan Zhang, Shuangshuang Lei, Peipei Huo, Zhihao Wang, et al. 2026. Empowering AI data scientists using a multi-agent LLM framework with self-evolving capabilities for autonomous, tool-aware biomedical data analyses. *Nature Biomedical Engineering* (2026), 1–16.

[9] Mihai Budiu, Tej Chajed, Frank McSherry, Leonid Ryzhyk, and Val Tannen. 2023. DBSP: Automatic Incremental View Maintenance for Rich Query Languages. *Proceedings of the VLDB Endowment* 16, 7 (2023), 1601–1614.

[10] Qingpeng Cai, Kaiping Zheng, H. V. Jagadish, Beng Chin Ooi, and James W. L. Yip. 2024. CohortNet: Empowering Cohort Discovery for Interpretable Healthcare Analytics. *Proceedings of the VLDB Endowment* 17, 10 (2024), 2487–2500.

[11] Tiffany J Callahan, Adrianne L Stefanski, Jordan M Wyrwa, Chenjie Zeng, Anna Ostropolets, Juan M Banda, William A Baumgartner Jr, Richard D Boyce, Elena Casiraghi, Ben D Coleman, et al. 2023. Ontologizing health systems data at scale: making translational discovery a reality. *npj Digital Medicine* 6, 1 (2023), 89.

[12] Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific data* 10, 1 (2023), 67.

[13] Adriane Chapman, Luca Lauro, Paolo Missier, and Riccardo Torlone. 2022. DPDS: assisting data science with data provenance. *Proceedings of the VLDB Endowment* 15, 12 (2022), 3614–3617.

[14] James Cheney, Laura Chiticariu, and Wang-Chiew Tan. 2009. Provenance in databases: Why, how, and where. *Foundations and trends in databases* 1, 4 (2009), 379–474.

[15] Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, et al. 2025. Verifying Facts in Patient Care Documents Generated by Large Language Models Using Electronic Health Records. *NEJM AI* 3, 1 (2025), AIdbp2500418.

[16] Jan Clusmann, Dyke Ferber, Isabella C Wiest, Carolin V Schneider, Titus J Brinker, Sebastian Foersch, Daniel Truhn, and Jakob Nikolas Kather. 2025. Prompt injection attacks on vision language models in oncology. *Nature Communications* 16, 1 (2025), 1239.

[17] Bernardo G Collaco, Syed Ali Haider, Srinivasagam Prabha, Cesar A Gomez-Cabello, Ariana Genovese, Nadia G Wood, Sanjay P Bagaria, Narayanan Gopala, Cui Tao, and Antonio Jorge Forte. 2026. The role of agentic artificial intelligence in healthcare: a scoping review. *npj Digital Medicine* (2026).

[18] European Commission. 2026. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[19] Scott A Crosby and Dan S Wallach. 2009. Efficient data structures for tamper-evident logging.. In *USENIX security symposium*. 317–334.

[20] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. 2025. Defeating prompt injections by design. *arXiv preprint arXiv:2503.18813* (2025).

[21] Rachel L Draelos, Samina Afreen, Barbara Blasko, Tiffany L Brazile, Natasha Chase, Dimple Patel Desai, Jessica Evert, Heather L Gardner, Lauren Herrmann, Aswathy Vaikom House, et al. 2026. Large language models provide unsafe answers to patient-posed medical questions. *npj Digital Medicine* (2026).

[22] Osnat Drien, Matanya Freiman, and Yael Amsterdamer. 2022. Activepdb: active probabilistic databases. *Proceedings of the VLDB Endowment* 15, 12 (2022), 3638–3641.

[23] Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M Coopersmith, Craig French, Flávia R Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C Prescott, et al. 2021. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Critical care medicine* 49, 11 (2021), e1063–e1143.

[24] Mariano Felisberto, Geovana dos Santos Lima, Ianka Cristina Celuppi, Miliane dos Santos Fantonelli, Wagner Luiz Zanotto, Júlia Meller Dias de Oliveira, Eduarda Talita Bramorski Mohr, Ranieri Alves Dos Santos, Daniel Henrique Scandolara, Célio Luiz Cunha, et al. 2024. Override rate of drug-drug interaction alerts in clinical decision support systems: A brief systematic review and meta-analysis. *Health Informatics Journal* 30, 2 (2024), 14604582241263242.

[25] Raul Castro Fernandez, Aaron J Elmore, Michael J Franklin, Sanjay Krishnan, and Chenhao Tan. 2023. How large language models will disrupt data management. *Proceedings of the VLDB Endowment* 16, 11 (2023), 3302–3309.

[26] Oscar Freyer, Rebecca Mathias, Hannah Sophie Muti, Henry Orlovsky, Stephan Buch, Max Ostermann, Anett Schönfelder, Akira-Sebastian Poncette, Adel Bassily-Marcus, and Stephen Gilbert. 2026. The regulation of artificial intelligence in intensive care units: from narrow tools to generalist systems. *NPJ Digital Medicine* (2026).

[27] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open* 7, 10 (2024), e2440969.

[28] Stefan Grafberger, Julia Stoyanovich, and Sebastian Schelter. 2021. Lightweight inspection of data preprocessing in native machine learning pipelines. In *Conference on Innovative Data Systems Research (CIDR)*.

[29] Todd J Green, Grigoris Karvounarakis, and Val Tannen. 2007. Provenance semirings. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 31–40.

[30] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security*. 79–90.

[31] Alon Halevy and Jane Dwivedi-Yu. 2023. Learnings from data integration for augmented language models. *arXiv preprint arXiv:2304.04576* (2023).

[32] Tianyu Han, Sven Nebelung, Firas Khader, Tianci Wang, Gustav Müller-Franzes, Christiane Kuhl, Sebastian Försch, Jens Kleesiek, Christoph Haarburger, Keno K Bressem, et al. 2024. Medical large language models are susceptible to targeted misinformation attacks. *NPJ digital medicine* 7, 1 (2024), 288.

[33] Ravi Jampani, Fei Xu, Mingxi Wu, Luis Leopoldo Perez, Christopher Jermaine, and Peter J Haas. 2008. Mcdb: a monte carlo approach to managing uncertain data. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 687–700.

[34] Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. 2025. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. *Nejm Ai* 2, 9 (2025), AIdbp2500144.

[35] Christos S Karapetis, Shirin Khambata-Ford, Derek J Jonker, Chris J O'Callaghan, Dongsheng Tu, Niall C Tebbutt, R John Simes, Haji Chalchal, Jeremy D Shapiro, Sonia Robitaille, et al. 2008. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine* 359, 17 (2008), 1757–1765.

[36] Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, et al. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777* (2025).

[37] Christoph Koch, Yanif Ahmad, Oliver Kennedy, Milos Nikolic, Andres Nötzli, Daniel Lupei, and Amir Shaikhha. 2014. DBToaster: higher-order delta processing for dynamic, frequently fresh views. *The VLDB Journal* 23, 2 (2014), 253–278.

[38] Kara L Larson, Bin Huang, Heidi L Weiss, Pam Hull, Philip M Westgate, Rachel W Miller, Susanne M Arnold, and Jill M Kolesar. 2021. Clinical outcomes of molecular tumor boards: a systematic review. *JCO precision oncology* 5 (2021), 1122–1132.

[39] Dung T Le, Jennifer N Uram, Hao Wang, Bjarne R Bartlett, Holly Kemberling, Aleksandra D Eyring, Andrew D Skora, Brandon S Luber, Nilofer S Azad, Dan Laheru, et al. 2015. PD-1 blockade in tumors with mismatch-repair deficiency. *New England Journal of Medicine* 372, 26 (2015), 2509–2520.

[40] Alexander W Lee, Justin Chan, Michael Fu, Nicolas Kim, Akshay Mehta, Deepti Raghavan, and Uğur Çetintemel. 2025. Semantic Integrity Constraints: Declarative Guardrails for AI-Augmented Data Processing Systems. *Proceedings of the VLDB Endowment* 18, 11 (2025), 4073–4080.

[41] Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 2022. Ehrsql: A practical text-to-sql benchmark for electronic health records. *Advances in Neural Information Processing Systems* 35 (2022), 15589–15601.

[42] Ro Woon Lee, Tae Joon Jun, Jeong-Moo Lee, Soo Ick Cho, Hyung Jun Park, and Jungyo Suh. 2025. Vulnerability of Large Language Models to Prompt Injection When Providing Medical Advice. *JAMA Network Open* 8, 12 (2025), e2549963.

[43] Guoliang Li, Jiayi Wang, Chenyang Zhang, and Jiannan Wang. 2025. Data+AI: LLM4Data and Data4LLM. In *Companion of the 2025 International Conference on Management of Data*. 837–843.

[44] Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, et al. 2023. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems* 36 (2023), 42330–42357.

[45] Chunwei Liu, Matthew Russo, Michael Cafarella, Lei Cao, Peter Baile Chen, Zui Chen, Michael Franklin, Tim Kraska, Samuel Madden, Rana Shahout, et al. 2025. Palimpzest: Optimizing ai-powered analytics with declarative query processing. In *Proceedings of the Conference on Innovative Database Research (CIDR)*. 2.

[46] Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. 2024. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*. 1831–1847.

[47] Yuyu Luo, Guoliang Li, Ju Fan, and Nan Tang. 2026. Data Agents: Levels, State of the Art, and Open Problems. *arXiv preprint arXiv:2602.04261* (2026).

[48] Nikita Mehandru, Brenda Y Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J Butte, and Ahmed Alaa. 2024. Evaluating large language models as agents in the clinic. *NPJ digital medicine* 7, 1 (2024), 84.

[49] Aastha Mehta, Eslam Elnikety, Katura Harvey, Deepak Garg, and Peter Druschel. 2017. Qapla: Policy compliance for database-backed systems. In *26th USENIX Security Symposium (USENIX Security 17)*. 1463–1479.

[50] Sadegh M Milajerdi, Rigel Gjomemo, Birhanu Eshete, and VN Venkatakrishnan. 2019. HOLMES: real-time apt detection through correlation of suspicious information flows. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 1137–1152.

[51] Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. 2022. Can Foundation Models Wrangle Your Data? *Proceedings of the VLDB Endowment* 16, 4 (2022), 738–746.

[52] Mahmud Omar, Vera Sorin, Jeremy D Collins, David Reich, Robert Freeman, Nicholas Gavin, Alexander Charney, Lisa Stump, Nicola Luigi Bragazzi, Girish N Nadkarni, et al. 2025. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Communications Medicine* 5, 1 (2025), 330.

[53] Liana Patel, Siddharth Jha, Melissa Z Pan, Harshit Gupta, Parth Asawa, Carlos Guestrin, and Matei Zaharia. 2025. Semantic Operators and Their Optimization: Towards AI-Based Data Analytics with Accuracy Guarantees. *Proc. VLDB Endow* 18, 11 (2025), 4171–4184.

[54] Mor Peleg. 2013. Computer-interpretable clinical guidelines: a methodological review. *Journal of biomedical informatics* 46, 4 (2013), 744–763.

[55] Rishabh Poddar, Sukrit Kalra, Avishay Yanai, Ryan Deng, Raluca Ada Popa, and Joseph M Hellerstein. 2021. Senate: A Maliciously-Secure MPC Platform for Collaborative Analytics. In *30th USENIX Security Symposium (USENIX Security 21)*. 2129–2146.

[56] Fotis Psallidas and Eugene Wu. 2018. Smoke: fine-grained lineage at interactive speed. *Proceedings of the VLDB Endowment* 11, 6 (2018), 719–732.

[57] Ihsan Ayyub Qazi, Ayesha Ali, Asad Ullah Khawaja, Muhammad Junaid Akhtar, Ali Zafar Sheikh, and Muhammad Hamad Alizai. 2026. Automation Bias in Large Language Model–Assisted Diagnostic Reasoning among Physicians Trained in AI Literacy—A Randomized Clinical Trial. *NEJM AI* 3, 5 (2026), AIoa2501001.

[58] Christian Reich, Anna Ostropolets, Patrick Ryan, Peter Rijnbeek, Martijn Schuemie, Alexander Davydov, Dmitry Dymshyts, and George Hripcsak. 2024. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. *Journal of the American Medical Informatics Association* 31, 3 (2024), 583–590.

[59] Theodoros Rekatsinas, Xu Chu, Ihab F Ilyas, and Christopher Ré. 2017. HoloClean: holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment* 10, 11 (2017), 1190–1201.

[60] Tonghui Ren, Chen Ke, Yuankai Fan, Yinan Jing, Zhenying He, Kai Zhang, and X Sean Wang. 2025. The Power of Constraints in Natural Language to SQL Translation. *Proceedings of the VLDB Endowment* 18, 7 (2025), 2097–2111.

[61] Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris Maddison, and Tatsunori Hashimoto. 2024. Identifying the risks of lm agents with an lm-emulated sandbox. In *International Conference on Learning Representations*, Vol. 2024. 27031–27098.

[62] Lukas Rupprecht, James C Davis, Constantine Arnold, Yaniv Gur, and Deepavali Bhagwat. 2020. Improving Reproducibility of Data Science Pipelines through Transparent Provenance Capture. *Proc. VLDB Endow.* 13, 12 (2020), 3354–3368.

[63] Sebastian Schelter, Stefan Grafberger, Shubha Guha, Olivier Sprangers, Bojan Karlaš, and Ce Zhang. 2022. Screening native ml pipelines with "arguseyes". In *Conference on Innovative Data Systems Research. CIDR*.

[64] Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. 2018. Automating large-scale data quality verification. *Proceedings of the VLDB Endowment* 11, 12 (2018), 1781–1794.

[65] Antony Seabra, Claudio Cavalcante, Nicolaas Ruberg, and Sergio Lifschitz. [n.d.]. Towards an SLM-based Auditing of Relational Schemas and Data Quality for Practical Data Governance. *Proceedings of the VLDB Endowment. ISSN* 2150

([n. d.]), 8097.

[66] Christopher W Seymour, Foster Gesten, Hallie C Prescott, Marcus E Friedrich, Theodore J Iwashyna, Gary S Phillips, Stanley Lemeshow, Tiffany Osborn, Kathleen M Terry, and Mitchell M Levy. 2017. Time to treatment and mortality during mandated emergency care for sepsis. *New England Journal of Medicine* 376, 23 (2017), 2235–2244.

[67] Ferdinandos Skoulidis, Bob T Li, Grace K Dy, Timothy J Price, Gerald S Falchook, Jürgen Wolf, Antoine Italiano, Martin Schuler, Hossein Borghaei, Fabrice Barlesi, et al. 2021. Sotorasib for lung cancers with KRAS p. G12C mutation. *New England Journal of Medicine* 384, 25 (2021), 2371–2381.

[68] Dennis J Slamon, Brian Leyland-Jones, Steven Shak, Hank Fuchs, Virginia Paton, Alex Bajamonde, Thomas Fleming, Wolfgang Eiermann, Janet Wolter, Mark Pegram, et al. 2001. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England journal of medicine* 344, 11 (2001), 783–792.

[69] Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. 2024. Large language models are poor medical coders—benchmarking of medical code querying. *Nejm Ai* 1, 5 (2024), AIdbp2300040.

[70] Howard R Strasberg, Bryn Rhodes, Guilherme Del Fiol, Robert A Jenders, Peter J Haug, and Kensaku Kawamoto. 2021. Contemporary clinical decision support standards using health level seven international fast healthcare interoperability resources. *Journal of the American Medical Informatics Association* 28, 8 (2021), 1796–1806.

[71] David R Sutton and John Fox. 2003. The syntax and semantics of the PROforma guideline modeling language. *Journal of the American Medical Informatics Association* 10, 5 (2003), 433–443.

[72] David Tamborero, Rodrigo Dienstmann, Maan Haj Rachid, Jorrit Boekel, Adria Lopez-Fernandez, Markus Jonsson, Ali Razzak, Irene Braña, Luigi De Petris, Jeffrey Yachnin, et al. 2022. The Molecular Tumor Board Portal supports clinical decisions and automated reporting for precision oncology. *Nature cancer* 3, 2 (2022), 251–261.

[73] Caitlyn Tan, Dinesh Visva Gunasekeran, Cheng Ooi Low, Gabrielle Sze Yee Sim, Danella Yaoxin Foo, Robert JT Morris, and Tien Yin Wong. 2026. Regulation of clinical Artificial Intelligence (AI) in the Age of Agents: Unconfined Non-Deterministic Clinical Software (UNDCS) systems for healthcare. *npj Digital Medicine* 9, 1 (2026), 186.

[74] Santiago Torres-Arias, Hammad Afzali, Trishank Karthik Kuppusamy, Reza Curtmola, and Justin Cappos. 2019. in-toto: Providing farm-to-table guarantees for bits and bytes. In *28th USENIX Security Symposium (USENIX Security 19)*. 1393–1410.

[75] Immanuel Trummer. 2022. From BERT to GPT-3 codex: harnessing the potential of very large language models for data management. *Proceedings of the VLDB Endowment* 15, 12 (2022), 3770–3773.

[76] Apostolia M Tsimberidou, Michael Kahle, Henry Hiep Vo, Mehmet A Baysal, Amber Johnson, and Funda Meric-Bernstam. 2023. Molecular tumour boards—current and future considerations for precision oncology. *Nature reviews Clinical oncology* 20, 12 (2023), 843–863.

[77] Matthias Urban and Carsten Binnig. 2024. CAESURA: Language Models as Multi-Modal Query Planners. In *CIDR*. www.cidrdb.org.

[78] U.S. Food and Drug Administration. 2025. https://www.fda.gov/media/166704/download.

[79] Robin van Kessel, Michael Anderson, Brian McMillan, Marc R Matthews, Paul Rust, Pauline Pearcy, Khurram Nasir, and Elias Mossialos. 2026. Omission and hallucination prevalence of clinical guidelines in diagnostic large language model outputs. *BMJ Health & Care Informatics* 33, 1 (2026), e101959.

[80] Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A Clifton, Gary S Collins, Spiros Denaxas, Alastair K Denniston, Livia Faes, Bart Geerts, Mudathir Ibrahim, et al. 2022. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *bmj* 377 (2022).

[81] Nikolaj Volgushev, Malte Schwarzkopf, Ben Getchell, Mayank Varia, Andrei Lapets, and Azer Bestavros. 2019. Conclave: secure multi-party computation on big data. In *Proceedings of the Fourteenth EuroSys Conference 2019*. 1–18.

[82] Xiaofei Wang, Zhuxin Xiong, Ke Zou, Sahana Srinivasan, Thaddaeus Wai Soon Lo, Yilan Wu, Minjie Zou, Nan Liu, Fares Antaki, Weizhi Ma, et al. 2026. Reasoning-driven large language models in medicine: opportunities, challenges, and the road ahead. *The Lancet Digital Health* (2026).

[83] Guohui Xiao, Emily Pfaff, Eric Prud'hommeaux, David Booth, Deepak K Sharma, Nan Huo, Yue Yu, Nansu Zong, Kathryn J Ruddy, Christopher G Chute, et al. 2022. FHIR-Ontop-OMOP: Building clinical knowledge graphs in FHIR RDF with the OMOP Common data Model. *Journal of biomedical informatics* 134 (2022), 104201.

[84] Guoli Xiong, Zhijiang Yang, Jiacai Yi, Ningning Wang, Lei Wang, Huimin Zhu, Chengkun Wu, Aiping Lu, Xiang Chen, Shao Liu, et al. 2022. DDInter: an online drug–drug interaction database towards improving clinical decision-making and patient safety. *Nucleic acids research* 50, D1 (2022), D1200–D1207.

[85] Tiantian Yang, Yihang Xiao, Zhijie Bao, Jianye Hao, and Jiajie Peng. 2025. The rise and potential opportunities of large language model agents in bioinformatics and biomedicine. *Briefings in Bioinformatics* 26, 6 (2025), bbaf601.

[86] Boya Zhang, Alban Bornet, Rui Yang, Nan Liu, and Douglas Teodoro. 2026. HealthContradict: Evaluating biomedical knowledge conflicts in language models. *npj Digital Medicine* (2026).

[87] Fuheng Zhao, Shaleen Deep, Fotis Psallidas, Avrilia Floratou, Divyakant Agrawal, and Amr El Abbadi. 2024. Sphinteract: Resolving Ambiguities in NL2SQL through User Interaction. *Proceedings of the VLDB Endowment* 18, 4 (2024), 1145–1158.

[88] Weike Zhao, Chaoyi Wu, Yanjie Fan, Pengcheng Qiu, Xiaoman Zhang, Yuze Sun, Xiao Zhou, Shuju Zhang, Yu Peng, Yanfeng Wang, et al. 2026. An agentic system for rare disease diagnosis with traceable reasoning. *Nature* (2026), 1–10.

[89] Kaiping Zheng, Shaofeng Cai, Horng-Ruey Chua, Wei Wang, Kee Yuan Ngiam, and Beng Chin Ooi. 2020. TRACER: A Framework for Facilitating Accurate and Interpretable Analytics for High Stakes Applications. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD)*.

[90] Kaiping Zheng, Gang Chen, Melanie Herschel, Kee Yuan Ngiam, Beng Chin Ooi, and Jinyang Gao. 2021. PACE: Learning Effective Task Decomposition for Human-in-the-loop Healthcare Delivery. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD)*.

[91] Kaiping Zheng, Horng-Ruey Chua, Melanie Herschel, H. V. Jagadish, Beng Chin Ooi, and James W. L. Yip. 2024. Exploiting Negative Samples: A Catalyst for Cohort Discovery in Healthcare Analytics. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

[92] Kaiping Zheng, Horng-Ruey Chua, and Beng Chin Ooi. 2025. Detecting Data Deviations in Electronic Health Records. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[93] Kaiping Zheng, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, and James W. L. Yip. 2017. Resolving the Bias in Electronic Medical Records. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[94] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2025. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. In *34th USENIX Security Symposium (USENIX Security 25)*. 3827–3844.