WHEN SPECIFICATIONS ARE ONLY PARTIALLY FORMALIZED: TOWARDS RELIABLE CLINICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern coding agents succeed when their environments provide reliable, machine-checkable feedback that can constrain model behavior. Software development exemplifies this regime: executable artifacts, compilation, testing, and execution create dense feedback signals that make actions verifiable and correctable. Clinical reasoning presents a different challenge. Clinical guidelines provide rich specifications, but only part of their content can be converted into machine-checkable structure, while the remainder requires context-dependent semantic judgment. The challenge is therefore to operationalize the executable portions of clinical specifications without forcing inherently semantic decisions into rigid rules. We introduce GEVHarness, a guideline harness for reliable clinical reasoning under partially formalized specifications. GEVHarness partitions guidelines into machine-checkable structure, which receives deterministic runtime authority over explicit state, and contextual judgments, which remain model-mediated. The model interprets evidence, while the runtime decides what follows, withholds commitment when evidence is incomplete, and makes every trajectory verifiable by replay. Across eight frozen benchmark panels spanning workflow control, open clinical reasoning, and evidence control, GEVHarness is the strongest interface among matched text and model-guided structured baselines for both evaluated backbone blocks; it improves the cases whose correct decisions change with evidence, and Controlled interventions further show that the value of formalization depends not only on how much of the specification is executable, but on where runtime authority is placed. These results position partially formalized specifications as a distinct inference-harness regime: a specification need not be a complete execution oracle to govern inference. More broadly, clinical reasoning exposes a persistent tension between *semantic flexibility* and *executable reliability*; GEVHarness offers one concrete step toward understanding how models, specifications, and runtimes should share control.

1 INTRODUCTION

The capability of modern agents increasingly depends not only on the model, but on the environment that surrounds it: the tools, runtimes, and feedback mechanisms that mediate actions, track state, and feed machine-checkable outcomes back into inference (Yang et al., 2024; Xie et al., 2024). Coding agents make this especially visible. Compilers, tests, type systems, and program execution provide dense external signals that can reject invalid actions and bound a model's freedom (Yao et al., 2022; Yang et al., 2024). In this setting, the environment itself supplies the executable feedback needed to transform an otherwise unconstrained model into a reliable agent. Many consequential reasoning domains are less accommodating. They often have rich specifications, such as guidelines, protocols, policies, or regulations, but no compiler or test suite that can mechanically determine whether an inference is sound. **Can such specifications still support reliable reasoning when they are only partially formalized?**

Clinical reasoning provides a particularly sharp instance of this problem. Clinical practice guidelines encode substantial normative structure: eligibility criteria, thresholds, required assessments, exceptions, sequencing constraints, and recommended actions (Boxwala et al., 2004; Brandt et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

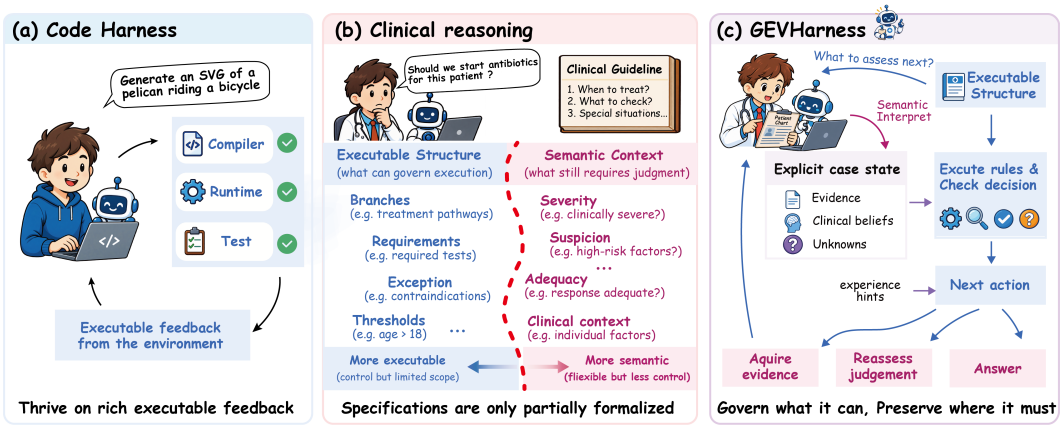


Figure 1: Rethinking inference harnesses from code to medicine. Unlike software specifications, clinical guidelines contain only partially formalized specifications; GEVHarness assigns runtime authority to machine-checkable structure while preserving context-dependent semantic judgment.

2020). Yet they also rely on judgments whose meaning depends on context. A condition such as age > 65, a laboratory threshold, or the presence of a medication can be evaluated directly from typed patient facts; whether a presentation is clinically suspicious, symptoms are severe, or a response is adequate cannot. **Clinical guidelines are therefore partially formalized specifications**: some of their content can directly constrain inference, while the remainder still requires semantic interpretation. The challenge is to **make the former operational without pretending that the latter has become mechanically decidable.**

Existing approaches largely use specifications in one of two ways (Li et al., 2025; Deng et al., 2026; Zhao et al., 2025; Sohn et al., 2025; Li et al., 2026). **One treats them as information**: guideline text or structured representations are placed in context, retrieved when relevant, or exposed as trees for a model to navigate. This preserves semantic flexibility, but leaves control of inference with the model; access to the correct guideline does not by itself prevent skipped assessments, unsupported transitions, ignored exceptions, or premature decisions. **The other treats the specification as a program**, translating as much of it as possible into executable logic. This gains deterministic control, but eventually requires contextual judgments to be replaced by hard operational proxies. These approaches differ in how aggressively they formalize the specification, but share a deeper limitation: neither explains how a specification that is incomplete as a program can nevertheless **exercise runtime authority over model inference.**

We introduce **GEVHarness, a harness for reasoning under partially formalized clinical specifications**. GEVHarness compiles a guideline into a typed intermediate representation that partitions predicates into those decidable from structured patient facts and those requiring contextual judgment. The model neither traverses the guideline nor emits a free-form conclusion; it evaluates the semantic predicates the current case requires, and those evaluations enter an explicit typed belief state. A deterministic runtime executes the formalized structure over that state, determining which branches hold, which exceptions apply, what remains unresolved, and whether a recommendation is licensed. **In short, the model interprets evidence; the runtime controls what follows.** Because state is explicit and execution deterministic, trajectories can be verified by replay, and unresolved state lets the runtime withhold commitment until the evidence establishes a decision. The specification thus serves not merely as knowledge, but as an operational substrate for control, verification, and decision readiness. The distinction between executable and semantic predicates determines where runtime authority applies; it is a design principle rather than an assumption that the guideline can be reduced to code.

We evaluate GEVHarness across clinical benchmarks spanning workflow control, clinical reasoning, and evidence control, and organize the results as three findings. **First, authority outperforms access.** GEVHarness improves over backbone-only inference on every benchmark for both backbones and is the best among the matched interface conditions on all eight, outperforming both text

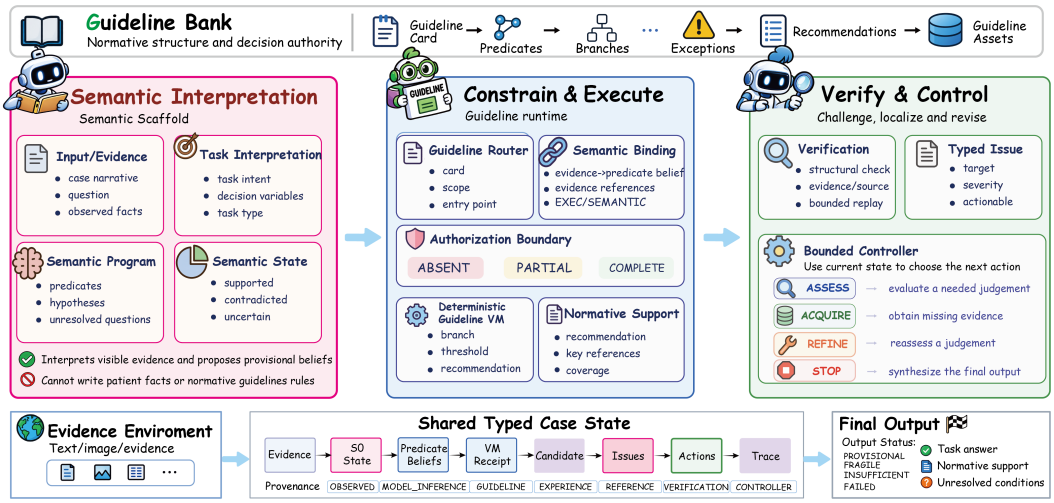


Figure 2: GEVHarness separates case interpretation (S0), scoped guideline execution (G), and verification with bounded control (V). A frozen experience bank supplies procedural hints. The shared state records evidence, revisable beliefs, execution receipts, issues, and actions; the task answer is distinct from its normative support.

access and model-controlled navigation over the same compiled representation. **Second, authority must be selective.** Authorizing machine-checkable structure raises the mechanism-panel score from 23.83% to 46.63%, semantic hardening is non-monotonic, and corrupting the allocation drops the score to 23.32%. **Third, the gains are robust across evidence-change subsets and backbones.** GEVHarness improves the three MedPIC evidence-change subsets by 22.53–50.00 points and transfers across four Qwen3.5 backbones with gains of 3.4–8.4 points.

Our contributions are threefold. **(1)** We formulate clinical reasoning with guidelines as inference under a *partially formalized specification*, separating executable from semantic predicates and making the allocation of runtime authority an explicit design variable. **(2)** We introduce **GEVHarness**, a specification-grounded harness that operationalizes this separation through typed state, deterministic execution, and replay-based verification while preserving non-executable guidance as model-mediated judgment. **(3)** We provide a controlled evaluation across workflow, open-reasoning, and evidence-control regimes, with matched baselines, evidence-change analyses, authority interventions, transfer studies, and an authority ablation that separates receipt visibility from normative control; claims are scoped to the frozen benchmark protocols.

2 The Guideline Harness: GEVHarness

GEVHarness separates what the visible case suggests, what a source-grounded rule licenses, and what the task asks the system to answer. A semantic scaffold interprets the case; a guideline program constrains covered decisions; verification and a bounded controller decide whether to assess, acquire evidence, refine, or stop (Figure 2). All three operate over an explicit state in which observed facts, model inferences, guideline support, and unresolved conditions keep separate provenance.

2.1 Source-Grounded Release and Selective Authority

An offline, versioned release $\mathcal{K}$ contains source-linked guideline units, entry points, and reference material. The executable part of a selected guideline is represented as

$$G = (P, D, C, X, R, Q), \tag{1}$$

where P is a set of typed predicates, D derived conditions, C guarded branches, X exceptions, R recommendations, and Q entry points identifying decision-relevant dependencies. Source spans, applicability, and an authority profile π are attached to the units; the router also returns relevant

non-executable references $\mathcal{R}$. Reference material can inform interpretation without becoming an executable rule, and formal criteria constrain only the propositions they cover.

Within an executable unit, we partition $P = P_{\text{exec}} \uplus P_{\text{sem}}$. An *executable* predicate is computed deterministically once its typed inputs are bound (extracting those inputs may still require model interpretation); a *semantic* predicate requires contextual judgment, such as whether symptoms are severe. Thresholds, Boolean combinations, and exceptions stay explicit rather than hidden inside an opaque judgment, and making one predicate executable does not authorize its source to decide unrelated options.

In our protocol, a strong general-purpose model proposes two source-linked artifacts: a typed candidate IR for executable structure and an Experience Bank for non-executable procedural guidance. Compiled rules are checked against the schema and VM tests, and the release process includes review and validation steps documented in the supplementary artifact (Appendix D). These checks test the compiled program's structure and behavior; they do not establish that every source statement was compiled faithfully. The release and any experience bank are frozen for a reported run, and source-to-IR fidelity is reported separately from runtime performance.

2.2 Case-Local State, Binding, and Guideline Execution

For every case x, including one with no applicable executable guideline, the model constructs a case-local scaffold $S_0(x)$. It records the task type, visible evidence, provisional clinical hypotheses and relations, option requirements when options exist, and open questions. These are auditable, revisable objects; a hypothesis is neither an observed fact nor a new rule. A router then selects applicable release units and entry points from the visible case and source scope; it may select no executable unit.

For a selected G, semantic binding maps evidence to predicate beliefs. Each belief $\beta_t(p)$ carries a typed value when available, a status in $\{\text{SUPPORTED}, \text{CONTRADICTED}, \text{UNCERTAIN}, \text{UNASSESSED}\}$, evidence references, and revision provenance. Values failing type, range, or evidence-reference checks are not admitted; missing evidence stays unknown rather than false; model judgments may update semantic beliefs but never observed facts, source rules, or a VM receipt. The executor is deterministic given a frozen program and validated beliefs:

$$J_t = \text{execute}(G, \beta_t). \tag{2}$$

Under three-valued logic, a branch is active, inactive, or unresolved; a recommendation can likewise be active, blocked, or conditional. An unresolved blocking exception cannot be treated as absent. Execution also exposes the live information needs

$$N(G, \beta_t) = \{p \in \text{req}(G, \beta_t) : \beta_t(p) \text{ is uncertain or unassessed}\}, \tag{3}$$

where req excludes dependencies rendered irrelevant by established branch conditions. The VM reads validated beliefs, not raw case prose, and makes no model call.

GEVHarness records a task answer y_t separately from the normative support n_t licensed by J_t. Coverage is ABSENT when no applicable executable unit contributes, PARTIAL when only some decision-relevant propositions are covered, and DECISION-COMPLETE only when the applicable rules resolve all propositions needed for the task decision. An active, applicable, target-matched constraint overrides a conflicting case-local inference, while a conditional or unmatched recommendation cannot settle an uncovered alternative; under absent or partial coverage, S_0 may still support a provisional answer with unresolved conditions, so producing an answer does not imply the guideline licensed it.

2.3 Experience, Verification, and Bounded Control

A frozen Experience Bank M supplies cross-case procedural hints for case interpretation, routing, checking, and reassessment. Its version and validation status are recorded for each run. Hints guide what to assess, reconsider, or verify, but cannot establish a patient fact, change or activate a rule, or supply an answer; bank creation and validation is an offline step, never an unobserved update during inference.

Verification V combines structural and evidence-aware checks. Structural replay reconstructs β_t and re-executes G to check whether the recorded branches, exceptions, and normative support agree with the VM. Evidence-aware checks may challenge a case interpretation, source applicability, predicate binding, option mapping, or proposed answer. Findings are typed and localized, with provenance and possible decision impact; they never silently overwrite the state they challenge—the controller decides whether reassessment is warranted.

At each step, the bounded controller sees the scaffold, evidence, beliefs, execution receipt, draft answer, verification findings, action history, and remaining budget. It can ASSESS a needed judgment from already visible evidence, ACQUIRE new evidence only if an interactive environment can return it, REFINE a targeted belief or answer mapping, or STOP. Rereading existing evidence is not a new observation, and static question-answering cases provide no acquisition transition; if no useful action remains, the system stops with unresolved conditions recorded. The recorded output includes the task answer, available normative support, status, and trace. The full control loop is given as Algorithm 1 in Appendix D.

Proposition 1 (Scoped structural conformance)**.** *If GEVHarness reports a recommendation as active* normative support *at termination, then under the final validated belief state an applicable branch containing that recommendation has a true guard, and every exception that blocks it evaluates to false. A reference-only unit or a case-local model relation cannot create that license.*

The guarantee concerns execution of the compiled program over the recorded state; source-to-IR fidelity and the clinical correctness of semantic beliefs require separate evaluation. The proof is in Appendix C.

3 EXPERIMENTS

We organize the results as three findings: runtime authority outperforms matched access-based interfaces (§3.3); the benefit depends on *where* authority is placed, not how much is formalized (§3.4); and the gains extend to evidence-change behavior and remain stable across backbones.

3.1 EXPERIMENTAL SETUP

We evaluate GEVHarness on eight benchmarks spanning three regimes: *Clinical Workflow Control* (MedPIC (Hou et al., 2026), MedGUIDE-P256 (Li et al., 2025), TriageBench (Wong, 2026): guideline-grounded rule activation, pathway selection, and next-action prediction), *Open Clinical Reasoning* (MedBullets (Chen et al., 2025), DiagnosisArena (Zhu et al., 2026): broader diagnostic reasoning), and *Evidence Control* (ClinCtrl-Acquire, -SafeStop, -Revise: inference under incomplete, acquired, and updated evidence). Dataset statistics, task definitions, and metric implementations are in Appendix E.

3.2 COMPARED INFERENCE INTERFACES

We compare four matched interfaces on the same frozen cases: direct LLM-only inference, Text-RAG over the original guideline text, model-guided navigation over the frozen IR, and GEVHarness with deterministic runtime authority over its machine-checkable structure. Codex and Claude Code are retained as non-matched agent-shell references. For the matched comparison, we hold the backbone, case evidence, guideline release, compiled representation, and task protocol fixed; the detailed access rules, budgets, and agent-shell configurations are specified in Appendix E.3.

3.3 SPECIFICATION AUTHORITY OUTPERFORMS SPECIFICATION ACCESS

Table 1 tests our central claim: placing the machine-checkable part of a specification under runtime authority should improve inference over supplying the same specification as context. **Across both backbone families, GEVHarness is the best among the matched interface conditions on all eight benchmarks, and the only interface that never falls below its backbone on any of them.** The gains are largest where the specification can constrain a decision transition: GEVHarness improves over direct inference by 20.77 points on MedPIC and 39.06 points on ClinCtrl-Acquire for

Table 1: **Main performance across eight benchmarks spanning clinical workflow control, clinical reasoning, and evidence control.** Backbone-only rows denote direct inference; "+" rows use the same backbone with the indicated interface (§3.2). Metrics are benchmark-specific percentage points and are not aggregated into a macro score. **Bold** marks the best result within each backbone block; Claude Code and Codex are non-matched agent-shell references, so the matched interface comparison is GEVHarness vs. Text-RAG and IR Navigation.

	Clinical Workflow Control			Open Clinical Reasoning		Evidence Control		
Method	**MedPIC** Acc. ↑	**MedGUIDE-P256** Acc. ↑	**TriageBench** Action Acc. ↑	**MedBullets** Acc. ↑	**DiagnosisArena** Dx Acc. ↑	**ClinCtrl Acquire** Success ↑	**ClinCtrl SafeStop** BAcc. ↑	**ClinCtrl Revise** Dx Acc. ↑
Reference models								
GPT-6-Sol	75.16	67.19	37.00	83.06	49.09	25.00	65.86	96.67
Gemini-3.5-flash-lite	62.96	57.81	32.00	72.58	11.48	23.44	59.41	95.83
DeepSeek-Flash-v4.1	45.82	57.81	38.00	65.32	25.90	28.12	61.43	90.00
+ Text-RAG	55.67	59.38	36.00	68.55	24.37	20.31	65.86	86.67
+ IR Navigation	74.33	58.59	37.00	69.35	25.79	25.00	64.52	89.17
+ Claude Code	60.81	57.81	38.00	72.58	38.36	67.19	59.41	90.83
+ Codex	60.81	**61.21**	39.00	**74.12**	43.06	68.75	**68.28**	90.31
+ Ours	**75.80**	60.94	**41.00**	73.39	**43.31**	**70.31**	66.32	**91.67**
Qwen3.8-Flash	62.96	64.06	45.00	78.23	19.45	35.94	62.63	94.17
+ Text-RAG	66.38	63.67	42.00	78.29	19.45	23.44	63.98	93.33
+ IR Navigation	81.58	53.12	37.00	79.84	20.55	29.69	60.22	92.50
+ Claude Code	64.67	61.72	43.00	76.61	**42.84**	71.88	66.13	91.67
+ Codex	61.24	58.20	32.00	76.61	36.83	60.94	47.31	92.10
+ Ours	**83.73**	**64.84**	**48.00**	**82.26**	41.94	**75.00**	**67.19**	**95.83**

Qwen3.8-Flash (Qwen Team, 2026), and by 29.98 and 42.19 points on the same tasks for DeepSeek-Flash-v4.1 (DeepSeek-AI, 2026).

Access to the specification is not the mechanism; authority over it is. The two matched controls fail in complementary ways. Text-RAG helps some workflow tasks but degrades the evidence-control endpoints (ClinCtrl-Acquire falls to 20.31/23.44 from 28.12/35.94 under direct inference for DeepSeek/Qwen): retrieved passages supply knowledge but cannot manage an evolving decision state. ClinCtrl-Acquire makes this distinction explicit: the harness can actively request hidden, case-bound resources, whereas passive interfaces only answer from what is already visible. IR Navigation exposes the very representation that GEVHarness executes, and shows how unstable model-controlled traversal is: for Qwen3.8-Flash it gains 18.62 points on MedPIC yet loses 10.94 on MedGUIDE-P256 and 8.00 on TriageBench with the same IR—the model may or may not follow the structure it is given. GEVHarness's advantage appears exactly when the structured part of the specification can reject an unsupported transition, record an unresolved requirement, or license a next action: specification access is useful, but specification authority changes the inference loop. Because GEVHarness and IR Navigation differ in more than the gate, we do not treat this comparison alone as isolating the mechanism; §3.4 intervenes on authority directly, holding the rest of the harness fixed.

Runtime authority narrows the capability gap. Against the frontier reference model, Qwen3.8-Flash with GEVHarness exceeds GPT-6-Sol on four of the eight benchmarks—including 83.73 vs. 75.16 on MedPIC, 48.00 vs. 37.00 on TriageBench, and 75.00 vs. 25.00 on ClinCtrl-Acquire—and is within one point on MedBullets and ClinCtrl-Revise. The remaining gap concentrates on DiagnosisArena, where open diagnosis leaves the specification least applicable. A harness thus buys a mid-scale backbone frontier-level behavior on the guideline-governed portion of the task.

On Open Clinical Reasoning, where the specification is only partially applicable, GEVHarness remains the strongest matched interface for both backbones (Qwen3.8-Flash: 78.23→82.26 on MedBullets, 19.45→41.94 on DiagnosisArena), so the harness does not require the task to be fully executable: it uses formalized structure where available and keeps the remaining judgment model-mediated. The generic agent shells are non-matched references rather than controls; they are competitive where free-form agency helps (Codex is higher in three DeepSeek settings, Claude Code in one Qwen setting) but can degrade their backbone elsewhere (Codex loses 13.00 points on TriageBench for Qwen3.8-Flash)—the instability the harness removes.

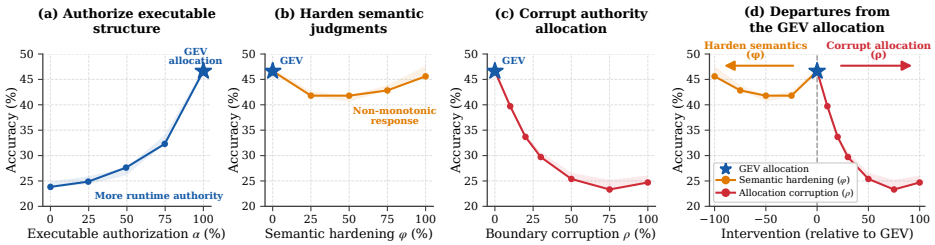


Figure 3: Runtime authority across amount and placement. (a) Authorization of executable structure; (b) semantic hardening; (c) boundary corruption; (d) composite visualization.

3.4 Authority Must Be Selective: Placement over Amount

We next intervene on the *allocation* of authority to distinguish two explanations of the main-table gains: performance could improve because the system receives more formal material, or because the right parts of that material are allowed to govern inference. If the benefit came from the amount of formalization alone, all three interventions below should move performance in the same direction; they do not. Figure 3 reports the intervention families on the frozen mechanism panel.

Authorizing machine-checkable structure consistently helps. Varying α, the realized fraction of eligible executable structure under runtime authority, raises the mechanism-panel score from 23.8342% at $\alpha = 0$ to 46.6321% at full authorization, with intermediate settings ordered by α. Two features of this curve matter. First, the underlying specification is held fixed: at $\alpha = 0$ the model sees the same compiled structure as information, so the curve isolates authority rather than access. Second, the trend is monotone: no intermediate regime rewards withholding executable authority.

Semantic hardening does not follow the same curve. Promoting semantic judgments to authoritative constraints produces a non-monotonic response: 46.63%, 41.80%, 41.80%, 42.83%, and 45.60% at $\phi \in \{0, 25, 50, 75, 100\}\%$; partial hardening is clearly harmful, while full hardening remains at or below the typed allocation. A natural mechanism explains this shape. A partially hardened judgment binds an uncertain contextual assessment too early, and because execution is deterministic, the committed value propagates through branches and exceptions without the revision that a semantic belief would receive. Full hardening restores internal consistency and recovers part of the loss, but never improves on the typed partition. Selective authority is therefore a property of *placement*, not a proxy for the amount of formalization: more formal control is not better control unless it lands on predicates whose truth conditions are actually checkable.

Table 2: **Component ablations.**

Variant	Workflow Control	Open Clinical Reasoning	Evidence Control
Ours	59.24	58.98	76.10
w/o G	53.57	57.40	74.90
w/o E	58.91	58.48	75.65
w/o V	56.57	57.48	74.48

Misplaced authority underperforms on the selected panel. Corrupting the assignment across the executable–semantic partition degrades the score from 46.63% at the typed allocation to 23.32% at $\rho = 75\%$, with a small rebound to 24.7% at $\rho = 100\%$. The corruption intervention keeps the overall control budget comparable while moving authority to the wrong predicates, and at $\rho = 75\%$ the corrupted harness has a lower point estimate than the $\alpha = 0$ system with no executable authority at all (23.83%). On this selected panel, the misplaced allocation has lower point estimates than the typed allocation and suggests that authority assigned across the executable–semantic boundary can be counterproductive.

Together, the interventions explain the main-table pattern: IR Navigation underperforms because access without authority leaves transitions unconstrained, and full formalization is no remedy because authority across the semantic boundary degrades. The operating rule is the design principle itself: enforce structure where its truth conditions are checkable; keep contextual judgments revisable until the evidence suffices.

Component ablations The ablation panel asks whether the three runtime roles are interchangeable. On the frozen mechanism panel, the full harness is highest in all three task regimes (59.24, 58.98, and 76.10). Removing the guideline program G causes the largest Workflow Control drop (59.24→53.57), while removing verification and bounded control V hurts more than removing the experience bank E in every regime (56.57 versus 58.91; 57.48 versus 58.48; and 74.48 versus 75.65). These comparisons are diagnostic rather than causal estimates—removing a component also changes the controller's available state—but they corroborate the main story: execution provides authority, verification protects the resulting state, and the experience bank supplies procedural support.

Executed specification results require runtime authority. We further examine whether the benefit of GEV comes from exposing structured execution results alone or from allowing those results to constrain model decisions. We evaluate four variants on a matched pool of MedPIC, MedGUIDE-P256, TriageBench, and MedBullets cases, covering medication rules, multi-step guideline pathways, triage decisions, and broader clinical reasoning.

Table 3: **Authority ablation and iterative control.** Mean acc, token usage, and calls per case.

Method	Mean accuracy	Tokens/case	Calls/case
Core(G+V)	73.07%	65,775	7.85
Core(G+V) (1 step)	61.88%	9,303	1.00
ReadOnly	60.08%	9,105	1.00
NoVM	61.46%	8,120	1.00

The full multi-step Core(G+V) configuration achieves 73.07% mean accuracy, substantially higher than the one-step Core(G+V) variant (61.88%), while requiring additional inference steps (7.85 vs. 1.00 calls per case). This comparison indicates that iterative execution and verification can improve decision quality when the harness is allowed to repeatedly update and check the intermediate state.

We next isolate the role of runtime authority. ReadOnly exposes the executed VM receipt to the model without allowing it to constrain the final decision, whereas NoVM removes the VM execution component entirely. ReadOnly does not improve over NoVM (60.08% vs. 61.46%; $\Delta = -1.37$ percentage points, 95% CI $[-2.85, 0.00]$), suggesting that exposing structured execution output alone is insufficient. In contrast, allowing the executed result to participate in the final decision improves performance over ReadOnly ($\Delta = +1.80$ percentage points, 95% CI $[+0.42, +3.27]$).

Together, these results support a narrower mechanism claim: the value of a specification-grounded harness does not arise merely from providing additional information or structured artifacts. Rather, gains emerge when executed specification results are granted appropriate runtime authority to constrain model inference.

3.5 Evidence-Conditioned Control and Backbone Transfer

Endpoint accuracy does not reveal whether the harness responds correctly when the evidence changes. We use these counterfactual and risk-changing subsets as an evidence-change behavior test, not as a clinical safety validation. Figure 4(a,b) isolates MedPIC cases whose correct answer must change with the evidence: GEVHarness improves over direct inference by 22.53–47.06 points across the counterfactual, risk-activation, and risk-deactivation subsets for Qwen3.8-Flash, and by 30.99–50.00 points for DeepSeek-Flash-v4.1, with the largest gains where a case departs from a familiar default. The evidence-control endpoints in Table 1 extend this to multi-state decisions: the runtime can defer a decision, request evidence, and incorporate a later release while preserving task-local state. GEVHarness thus changes how inference proceeds between evidence states, not only which answer is finally emitted; per-case transition statistics accompany the released evaluation traces.

GEVHarness transfers across model capacity. A harness claim should survive changes in the backbone. With the specification, evidence access, and protocol held fixed, GEVHarness improves all four Qwen3.5 backbones—by 3.4, 8.0, 8.4, and 5.3 points for 4B, 9B, 27B, and 35B-A3B, respectively (Figure 4(c)). The gains are not monotone in model size, so we read this as transfer across backbones within the evaluated family rather than a scaling law.

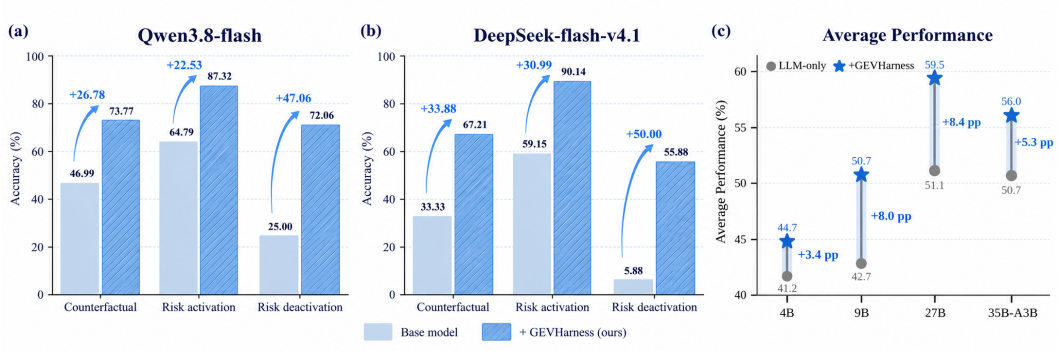


Figure 4: Evidence-conditioned control and backbone transfer. (a,b) Backbone-only vs. GEVHarness accuracy on MedPIC counterfactual, risk-activation, and risk-deactivation subsets for Qwen3.8-Flash and DeepSeek-Flash-v4.1; arrows show paired improvements. (c) Average performance across Qwen3.5 backbones, computed as the direct mean of the three task regimes.

4 Discussion

Specifications need not be complete oracles. Clinical guidelines contain enough machine-checkable structure to constrain traversal, gate decisions, and support replay-based verification, even though part of their meaning remains contextual. The authority interventions expose the underlying tension between *semantic flexibility* and *executable reliability*: context-sensitive concepts need interpretive flexibility, while thresholds, exceptions, and decision structure benefit from deterministic enforcement. Neither unrestricted model control nor indiscriminate hardening captures both; selective runtime authority navigates the tension without resolving it. The price is deliberate: an unlicensed recommendation is never emitted through the authoritative path, and an unresolved requirement may leave the system NOT-READY despite high model confidence.

For clinical reasoning, this division of labor suggests a broader role for a harness beyond maximizing endpoint accuracy. The base model can supply flexible interpretation and multimodal evidence understanding, while the harness can decide which judgments need to be revisited, which transitions are licensed, and what support remains unresolved. A useful future harness should therefore make the intermediate decision state, evidence provenance, authority boundaries, and unresolved requirements inspectable alongside the final answer. Our results provide an initial design signal: selective authority improves several decision transitions, active evidence acquisition changes what the system can know, and replayable state makes these behaviors auditable. These observations do not establish clinical safety, but they suggest that progress in clinical reasoning may be measured by a combination of performance, traceability, and calibrated refusal to overclaim.

Limitations, including compilation fidelity and the scope of the verification guarantee, are discussed in Appendix A.

5 Conclusion

We studied how partially formalized specifications can support reliable clinical reasoning. Using clinical guidelines as the substrate, GEV turns machine-checkable structure into deterministic runtime control while keeping context-dependent judgments model-mediated, improving endpoint performance, evidence-responsive behavior, and backbone transfer. The broader problem is how models, specifications, and runtimes should share authority when formalization is necessarily incomplete: a specification need not be a complete execution oracle to govern inference, and reliable reasoning depends on knowing when to exercise authority and when to preserve judgment.

AI USE STATEMENT

Generative AI tools were used for language editing, manuscript organization, and assistance with implementation-related drafting. The authors designed the method, constructed the evaluation protocol, verified the reported results, and take responsibility for the final manuscript and all claims.

ETHICS STATEMENT

This work studies inference-time control on public benchmark-style data and is not a clinical validation study or a medical device. It does not use private patient data. The system may emit provisional or unresolved states, but the runtime conformance analysis does not establish clinical correctness. Any deployment in clinical settings would require independent clinical, regulatory, and safety evaluation.

REPRODUCIBILITY STATEMENT

The paper specifies the typed guideline representation, belief-state contract, runtime actions, authority interventions, benchmark panels, and evaluation protocol. The accompanying artifact will include the frozen guideline release, experience-bank version, prompts, exact model identifiers and decoding settings, per-case traces, and evaluation scripts required to reproduce and audit the reported results, including the point estimates and bootstrap intervals in Figure 3.

REFERENCES

Aziz A Boxwala, Mor Peleg, Samson Tu, Omolola Ogunyemi, Qing T Zeng, Dongwen Wang, Vimla L Patel, Robert A Greenes, and Edward H Shortliffe. Glif3: a representation format for sharable computer-interpretable clinical practice guidelines. *Journal of biomedical informatics*, 37(3):147–161, 2004.

Pascal S Brandt, Richard C Kiefer, Jennifer A Pacheco, Prakash Adekkanattu, Evan T Sholle, Faraz S Ahmad, Jie Xu, Zhenxing Xu, Jessica S Ancker, Fei Wang, et al. Toward cross-platform electronic health record-driven phenotyping using clinical quality language. *Learning Health Systems*, 4(4):e10233, 2020.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3563–3599, 2025.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026.

Ruiqi Deng, Geoffrey Martin, Tony Wang, Gongbo Zhang, Yi Liu, Chunhua Weng, Yanshan Wang, Justin F Rousseau, and Yifan Peng. Cpgprompt: Translating clinical guidelines into llm-executable decision support. *arXiv preprint arXiv:2601.03475*, 2026.

Chengze Ge, Ruiqing Zhang, Yining Wang, Shengping Liu, Liang Jiaen, Weihua Chen, et al. Guidetree: Guideline-induced review trees for long medical records. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pp. 1589–1604, 2026.

Alon Gorenshtein, Mahmud Omar, Yiftach Barash, Girish N Nadkarni, and Eyal Klang. Neuro-symbolic llm integration in clinical medicine: a systematic review and taxonomy. *Research Square*, pp. rs–3, 2026.

Zhitian Hou, Yuhang Liu, Pengkai Wang, Zeyu Liu, Guanghao Zhu, Zheng Liu, Shuo Cai, Congkai Xie, Zhijie Sang, Kun Zeng, et al. Evaluating counterfactual sensitivity to patient information in medication-safety reasoning. *arXiv preprint arXiv:2608.03028*, 2026.

George Hripcsak. Writing arden syntax medical logic modules. *Computers in biology and medicine*, 24(5):331–363, 1994.

George Hripcsak, Paul D Clayton, T Allan Pryor, Peter Haug, Ove B Wigertz, and Johan Van der Lei. The arden syntax for medical logic modules. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 200, 1990.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Lei Li, Xiao Zhou, Yingying Zhang, and Xian Wu. From retrieval to generation: Unifying external and parametric knowledge for medical question answering. In *Proceedings of the ACM Web Conference 2026*, pp. 2419–2430, 2026.

Xiaomin Li, Mingye Gao, Yuexing Hao, Taoran Li, Guangya Wan, Zihan Wang, and Yijun Wang. Medguide: benchmarking clinical decision-making in large language models. *arXiv preprint arXiv:2505.11613*, 2025.

Qwen Team. Qwen3.8-Flash-Next: A new architecture, towards ultimate cost-efficiency, August 2026. URL `https://qwen.ai/blog?id=qwen3.8-flash-next`.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Yuhao Shen, Lang Cao, Simo Du, Yuqing Wang, Juexiao Zhou, Hao Peng, and Yue Guo. Medguidex: Internalizing decision logic from executable guidelines into large language models for clinical reasoning. *arXiv preprint arXiv:2605.26567*, 2026.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. Rationale-guided retrieval augmented generation for medical question answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12739–12753, 2025.

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

Qi Han Wong. Implicit geographic inference in llm medical triage: Language-driven disparities in emergency recommendations. *arXiv preprint arXiv:2606.01204*, 2026.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Tianshi Xu, Huifeng Wen, and Meng Li. Adapting the interface, not the model: Runtime harness adaptation for deterministic llm agents. *arXiv preprint arXiv:2605.22166*, 2026.

Chenyang Yang, Xinran Zhao, Tongshuang Wu, and Christian Kästner. Better harnesses, smaller models: Building 90% cheaper agents via automated harness adaptation. *arXiv preprint arXiv:2607.08938*, 2026.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Daniel Shao, Yong Hoe Koo, Ko Minhyeok, Qingyu Chen, Mark Gerstein, et al. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16565–16582, 2025.

Yichi Zhang, Nabeel Seedat, Yinpeng Dong, Peng Cui, Jun Zhu, and Mihaela van der Schaar. Glean: Guideline-grounded evidence accumulation for high-stakes agent verification. In *Forty-third International Conference on Machine Learning*.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pp. 4442–4457, 2025.

Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. Diagnosisarena: benchmarking diagnostic reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 3074–3098, 2026.

A Limitations

First, GEVHarness operates on compiled guideline representations. Although the compilation pipeline includes schema validation and structural checks, the resulting representation may not capture every aspect of the original source. The runtime guarantees conformance to the compiled specification and recorded state, rather than correctness of the underlying guideline interpretation or clinical decisions. Second, the executable–semantic partition is an operational abstraction that enables selective runtime authority. The boundary may require domain-specific judgment and should not be interpreted as a universal decomposition of clinical concepts. Third, replay-based verification improves traceability by checking consistency between explicit state and deterministic execution. However, it does not replace independent evaluation of evidence interpretation or semantic judgments made by the model. Fourth, our evaluation uses benchmark-style tasks and does not establish clinical safety, effectiveness, or readiness for deployment. Real-world use would require additional validation under appropriate clinical, regulatory, and safety frameworks. Finally, backbone transfer demonstrates robustness across the evaluated model family, but does not imply a general scaling law or performance guarantees for unseen models.

B Related Work

Computable clinical guidelines. Representing clinical knowledge in machine-interpretable form is a longstanding goal: the Arden Syntax encodes executable medical logic modules (Hripcsak et al., 1990; Hripcsak, 1994), GLIF3 and PROforma represent decision logic and care processes in computable form (Boxwala et al., 2004; Sutton & Fox, 2003), and CQL standardizes machine-readable decision-support logic (Brandt et al., 2020). This literature establishes that substantial clinical structure can be made explicit enough for deterministic computation. GEVHarness builds on it but studies a different setting: the guideline is compiled into a common typed representation without requiring every clinical predicate to become mechanically executable, so runtime authority is assigned selectively.

LLMs with clinical guidelines. Recent work incorporates guidelines into LLM reasoning in several roles. MedGUIDE benchmarks decision consistency with structured guideline pathways (Li et al., 2025). CPGPrompt translates guidelines into decision trees navigated by an LLM, and reports failures on negation and temporal criteria (Deng et al., 2026). GuideTree compiles guidelines into fixed, auditable review trees for long medical records (Ge et al., 2026). MedGuideX uses executable guideline logic as supervision, generating factual and counterfactual training data (Shen et al., 2026). Guidelines also evaluate intermediate reasoning, via guideline-grounded process rewards in Med-PRM (Yun et al., 2025) and calibrated trajectory verification in GLEAN (Zhang et al.). These systems use guidelines as context, workflows, supervision, or verification. GEVHarness asks the complementary question of *how a partially formalized specification can itself participate in inference control*: deterministic authority for machine-checkable structure, typed semantic judgments for the model.

Inference harnesses. Agent research increasingly treats the infrastructure surrounding a model as a first-class determinant of behavior. SWE-agent showed that the design of the agent–computer

interface can substantially alter the performance of a fixed language model on software-engineering tasks (Yang et al., 2024). Related work further treats the harness itself as an object of adaptation or generation, modifying runtime interfaces and trajectory control while keeping model weights fixed (Xu et al., 2026; Yang et al., 2026). Related neuro-symbolic clinical systems similarly place explicit rules, knowledge structures, or symbolic validation around LLM inference, with designs differing in how much authority the symbolic component receives (Gorenshtein et al., 2026). GEVHarness studies a distinct regime in which the specification is only partially formalized—enough structure to constrain and verify inference, but not enough to act as a complete execution oracle—so the question is not only how to place structure around a model, but *which parts of it should be allowed to govern inference*.

C Proof of Proposition 1

Let β_t be the validated belief state at iteration t and let $J_t = \mathrm{execute}(G, \beta_t)$ be the deterministic execution receipt. The validation step admits only typed values with valid ranges and evidence references; it cannot modify G, its source spans, or the authority profile π. Therefore every branch and exception recorded by J_t is evaluated from the fixed program and the final validated beliefs.

By the definition of $\mathrm{authorizedSupport}$, a recommendation can enter the normative-support channel only if J_t contains an applicable branch whose guard is true and every attached blocking exception is false. An unknown or unresolved blocking exception is not false under the three-valued execution semantics, so it prevents the recommendation from entering that channel. The authority gate reads only this receipt and the source scope; reference-only cards, case-local hypotheses, and the task synthesizer have no write access to normative support. After every accepted state update, the executor is called again before a new support record is emitted. Induction over the loop therefore shows that any recommendation reported as active normative support at termination satisfies the stated conditions. This proves structural conformance to the compiled program. It does not prove source-to-IR fidelity, correctness of the model's semantic judgments, or correctness of the task answer.

D Implementation and Release Details

Algorithm 1 GEVHarness inference with selective guideline authority

Require: visible case x, frozen release $\mathcal{K}$, frozen bank M, budget B covering initialization and one complete evaluation

$e_0 \leftarrow \mathrm{visibleEvidence}(x)$; $S_0 \leftarrow \mathrm{scaffold}(x, e_0; M)$
$(G, \pi, \mathcal{R}) \leftarrow \mathrm{route}(x, S_0, \mathcal{K}; M)$; $\beta_0 \leftarrow \mathrm{bind}(G, S_0, e_0)$; $h_0 \leftarrow \varnothing$
for $t = 0, 1, \ldots$ **do**
  $J_t \leftarrow \mathrm{execute}(G, \beta_t)$ ▷ empty receipt if $G = \varnothing$
  $(n_t, c_t) \leftarrow \mathrm{authorizedSupport}(J_t, \pi, S_t, x)$
  $y_t \leftarrow \mathrm{synthesize}(S_t, e_t, n_t, c_t, \mathcal{R})$
  $f_t \leftarrow \mathrm{verify}(G, \beta_t, J_t, y_t, e_t; M)$
  $a_t \leftarrow \mathrm{control}(e_t, S_t, \beta_t, J_t, y_t, f_t, c_t, N(G, \beta_t), h_t, B; M)$
  if $a_t =$ STOP or B cannot fund another update and evaluation **then**
    break
  end if
  $(e_{t+1}, S_{t+1}, \beta_{t+1}, h_{t+1}) \leftarrow \mathrm{apply}(a_t, e_t, S_t, \beta_t, h_t)$
end for
return $(y_t, n_t, c_t, \mathrm{status}_t, \mathrm{trace})$

The algorithm states the shared control contract, not that every task uses every action. For $G = \varnothing$, we define an empty execution receipt and $N(G, \beta) = \varnothing$; S_0, verification, and task synthesis still operate. All calls and actions debit B; if the initial budget cannot cover the first complete evaluation, the run records a budget failure rather than returning an undefined candidate. Otherwise, the controller returns the last fully evaluated state if the remaining budget cannot complete another update and evaluation. The frozen Full configuration also enabled optional case-local relation expansion; the ablation comparison is reported on the full evaluation dataset.

D.1 Source-Grounded IR Preparation

Each guideline package is tied to a versioned source record containing its title, locator, and source spans. An LLM-assisted compiler, implemented with GPT-5.6-Luna in the reported release, proposes a guideline card and a typed program containing predicates, derived conditions, branches, exceptions, recommendations, and entry points. Candidate packages undergo schema, identifier, dependency, and source-span checks. Cited excerpts must be literal spans of the supplied source, and card entry points must agree with those in the program. VM witnesses and related validation records are retained as part of the release process.

Review records track source applicability, clinical meaning, rule coverage, and executable behavior. The release manifest identifies the exact package versions admitted to a reported run; compilation drafts are not loaded as runtime guideline programs.

We distinguish a source's authority from the evaluation type of a predicate. A source-grounded threshold may be evaluated deterministically once its inputs are bound, whereas a judgment such as clinical severity remains semantic and revisable. Reference material may support interpretation and verification without acquiring the authority to activate a normative VM rule. This separation allows useful source material to enter the release without treating every statement as an executable recommendation.

D.2 Staged Guideline Disclosure and Routing

The router initially sees compact guideline cards rather than the full program library. A card identifies the guideline, its version, target population, applicability cues, and available entry points; routing metadata may also include source titles and predicate labels. Libraries containing at most 32 cards expose the full card catalog. For larger libraries, a deterministic filter uses visible case and task text together with card metadata to present at most 12 candidates. Neither the candidate ranking nor the card itself establishes a patient fact or activates a rule.

The router returns guideline and entry-point identifiers. The runtime validates these identifiers against the frozen registry before loading the selected program and its source records. Subsequent semantic binding and execution operate on this loaded program. An unmatched route leaves the case-local semantic scaffold active, with no guideline recommendation invented to fill the gap.

D.3 Binding, Execution, and Partial Coverage

For each selected predicate, binding records a typed value when available, an assessment status, evidence references, and the provenance of any revision. Invalid values or references are rejected before VM execution; missing evidence is represented as unknown. Given the frozen program and validated beliefs, branch evaluation is deterministic. If a branch requires two propositions and one remains unassessed, it cannot yield unconditional normative support. Its unresolved dependency instead becomes an information need that the controller may assess from existing evidence or acquire from an interactive environment.

Execution authority is scoped to the propositions covered by the loaded source. Under partial coverage, an applicable rule can constrain its covered proposition without deciding unrelated alternatives. The system records the task answer and its normative support separately: a provisional answer may be produced from semantic state even when the guideline does not resolve the complete task decision. The runtime maintains separate provenance for observed evidence, model inference, guideline support, experience hints, verification findings, and controller actions.

D.4 Experience, Verification, and Control

The Experience Bank is prepared and frozen before a reported run. Each item specifies a procedural hint, retrieval triggers, provenance, and validation status. At inference time, eligible hints are retrieved from currently visible case context for their declared stage, such as routing or verification. The trace records whether a hint was retrieved and consumed. Hints may direct attention to a scope condition or a judgment worth rechecking; they do not supply patient observations, rewrite a guideline program, or determine the final answer.

Verification checks consistency among recorded beliefs, the VM receipt, and the proposed decision. When enabled, evidence-aware checks can also localize a questionable interpretation, source application, or answer mapping. A finding is recorded as an issue for the controller rather than silently overwriting the challenged state. The controller selects among assessment of visible evidence, acquisition of genuinely new evidence when an environment permits it, targeted refinement, and stopping. Its externally visible action set is {ASSESS, ACQUIRE, REFINE, STOP}. Every model call and action is charged to the configured budget. Static question-answering tasks have no acquisition transition; a run returns the last fully evaluated state when another complete update and evaluation cannot be funded, or records budget failure if the initial evaluation cannot be completed.

D.5 COMPARATOR ACCESS AND RUN RECORDS

Text-RAG retrieves passages from the original source-text collection and passes them to the model for a direct-style answer. In the BM25 condition, retrieval uses the visible case, question, and options as its query and returns the top three passages. It does not receive compiled predicates or execute the VM. IR Navigation instead accesses the frozen IR library through guideline cards and a read-only `load_guideline` action. The model chooses which entry to load and interprets the resulting structure itself; no external selector or deterministic VM makes its decision.

For each reported condition, we retain the input and source-collection identifiers, release and Experience Bank versions, effective component configuration, prompts, model and decoding settings, call and action budgets, per-case outputs, provider request and token usage, failures, retries, tool activity, and evaluation protocol. Agent-shell runs additionally retain raw events and final outputs. Gold labels and final scoring remain outside the runtime workspace. Deterministic workspace and label-isolation checks remain the primary controls.

D.6 COMPUTE AND SERVING CONFIGURATION

Experiments were conducted on a single Ubuntu 24.04.4 server with two Intel Xeon Platinum 8558 CPUs (96 physical cores, 192 threads), 2 TiB of RAM, and eight NVIDIA H20-3e GPUs (143,771 MiB memory per GPU). For cross-scale experiments, Qwen3.5-4B and Qwen3.5-9B were served with SGLang using two data-parallel replicas each; Qwen3.5-27B used four data-parallel SGLang replicas. These SGLang deployments used a 65,536-token context and FP8 E4M3 KV caches. Qwen3.5-35B-A3B was served using vLLM 0.19.0 with four-way tensor parallelism and expert parallelism, a 262,144-token context, BF16 computation, and an FP8 E4M3 KV cache.

Qwen3.8-Flash-Next-FP8 was evaluated separately through two independent four-GPU SGLang replicas, each with tensor parallelism 4 and expert parallelism 4; evaluation shards were distributed across the replicas. Its serving context length was 262,144 tokens, with a static GPU memory fraction of 0.85 and an 8,192-token chunked prefill size.

D.7 RELEASE SCOPE AND LIMITS

A clean schema or VM test does not prove faithful source-to-IR compilation, and a frozen Experience Bank does not guarantee semantic completeness. In the authority ablation, the receipt-only contrast is not reliably positive, whereas allowing the executed result to constrain the decision yields a positive paired contrast on the matched pool. The MedPIC-Bench release already contains patient-specific counterfactual rule activation and deactivation cases (Hou et al., 2026); our claim concerns the inference-time authority allocation applied to that setting. The three ClinCtrl panels are constructed evaluation panels, so their provenance, case-level files, and parent-case mappings must be included in the artifact for external audit.

The main table reports point estimates on frozen benchmark panels. Exact model identifiers and decoding settings are fixed for each reported comparison.

E EXPERIMENTAL SETUP

E.1 BENCHMARKS AND EVALUATION REGIMES

We evaluate GEVHarness on eight benchmarks spanning three complementary regimes of clinical inference: *workflow control*, *clinical reasoning*, and *evidence control*. These regimes differ in how directly a formalized specification can govern the target decision, allowing us to evaluate both strongly guideline-structured settings and tasks in which formalized structure covers only part of the reasoning process.

Clinical workflow control. **MedPIC** uses the 467-case MedPIC-Bench release (Hou et al., 2026), including 284 guideline-following and 183 counterfactual cases. The task requires predicting the exact set of applicable decisions, testing whether patient-specific evidence correctly activates or deactivates risks, contraindications, and interactions. **MedGUIDE-P256** is a fixed 256-case subset constructed for path coverage from the larger MedGUIDE pool (Li et al., 2025). Cases require decisions involving eligibility, thresholds, exceptions, temporal conditions, and guideline routing. **TriageBench** uses 100 staged triage states from the public release (Wong, 2026), balanced between cases that require further questioning and cases that should terminate with an assessment, and evaluates the correctness of the next action.

Clinical reasoning. **MedBullets** is a fixed mixed pool of 124 multiple-choice questions drawn from the public MedBullets release (Chen et al., 2025), spanning diagnosis, testing, management, and clinical knowledge. **DiagnosisArena** uses all 915 cases in the publicly released test split (Zhu et al., 2026). We evaluate one Top-1 diagnosis per case with a fixed DeepSeek-Flash-v4.1 terminology-equivalence judge.

Evidence control. We additionally evaluate whether the harness can govern inference as evidence becomes available. **ClinCtrl-Acquire** is our frozen 64-case evidence-control panel in which the system initially observes partial patient information and may request additional case-specific resources before producing a final diagnosis. **ClinCtrl-SafeStop** is our frozen panel of 248 blinded evidence states derived from 62 parent cases and evaluates whether additional evidence should be acquired or the current evidence is sufficient to stop. **ClinCtrl-Revise** is our frozen panel of 120 cases in which image evidence is released after an initial text-based assessment, testing whether the system appropriately updates its final diagnosis after new evidence arrives. This track also tests multimodal revision: the system must incorporate a case-linked image released after the initial assessment. These three ClinCtrl panels are constructed from public AgentClinic assets (Schmidgall et al., 2024). Their source-to-task transformations and separate scoring rules are specified in Appendix E.2.

E.2 CLINCTRL: AGENTCLINIC-DERIVED EVIDENCE CONTROL

ClinCtrl adapts public AgentClinic MedQA and NEJM-Extended case assets into three deterministic evidence-replay tracks (Schmidgall et al., 2024). It is not a newly collected patient cohort. The environment releases only case-bound material present in the source record; it does not synthesize patient findings. The three tracks have different tasks and metrics and are never pooled into one accuracy.

Source-to-task transformation. We pin the AgentClinic source revision and convert each eligible record into three separate artifacts: a runtime-visible case, a case-bound evidence environment, and an evaluator-only answer or action label. For MedQA-Extended, demographics, the presenting symptom, and the diagnostic objective form the initial case. Nonempty history, examination, and diagnostic-test fields become typed resources; the model initially sees their IDs, types, and labels but receives their original contents only after a valid request. The diagnosis label remains outside the runtime state. Acquire uses this interactive view; SafeStop reuses its parent cases and changes only which source resources are visible or requestable. For NEJM-Extended, the case text and five source options form the initial view, while the original linked image is either hidden, shown from the start, or released after the first assessment. These transformations preserve the source evidence and make evidence access, rather than newly authored patient facts, the experimental intervention. The frozen manifests record source revision, case IDs, construction rules, and input hashes.

Table 4: Three ClinCtrl tasks derived from public AgentClinic assets. SafeStop counts evidence states, not independent patients. Each track is scored separately.

Track	Source and unit	Evidence transformation	Required decision	Primary metric
Acquire	64 MedQA cases	Hide source history, examination, and tests behind a resource catalog	Request or stop, then name a diagnosis	Final diagnosis accuracy
SafeStop	62 MedQA parents $\times$ 4 states	Remove or expose one source resource; vary whether it is requestable	Acquire or stop in the current state	Binary action balanced accuracy
Revise	120 NEJM cases	Hold back the case-linked image until a defined release point	Reassess and choose one of five diagnoses	Final option accuracy after image access

Acquire. The 64 episodes are selected from schema-valid AgentClinic MedQA cases with at least three hidden resources, using a fixed case ordering. Initially visible fields are demographics, primary symptom, diagnostic objective, and a catalog of resource IDs, types, labels, and costs; resource content remains hidden until requested. A system may request available resources, decide when to stop, and name one final diagnosis. Table 1 reports the fraction of 64 cases with a correct final diagnosis, not the correctness of each resource choice. The current frozen scorer uses normalized exact diagnosis strings; any semantic-equivalence judge defines a separate scorer version. This track has no physician-annotated optimal-resource sequence.

SafeStop. Sixty-two parent cases from Acquire define the SafeStop panel. Each parent yields four blinded states: HARD-GAP (a useful available resource is missing; target ACQUIRE), DISTRACTOR-GAP (a noncritical resource is missing; target STOP), COMPLETE (target STOP), and UNAVAILABLE-GAP (the missing resource cannot be requested; target STOP). The system chooses ACQUIRE or STOP from the current visible state. The runtime sees neither the state type nor parent identifier, expected action, or critical-resource label. These 248 states are repeated observations of 62 cases. The primary metric is binary action balanced accuracy: one half of hard-gap acquisition recall plus one half of stop accuracy across the other three states. It does not score the final diagnosis. Uncertainty estimates cluster by parent case. The Qwen Direct/controller panel is distinct from the archived agent panel and is not pooled with it.

Revise. We use 120 NEJM-Extended cases from a fixed AgentClinic source revision, each with five diagnostic options and one case-linked medical image. Text-only, image-always-visible, and staged-release conditions distinguish initial reasoning from revision after the image arrives. Images are cached with hashes and dimensions for reproducibility but are not redistributed with the manuscript. Because each case has one primary image, this track probes grounding and revision after new modality evidence, not choice among many resources. Table 1 reports final exact-option accuracy descriptively.

E.3 COMPARED INFERENCE INTERFACES

Our primary comparisons vary how the same underlying model interacts with the available clinical specification.

LLM-only. The model answers directly from the available case evidence without access to the formalized guideline representation.

BM25 Text-RAG. We retrieve the top three passages from the original guideline source with BM25 and provide them under the task output protocol. This is a controlled retrieval condition related to retrieval-augmented generation and medical guideline retrieval (Lewis et al., 2020; Zhao et al., 2025).

LLM-guided IR Navigation. The model receives the same frozen guideline representation used by GEVHarness, but retains control over traversal and decision making. It selects and interprets

entries without running the deterministic VM or writing normative support through the authority gate. This is our controlled implementation inspired by CPGPrompt, not a reproduction of it.

GEVHarness (ours). GEVHarness uses the same formalized specification as IR Navigation, but places machine-checkable structure under deterministic runtime control and restricts the model to typed semantic judgments requested by the runtime.

Agent-shell reference interfaces. Codex CLI v0.155.0 and Claude Code v2.1.280 are evaluated as inference-time shells connected to the specified base model for each backbone block. Each unit runs in an isolated workspace with blocked external web access, a fixed output schema, and only the permitted resource catalogue or clinical image. Codex may use shell tools, whereas Claude Code is restricted to Read. Each unit is limited to 12 provider requests, including failed requests, and raw agent events, tool activity, failures, and final outputs are retained. Gold labels and final scoring remain outside the agent workspace. These are reference interfaces rather than matched controls.

For matched interface comparisons, we hold fixed the backbone model, case evidence, guideline source, compiled representation, task-specific evaluation protocol, decoding settings, and inference budget; only the inference interface changes.

E.4 Evaluation Protocol and Metrics

All matched conditions are evaluated on the same frozen cases with identical evidence access and task-specific output constraints. Failed calls, invalid outputs, abstentions where a task requires an answer, and evaluator failures are retained according to the benchmark-specific scoring protocol rather than silently discarded.

We report the native primary metric of each benchmark. MedPIC uses exact-set accuracy; MedGUIDE-P256 and MedBullets-124 use exact option accuracy; TriageBench uses next-action accuracy; DiagnosisArena uses Top-1 terminology-equivalence accuracy under the fixed DeepSeek-Flash-v4.1 judge; ClinCtrl-Acquire reports final diagnostic success after evidence acquisition; ClinCtrl-SafeStop reports balanced accuracy for acquire-versus-stop decisions; and ClinCtrl-Revise reports final diagnosis accuracy after image release.

For DiagnosisArena, all compared methods use the complete publicly released 915-case test split, the same input representation, and the same fixed judge model and prompt. These results correspond to our Top-1 evaluation protocol and are not directly comparable to the benchmark's original ordered Top-5 setting.

For ClinCtrl-SafeStop, the 248 evaluation states are derived from 62 parent cases; uncertainty estimates are therefore computed at the parent-case level rather than treating the derived states as independent patients.

The main table reports point estimates on the frozen panels. Confidence bands are shown for the authority-intervention curves in Figure 3; the SafeStop panel uses parent cases as the resampling unit because its 248 states are derived from 62 parents. We do not aggregate heterogeneous native metrics into a single headline score.

E.5 Mechanism, Transfer, and Ablation Studies

Beyond endpoint performance, we evaluate whether GEVHarness exhibits the behaviors expected of a specification-grounded inference harness.

Behavior under changing and incomplete evidence. We analyze counterfactual and risk-changing cases in MedPIC and decompose the evidence-control tasks into acquisition, stopping, and revision behavior. These analyses test whether the harness responds appropriately when clinically relevant evidence is added, removed, or remains unresolved, rather than evaluating endpoint accuracy alone. For ClinCtrl-Acquire, we additionally track changes from the initial to final diagnosis and resource-use behavior; for ClinCtrl-SafeStop, we separate acquisition of critical missing evidence from correct stopping; and for ClinCtrl-Revise, we examine within-case changes before and after image evidence is released.

Table 5: Updated authority ablation on the matched-state pool.

Contrast	Interpretation	Paired difference
ReadOnly − NoVM	Receipt without normative authority	-1.37 pp 95% CI $[-2.85, 0.00]$
Full − ReadOnly	Normative authority after receipt	$+1.80$ pp 95% CI $[+0.42, +3.27]$

Selective execution. To test whether GEVHarness benefits from selective rather than maximal runtime authority, we intervene on how formal control is allocated across the typed specification. We study three complementary perturbations. First, we vary the amount of machine-checkable structure granted execution authority. Second, beginning from the typed GEVHarness allocation, we progressively promote semantic predicates to hard constraints. Third, at a fixed authority budget, we perturb which predicates receive control by swapping assignments across the executable–semantic partition. These interventions test whether performance depends on selective authority allocation rather than simply on the quantity of formal control.

We denote by α the fraction of eligible executable structure granted runtime authority, by ϕ the extent to which semantic predicates are promoted to hard constraints, and by ρ the fraction of a fixed authority budget whose assignment is perturbed across the typed partition. The exact construction of the intervention sets, sampling procedure, seeds, and evaluated values are given in Appendix H.

Model transfer. To test whether the harness benefit depends on a particular backbone, we repeat the matched LLM-only versus GEVHarness comparison across models of different capacities. All other components of the inference protocol are held fixed where supported by the model interface. We report both absolute performance and the within-backbone gain

$$\Delta_{\mathrm{GEVHarness}} = \mathrm{Perf}(\mathrm{GEVHarness}) - \mathrm{Perf}(\mathrm{LLM\text{-}only}).$$

Component ablations. Finally, we ablate individual components of GEVHarness while keeping the remaining inference protocol fixed. These experiments are intended to identify which parts of the harness account for its observed behavior, rather than to redefine the selective-execution principle itself. The exact ablation variants are specified in Section 3.4 and Appendix I.

F Experience Bank Construction Rules

The Experience Bank is a versioned collection of procedural hints rather than a second source of clinical facts. Each hint is linked to an intended use such as case interpretation, route selection, verification, or reassessment. The release validator rejects hints that contain patient-specific assertions, write operations over observed evidence or guideline rules, rule-activation instructions, or final-answer assignments. A hint may recommend an assessment or a check; it cannot decide the result of that assessment. The bank is validated as a procedural hint collection.

G Authority Ablation Protocol

The updated authority ablation uses a matched pool drawn from MedPIC, MedGUIDE-P256, TriageBench, and MedBullets. We selected these datasets because they span patient-specific medication rules, multi-step guideline pathways, triage actions, and mixed clinical reasoning while sharing fixed options and deterministic scoring. All conditions start from the same frozen case state and use the same task protocol. Full executes the typed program and allows its receipt to constrain the decision; ReadOnly executes the same program and exposes the receipt without normative authority; NoVM disables execution and receipt generation. The multi-step G+V reference reaches 73.07% mean accuracy at 65,775 tokens and 7.85 calls per case. Full, ReadOnly, and NoVM use one call per case with 9,303, 9,105, and 8,120 tokens, respectively.

The first contrast does not show a reliable benefit from placing the receipt in context alone. The second shows a positive paired effect when the executed result is allowed to constrain the decision. We do not use Full − NoVM as an additional claim.

H AUTHORITY-INTERVENTION CONSTRUCTION

Dual-opportunity panel. We constructed one mechanism panel from five frozen public-test pools. A case was eligible if its replayable pre-synthesis state contained both at least one source-defined executable predicate and at least one otherwise-soft case-local semantic condition. These two objects are required to implement authorization (α), semantic hardening (ϕ), and a within-case, budget-matched authority swap (ρ) on the *same cases*. We exhaustively included all 193 eligible cases among 1,842 replayable cases: 93 MedPIC, 85 MedGUIDE-Path256, 9 TriageBench, and 6 MedBullets-mixed124 cases.

Frozen-state policy interventions. We evaluated all three axes on these 193 cases using frozen pre-synthesis semantic states from Qwen3.8-Flash-Next. In the authorization axis (α), we progressively granted VM authority to source-defined executable predicates; predicates outside the authorization mask were projected to unknown for VM execution. In the semantic-hardening axis (ϕ), all executable predicates remained authorized while increasingly many otherwise-soft case-local conditions were promoted to hard option guards. In the boundary-corruption axis (ρ), each feasible swap demoted a source-defined executable predicate and promoted an otherwise-soft semantic condition from the same case, keeping the number of hard constraints fixed. The ρ x-axis denotes the fraction of feasible within-case swaps, not the fraction of all executable predicates.

We reran structured answer mapping only for distinct case–execution states and scored unresolved or failed outputs as incorrect. The common execution state at $\alpha = 100\%$, $\phi = 0\%$, and $\rho = 0\%$ matched per case. Because hardening is enforced through both the mapping instruction and a controller, these contrasts estimate the effect of the implemented authority policy rather than an isolated VM-only effect.

Evaluated settings. The plotted scores are 23.83, 24.87, 27.63, 32.30, and 46.63% at $\alpha = 0, 25, 50, 75, 100\%$; 46.632, 41.80, 41.80, 42.83, and 45.60% at $\phi = 0, 25, 50, 75, 100\%$; and 46.63, 39.72, 33.68, 29.71, 25.39, 23.32%, and 24.70% at $\rho = 0, 10, 20, 30, 50, 75, 100\%$. These interventions measure the implemented allocation policies on this selected panel; they do not identify a universal optimal allocation for all guidelines or domains.

Figure 4(c) reports the direct mean across the three task regimes for each Qwen3.5 backbone.

I COMPONENT ABLATIONS

The ablation rows use the same frozen evaluation panel, case evidence, and inference budget. Full$-G$ retains the case-local scaffold S_0 and raw text from the same guideline source, replacing executable IR with that text. It does not execute the VM and disables assessment and verification actions that depend on executable G. This is a *component-family replacement*, not the removal of all guideline information. Full$-E$ neither retrieves nor injects the frozen Experience Bank; all other settings remain as in Full. Full$-V$ neither generates nor consumes dedicated verification reports or revision feedback; schema checks, type checks, and intrinsic VM checks remain active. The full configuration has the highest reported score in Workflow Control (59.24%), Open Clinical Reasoning (58.98%), and Evidence Control (76.10%). Because these interventions alter different state channels and available feedback, the comparisons are diagnostic rather than additive causal effects.

J IR SOURCE DOCUMENTS

The IR source document registry contains 568 entries, including clinical guidelines, prescribing information, and medical reference material. The linked titles and their publishers or hosts are listed below in the source registry order.

2015 ATA thyroid nodule guidelines (American Thyroid Association / Thyroid), 2019 ESC Guidelines for the management of patients with supraventricular tachycardia (European Society of Cardiology / European Heart Journal), 2022 ACC/AHA Guideline for the Diagnosis and Management of Aortic Disease (ACC/AHA / Circulation), 2023 ESC Guidelines for the management of cardiomyopathies (European Society of Cardiology / European Heart Journal), 2024 ACC/AHA lower-extremity PAD guideline slide set (American Heart Association), 2025 American Thyroid Association Management Guidelines for Adult Patients with Differentiated

Thyroid Cancer (American Thyroid Association / Thyroid), AASLD Practice Guidance on prevention, diagnosis, and treatment of hepatocellular carcinoma (AASLD / Hepatology), Acute Intermittent Porphyria (NCBI Bookshelf), Acute Myeloid Leukemia Treatment (PDQ®): Treatment of Recurrent APL (National Cancer Institute), AGA patient guidance: constipation (American Gastroenterological Association), American Geriatrics Society 2023 updated AGS Beers Criteria® for potentially inappropriate medication use in older adults (American Geriatrics Society / Journal of the American Geriatrics Society), American Society of Hematology 2025 guidelines for treating newly diagnosed acute myeloid leukemia in older adults: Educational Slide Set (American Society of Hematology), Aminoguanidine ameliorates overexpression of prosclerotic growth factors and collagen deposition in experimental diabetic nephropathy (Journal of the American Society of Nephrology), An Overview of Uncommon Cutaneous Malignancies, Including Skin Appendageal (Adnexal) Tumours and Sarcomas (Clinical Oncology), Biliary tract cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up (ESMO / Annals of Oncology), British Society of Gastroenterology guidelines for the management of hepatocellular carcinoma in adults (British Society of Gastroenterology), BSH drug-induced immune and secondary autoimmune haemolytic anaemia guideline (British Society for Haematology), BTS clinical statement for the diagnosis and management of ocular tuberculosis (BMJ open respiratory research), Clinical Practice Algorithm for the Follow-Up of Unrepaired and Repaired Patent Ductus Arteriosus (American College of Cardiology), Consensus on acromegaly complications: an update (Pituitary), DailyMed diclofenac sodium/misoprostol prescribing information (DailyMed), DailyMed Leflunomide prescribing information (DailyMed), DailyMed meloxicam prescribing information (DailyMed), DailyMed Methotrexate tablet prescribing information (DailyMed), DailyMed metronidazole tablet prescribing information (DailyMed), DailyMed Multaq (dronedarone) prescribing information (DailyMed), DailyMed ribavirin capsule prescribing information (DailyMed), DailyMed sulfamethoxazole and trimethoprim prescribing information (DailyMed), DailyMed Thalomid (thalidomide) prescribing information (DailyMed), DailyMed Warfarin sodium prescribing information (DailyMed), Diagnosis and management of AML in adults: 2022 ELN recommendations (author-network PDF) (European LeukemiaNet), Diagnosis and management of mitochondrial disease: a consensus statement from the Mitochondrial Medicine Society (Genetics in medicine : official journal of the American College of Medical Genetics), Diagnosis and Management of Patent Ductus Arteriosus (NeoReviews / American Academy of Pediatrics), Donor splice-site mutations in WT1 are responsible for Frasier syndrome (Nature Genetics), Drug-Induced Acute Interstitial Nephritis (Clinical journal of the American Society of Nephrology : CJASN), EANO guideline on the diagnosis and management of meningiomas (Neuro-oncology), EAU Guidelines on Muscle-Invasive Bladder Cancer 2025 (European Association of Urology), EAU Guidelines on Non-muscle-invasive Bladder Cancer 2025 (European Association of Urology), EAU Guidelines on Primary Urethral Carcinoma (archive 2025; limited update April 2024) (European Association of Urology), EAU Guidelines on Renal Cell Carcinoma 2025 (European Association of Urology), EAU Guidelines on Testicular Cancer 2025 (European Association of Urology), EAU Guidelines on Upper Urinary Tract Urothelial Carcinoma 2025 (European Association of Urology), EAU-ASCO Guidelines on Penile Cancer 2025 (European Association of Urology), Endocrine Society: Acromegaly Clinical Practice Guideline (Endocrine Society), ESGO/ESTRO/ESP Guidelines for the management of patients with cervical cancer - Update 2023 (International journal of gynecological cancer : official journal of the International Gynecological Cancer Society), ESMO Clinical Practice Guideline interim update on the use of targeted therapy in acute lymphoblastic leukaemia (ESMO / Annals of Oncology), European 2015 congenital hypogonadotropic hypogonadism consensus (Nature Reviews Endocrinology), European Society of Gynaecological Oncology Guidelines for the Management of Patients with Vulvar Cancer - Update 2023 (International journal of gynecological cancer : official journal of the International Gynecological Cancer Society), Fabry Disease (NCBI Bookshelf), FDA Drug Safety Communication: NSAIDs in pregnancy at 20 weeks or later (FDA), FDA Isotretinoin Capsule Information (FDA), FDA nivolumab indication for resected esophageal or GEJ cancer (FDA), FDA Pregnancy and Lactation Labeling Rule resources (FDA), FDA statin pregnancy Drug Safety Communication (2021) (FDA), FDA: Pediatric Information Incorporated Into Human Prescription Drug and Biological Product Labeling (FDA), FDA: Should You Give Kids Medicine for Coughs and Colds? (FDA), Fluorouracil Cream USP 5% prescribing information (DailyMed), Glycogen Storage Disease Type I (NCBI Bookshelf), GMC remote consultations (General Medical Council), Guideline for the management of myasthenic syndromes (Therapeutic advances in neurological disorders), Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017 (Fleischner Society / Radiology), Guidelines for the treatment of hypothyroidism: prepared by the american thyroid association task force on thyroid hormone replacement (Thyroid : official journal of the American Thyroid Association), Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents With HIV (NIH HIV Clinical Guidelines), Guidelines of care for the management of cutaneous squamous cell carcinoma (Journal of the American Academy of Dermatology), Hodgkin lymphoma: EHA Clinical Practice Guidelines for diagnosis, treatment, and follow-up (HemaSphere), Immunotherapy and Targeted Therapy for Advanced Gastroesophageal Cancer: ASCO Guideline Update (ASCO / Journal of Clinical Oncology), International, evidence-based consensus diagnostic criteria for HHV-8-negative/idiopathic multicentric Castleman disease (Blood), Key Potentially Inappropriate Drugs in Pediatrics: The KIDs List (The journal of pediatric pharmacology and therapeutics : JPPT : the official journal of PPAG), Management of Abdominal Aortic Aneurysms: ESVS clinical practice guideline (European Society for Vascular Surgery), Management of acute promyelocytic leukemia: updated recommendations from an expert panel of the European LeukemiaNet (European LeukemiaNet / Blood), Medical management of Langerhans cell histiocytosis from diagnosis

to treatment (Expert Opinion on Pharmacotherapy), Metastatic Pancreatic Cancer: ASCO Guideline Update (ASCO / Journal of Clinical Oncology), Minimal/measurable residual disease in AML: a consensus document from the European LeukemiaNet MRD Working Party (Blood), Multidisciplinary consensus guideline for the diagnosis and management of spontaneous intracranial hypotension (Journal of Neurology, Neurosurgery & Psychiatry / BMJ), Nanophthalmos: A Review of the Clinical Spectrum and Genetics (Journal of ophthalmology), NCCN Clinical Practice Guidelines in Oncology: Acute Myeloid Leukemia, Version 2.2021 (NCCN (archived PDF mirror)), NCCN Clinical Practice Guidelines in Oncology: Central Nervous System Cancers (NCCN), NCCN Clinical Practice Guidelines in Oncology: Central Nervous System Cancers, Version 1.2023 (NCCN (archived PDF mirror)), NCCN Clinical Practice Guidelines in Oncology: Chronic Myeloid Leukemia (NCCN), NCCN Clinical Practice Guidelines in Oncology: Esophageal and Esophagogastric Junction Cancers (NCCN), NCCN Clinical Practice Guidelines in Oncology: Gastric Cancer (NCCN), NCCN Clinical Practice Guidelines in Oncology: Pancreatic Adenocarcinoma (NCCN), NCCN Clinical Practice Guidelines in Oncology: Thyroid Carcinoma (NCCN), NCI: How Is Breast Cancer Diagnosed? (National Cancer Institute), NCI: Staging for Breast Cancer—The TNM System (National Cancer Institute), Neonatal hypertension: cases, causes, and clinical approach (Pediatric nephrology (Berlin, Germany)), NHS common health problems in pregnancy (NHS), NHS Symptoms A to Z (NHS), NHS: Allergies (NHS), NHS: Balanitis (NHS), NHS: Blepharitis (NHS), NHS: Conjunctivitis (NHS), NHS: Earache (NHS), NHS: Headaches (NHS), NHS: Heartburn and acid reflux (NHS), NHS: Knee pain (NHS), NHS: Meningitis symptoms (NHS), NHS: Mouth ulcers (NHS), NHS: Urinary tract infections (UTIs) (NHS), NHS: Vaginal discharge (NHS), NICE CG84: Diarrhoea and vomiting caused by gastroenteritis in under 5s (NICE), NICE NG12 suspected oesophageal cancer referral (NICE), NICE NG120: Acute cough antimicrobial prescribing (NICE), NICE NG122 lung cancer diagnosis and staging (NICE), NICE NG143: Fever in under 5s (NICE), Ocular findings in posterior microphthalmos (Saudi journal of ophthalmology : official journal of the Saudi Ophthalmological Society), Oral tuberculosis - Current concepts (Journal of family medicine and primary care), Pan-Asian adapted ESMO Clinical Practice Guidelines for the diagnosis, staging and treatment of patients with metastatic breast cancer (ESMO open), Pan-Asian adapted ESMO Clinical Practice Guidelines for the diagnosis, treatment and follow-up of patients with endometrial cancer (ESMO open), Pan-Asian adapted ESMO Clinical Practice Guidelines for the diagnosis, treatment and follow-up of patients with oncogene-addicted metastatic non-small-cell lung cancer (ESMO open), Pediatric Pharmacy Association 2025 KIDs List of Key Potentially Inappropriate Drugs in Pediatrics (The journal of pediatric pharmacology and therapeutics : JPPT : the official journal of PPAG), Pneumocystis Pneumonia in Human Immunodeficiency Virus-infected Adults and Adolescents: Current Concepts and Future Directions (Clinical medicine insights. Circulatory, respiratory and pulmonary medicine), Pneumocystis Pneumonia: Adult and Adolescent OIs (NIH HIV Clinical Guidelines), Postmastectomy Radiation Therapy: An ASTRO-ASCO-SSO Clinical Practice Guideline (ASTRO–ASCO–SSO / Journal of Clinical Oncology), Primary central nervous system lymphomas: EHA-ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up (EHA–ESMO / HemaSphere), Protocols on classification, monitoring and therapy in children's rheumatology (PRO-KIND): results of the working group Polyarticular juvenile idiopathic arthritis (Pediatric rheumatology online journal), Radiotherapy of Breast Cancer—Professional Guideline, Central-Eastern European Professional Consensus Statement (Pathology & Oncology Research), S2k guideline basal cell carcinoma of the skin (update 2023) (Journal der Deutschen Dermatologischen Gesellschaft = Journal of the German Society of Dermatology : JDDG), SEOM Clinical Guideline of management of soft-tissue sarcoma (2020) (Clinical & translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico), SEOM-TTCC clinical guidelines for the treatment of head and neck cancer (2025) (Clinical & translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico), Small-cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up(⋆) (Annals of oncology : official journal of the European Society for Medical Oncology), Society for Vascular Surgery practice guideline: patients with abdominal aortic aneurysm (Society for Vascular Surgery), Systemic Treatment of Patients With Metastatic Breast Cancer: ASCO Resource-Stratified Guideline (JCO global oncology), Systemic Treatment of Thyroid Cancer: ASCO Guideline (ASCO / Journal of Clinical Oncology), Thyroid Health: Molecular testing in thyroid nodules (American Thyroid Association), Tolak (fluorouracil) cream prescribing information (DailyMed), Transient loss of consciousness ('blackouts') in over 16s (NICE CG109) (NICE), Treatment of Metastatic Colorectal Cancer: ASCO Guideline (Journal of clinical oncology : official journal of the American Society of Clinical Oncology), USMSTF post-resection CRC surveillance (U.S. Multi-Society Task Force / American Gastroenterological Association), WHO Classification of Tumours Online: Skin Tumours (WHO Classification of Tumours / IARC), Abdominal Aortic Aneurysm (StatPearls / NCBI Bookshelf), Abducens Nerve Palsy (StatPearls / NCBI Bookshelf), Actinic Prurigo (StatPearls / NCBI Bookshelf), Acute Abdomen (StatPearls / NCBI Bookshelf), Acute Aortic Syndrome (StatPearls / NCBI Bookshelf), Acute Coronary Syndrome (StatPearls / NCBI Bookshelf), Acute Kidney Injury (StatPearls / NCBI Bookshelf), Acute Myocardial Infarction (StatPearls / NCBI Bookshelf), Acute Myocarditis (StatPearls / NCBI Bookshelf), Acute Pancreatitis (StatPearls / NCBI Bookshelf), Acute Pulmonary Embolism (StatPearls / NCBI Bookshelf), Acute Respiratory Distress Syndrome (StatPearls / NCBI Bookshelf), Acute-on-Chronic Liver Failure (StatPearls / NCBI Bookshelf), Adalimumab (StatPearls / NCBI Bookshelf), ADAMTSL4-Related Eye Disorders (GeneReviews / NCBI Bookshelf), Adenocarcinoma (Archived) (StatPearls / NCBI Bookshelf), Adrenal Metastasis (StatPearls / NCBI Bookshelf), ALK Negative Anaplastic Large Cell Lymphoma (StatPearls / NCBI Bookshelf), Amyloidosis (StatPearls

/ NCBI Bookshelf), ANCA-Associated Vasculitis (StatPearls / NCBI Bookshelf), Anetoderma (StatPearls / NCBI Bookshelf), Anorexia Nervosa (StatPearls / NCBI Bookshelf), Anterior Mediastinal Mass (StatPearls / NCBI Bookshelf), Aortic Dissection (StatPearls / NCBI Bookshelf), Aortic Rupture (Archived) (StatPearls / NCBI Bookshelf), Aortic Valve Endocarditis (StatPearls / NCBI Bookshelf), Aortico-Left Ventricular Tunnel (StatPearls / NCBI Bookshelf), Aphasia (StatPearls / NCBI Bookshelf), Apical Periodontitis (StatPearls / NCBI Bookshelf), Arrhythmogenic Right Ventricular Cardiomyopathy (StatPearls / NCBI Bookshelf), Arrhythmogenic Right Ventricular Cardiomyopathy Overview (GeneReviews / NCBI Bookshelf), Aseptic Meningitis (StatPearls / NCBI Bookshelf), Aspergillus Fumigatus (Archived) (StatPearls / NCBI Bookshelf), Asthma: Allergic, Nonallergic, and Environmental Phenotypes (StatPearls / NCBI Bookshelf), Atopic Dermatitis (StatPearls / NCBI Bookshelf), Atrial Fibrillation (StatPearls / NCBI Bookshelf), Atrial Flutter (StatPearls / NCBI Bookshelf), Atrial Septal Defect (StatPearls / NCBI Bookshelf), Atrioventricular Block (StatPearls / NCBI Bookshelf), Atrioventricular Block (Nursing) (StatPearls / NCBI Bookshelf), Atrioventricular Nodal Reentry Tachycardia (StatPearls / NCBI Bookshelf), Autoimmune Encephalitis (StatPearls / NCBI Bookshelf), Autoimmune Myopathies (StatPearls / NCBI Bookshelf), Bacteremia (StatPearls / NCBI Bookshelf), Baker's Cyst (StatPearls / NCBI Bookshelf), Benign Familial Pemphigus (Hailey-Hailey Disease) (StatPearls / NCBI Bookshelf), Bicuspid Aortic Valve (Archived) (StatPearls / NCBI Bookshelf), Biochemistry, Histamine (StatPearls / NCBI Bookshelf), Birth Trauma (StatPearls / NCBI Bookshelf), Blastic Plasmacytoid Dendritic Cell Neoplasm (StatPearls / NCBI Bookshelf), Bowel Ischemia (StatPearls / NCBI Bookshelf), Brain Abscess (StatPearls / NCBI Bookshelf), Branch Retinal Vein Occlusion (StatPearls / NCBI Bookshelf), BRASH Syndrome (StatPearls / NCBI Bookshelf), BRCA1 and BRCA2 Mutations (StatPearls / NCBI Bookshelf), Breast Implant Rupture (StatPearls / NCBI Bookshelf), Breast Myofibroblastoma (StatPearls / NCBI Bookshelf), Bronchiectasis (StatPearls / NCBI Bookshelf), Brucellosis (StatPearls / NCBI Bookshelf), Budd-Chiari Syndrome (StatPearls / NCBI Bookshelf), Bullous Emphysema (StatPearls / NCBI Bookshelf), Bullous Pemphigoid (StatPearls / NCBI Bookshelf), Calcinosis Cutis (StatPearls / NCBI Bookshelf), Carcinoid Syndrome (StatPearls / NCBI Bookshelf), Cardiac Arrest (StatPearls / NCBI Bookshelf), Cardiac Manifestations of Coronavirus (COVID-19) (StatPearls / NCBI Bookshelf), Cardiac Rhabdomyoma (StatPearls / NCBI Bookshelf), Cardiac Syncope (Archived) (StatPearls / NCBI Bookshelf), Cardiac Tamponade (StatPearls / NCBI Bookshelf), Cardioembolic Stroke (Archived) (StatPearls / NCBI Bookshelf), Caroli Disease (StatPearls / NCBI Bookshelf), Carotid Cavernous Fistula (StatPearls / NCBI Bookshelf), Catatonia (StatPearls / NCBI Bookshelf), Catheter Management of Mitral Regurgitation (StatPearls / NCBI Bookshelf), Central Retinal Vein Occlusion (StatPearls / NCBI Bookshelf), Central Serous Chorioretinopathy (StatPearls / NCBI Bookshelf), Cerebral Amyloid Angiopathy (StatPearls / NCBI Bookshelf), Cerebrospinal Fluid Leak (StatPearls / NCBI Bookshelf), Chest Pain (Archived) (StatPearls / NCBI Bookshelf), Choanal Atresia (StatPearls / NCBI Bookshelf), Cholangiocarcinoma (StatPearls / NCBI Bookshelf), Cholinergic Crisis (StatPearls / NCBI Bookshelf), Chronic Cough (StatPearls / NCBI Bookshelf), Chronic Diarrhea (StatPearls / NCBI Bookshelf), Chronic Inflammatory Demyelinating Polyradiculoneuropathy (StatPearls / NCBI Bookshelf), Chronic Kidney Disease (StatPearls / NCBI Bookshelf), Chronic Myelogenous Leukemia (StatPearls / NCBI Bookshelf), Chronic Sinusitis (StatPearls / NCBI Bookshelf), Chronic Urticaria (StatPearls / NCBI Bookshelf), CNS Lymphoma (Archived) (StatPearls / NCBI Bookshelf), Coccidioidomycosis (StatPearls / NCBI Bookshelf), Colon Cancer (StatPearls / NCBI Bookshelf), Constrictive-Effusive Pericarditis (Archived) (StatPearls / NCBI Bookshelf), Corneal Dystrophy (StatPearls / NCBI Bookshelf), Coronary Artery Disease (StatPearls / NCBI Bookshelf), Coronary Artery Dissection (Archived) (StatPearls / NCBI Bookshelf), Crescentic Glomerulonephritis (Archived) (StatPearls / NCBI Bookshelf), Crohn Disease (StatPearls / NCBI Bookshelf), Cryoglobulinemic Vasculitis (Archived) (StatPearls / NCBI Bookshelf), Cryptococcus (Archived) (StatPearls / NCBI Bookshelf), Cushing Syndrome (StatPearls / NCBI Bookshelf), Cutaneous Crohn Disease (StatPearls / NCBI Bookshelf), Cutaneous Cryptococcus (StatPearls / NCBI Bookshelf), Cutaneous Endometriosis (StatPearls / NCBI Bookshelf), Cutaneous Leiomyomas (StatPearls / NCBI Bookshelf), Cyanosis (Archived) (StatPearls / NCBI Bookshelf), Cystinosis (GeneReviews / NCBI Bookshelf), Cytomegalovirus Colitis (Archived) (StatPearls / NCBI Bookshelf), Cytomegalovirus Esophagitis (Archived) (StatPearls / NCBI Bookshelf), Cytomegalovirus Infections (StatPearls / NCBI Bookshelf), Dacryocystitis (StatPearls / NCBI Bookshelf), Darier Disease (StatPearls / NCBI Bookshelf), Degenerative Myopia (StatPearls / NCBI Bookshelf), Dengue Fever (StatPearls / NCBI Bookshelf), Dermatomyositis (StatPearls / NCBI Bookshelf), Dextrocardia (StatPearls / NCBI Bookshelf), Diarrhea (Archived) (StatPearls / NCBI Bookshelf), Diffuse Large B-Cell Lymphoma (StatPearls / NCBI Bookshelf), Dilated Cardiomyopathy (StatPearls / NCBI Bookshelf), Diplopia (StatPearls / NCBI Bookshelf), Discoid Lupus Erythematosus (StatPearls / NCBI Bookshelf), Diving Gas Embolism (StatPearls / NCBI Bookshelf), Drug-Induced Pigmentation (StatPearls / NCBI Bookshelf), Duodenal Perforation (StatPearls / NCBI Bookshelf), Dural Arteriovenous Fistula (StatPearls / NCBI Bookshelf), Dysphagia (StatPearls / NCBI Bookshelf), Dysphonia (StatPearls / NCBI Bookshelf), Dysplastic Nevi (Archived) (StatPearls / NCBI Bookshelf), Dystrophic Epidermolysis Bullosa (GeneReviews / NCBI Bookshelf), ECG T Wave (StatPearls / NCBI Bookshelf), Ectropion (StatPearls / NCBI Bookshelf), Eczema (StatPearls / NCBI Bookshelf), Embolic Stroke (Archived) (StatPearls / NCBI Bookshelf), Emerging Variants of SARS-CoV-2 and Novel Therapeutics Against Coronavirus (COVID-19) (Archived) (StatPearls / NCBI Bookshelf), Empyema (Archived) (StatPearls / NCBI Bookshelf), End-Stage Renal Disease (StatPearls / NCBI Bookshelf), Endocarditis (StatPearls / NCBI Bookshelf), Endogenous Endophthalmitis (StatPearls / NCBI Bookshelf), Endometrial Cancer (StatPearls / NCBI Bookshelf), Endophthalmitis (StatPearls / NCBI Bookshelf), Eosinophilia (StatPearls / NCBI Bookshelf), Epidermolysis Bullosa

(StatPearls / NCBI Bookshelf), Epidermolytic Hyperkeratosis (StatPearls / NCBI Bookshelf), Epididymitis (StatPearls / NCBI Bookshelf), Epiretinal Membrane (StatPearls / NCBI Bookshelf), Epithelial Downgrowth (StatPearls / NCBI Bookshelf), Epithelial Ovarian Cancer (StatPearls / NCBI Bookshelf), Epstein-Barr Virus (StatPearls / NCBI Bookshelf), Esophageal Cyst (StatPearls / NCBI Bookshelf), Extracorporeal Membrane Oxygenation in Adults (StatPearls / NCBI Bookshelf), Extramammary Paget Disease (StatPearls / NCBI Bookshelf), Extrapyramidal Side Effects (StatPearls / NCBI Bookshelf), Exudative Retinal Detachment (StatPearls / NCBI Bookshelf), Facial Nerve Palsy (StatPearls / NCBI Bookshelf), Familial Cerebral Cavernous Malformations (GeneReviews / NCBI Bookshelf), Fanconi Syndrome (StatPearls / NCBI Bookshelf), Features, Evaluation, and Treatment of Coronavirus (COVID-19) (StatPearls / NCBI Bookshelf), Febrile Neutropenia (StatPearls / NCBI Bookshelf), Fever of Unknown Origin (StatPearls / NCBI Bookshelf), Fibrinous Pericarditis (Archived) (StatPearls / NCBI Bookshelf), Foot Drop (StatPearls / NCBI Bookshelf), Forearm Fractures (StatPearls / NCBI Bookshelf), Fumarate Hydratase Deficiency (GeneReviews / NCBI Bookshelf), Fungal Infections of the Oral Mucosa (StatPearls / NCBI Bookshelf), Galactorrhea (StatPearls / NCBI Bookshelf), Galeazzi Fractures (StatPearls / NCBI Bookshelf), Gallstone Ileus (StatPearls / NCBI Bookshelf), Gastric Cancer (StatPearls / NCBI Bookshelf), Generalized Pustular Psoriasis (StatPearls / NCBI Bookshelf), Genetic Atypical Hemolytic-Uremic Syndrome (GeneReviews / NCBI Bookshelf), Giant Cell Arteritis (Temporal Arteritis) (StatPearls / NCBI Bookshelf), Gleason Score (StatPearls / NCBI Bookshelf), Glioblastoma Multiforme (StatPearls / NCBI Bookshelf), Glomeruloid Hemangioma (StatPearls / NCBI Bookshelf), Glycogen Storage Disease (StatPearls / NCBI Bookshelf), Granuloma Faciale (StatPearls / NCBI Bookshelf), Granulomatous Uveitis (StatPearls / NCBI Bookshelf), Graves Disease (StatPearls / NCBI Bookshelf), Hand, Foot, and Mouth Disease (StatPearls / NCBI Bookshelf), Head Trauma (Archived) (StatPearls / NCBI Bookshelf), Hearing Loss (StatPearls / NCBI Bookshelf), Hemolytic Anemia (Archived) (StatPearls / NCBI Bookshelf), Hemophagocytic Lymphohistiocytosis (StatPearls / NCBI Bookshelf), Hemorrhagic Fever Renal Syndrome (StatPearls / NCBI Bookshelf), Hemorrhagic Shock (StatPearls / NCBI Bookshelf), Hereditary Myopathy with Early Respiratory Failure (GeneReviews / NCBI Bookshelf), Hirsutism (StatPearls / NCBI Bookshelf), HIV in Pregnancy (StatPearls / NCBI Bookshelf), HIV-1–Associated Opportunistic Infections (StatPearls / NCBI Bookshelf), Horner Syndrome (StatPearls / NCBI Bookshelf), Horseshoe Kidney (StatPearls / NCBI Bookshelf), Human Papillomavirus (StatPearls / NCBI Bookshelf), Hypercortisolism (Cushing Syndrome) (StatPearls / NCBI Bookshelf), Hyperhidrosis (StatPearls / NCBI Bookshelf), Hyperkalemia (StatPearls / NCBI Bookshelf), Hyperthyroidism in Pregnancy (StatPearls / NCBI Bookshelf), Hypertriglyceridemia (StatPearls / NCBI Bookshelf), Hypertrophic Cardiomyopathy (StatPearls / NCBI Bookshelf), Hypertrophic Obstructive Cardiomyopathy (Archived) (StatPearls / NCBI Bookshelf), Hypogammaglobulinemia (StatPearls / NCBI Bookshelf), Hypomagnesemia (StatPearls / NCBI Bookshelf), Hypoparathyroidism (StatPearls / NCBI Bookshelf), Hypopigmented Macules (StatPearls / NCBI Bookshelf), Ictal-Interictal Continuum (StatPearls / NCBI Bookshelf), Idiopathic Guttate Hypomelanosis (StatPearls / NCBI Bookshelf), IgG4-Related Disease (StatPearls / NCBI Bookshelf), Infective Endocarditis (StatPearls / NCBI Bookshelf), Inferior Myocardial Infarction (StatPearls / NCBI Bookshelf), Inflammatory Breast Cancer (StatPearls / NCBI Bookshelf), Interpretation of Blood Clotting Studies and Values (PT, PTT, aPTT, INR, Anti-Factor Xa, D-Dimer) (StatPearls / NCBI Bookshelf), Intestinal Perforation (Archived) (StatPearls / NCBI Bookshelf), Intracranial Hypotension (StatPearls / NCBI Bookshelf), Intramedullary Spinal Cord Tumors (StatPearls / NCBI Bookshelf), Intravitreal Implants (StatPearls / NCBI Bookshelf), Invasive Lobular Carcinoma (StatPearls / NCBI Bookshelf), Iodine-131 Uptake Study (StatPearls / NCBI Bookshelf), Ischemic Stroke (StatPearls / NCBI Bookshelf), Junctional Epidermolysis Bullosa (GeneReviews / NCBI Bookshelf), Junctional Rhythm (StatPearls / NCBI Bookshelf), Juvenile Dermatomyositis (StatPearls / NCBI Bookshelf), Juvenile Xanthogranuloma (StatPearls / NCBI Bookshelf), Keratoacanthoma (StatPearls / NCBI Bookshelf), Keratoconjunctivitis (StatPearls / NCBI Bookshelf), Ketamine Toxicity (StatPearls / NCBI Bookshelf), Klebsiella Pneumonia (StatPearls / NCBI Bookshelf), Laboratory Evaluation of Immune Hemolytic Anemias (StatPearls / NCBI Bookshelf), Lactic Acidosis (StatPearls / NCBI Bookshelf), Langerhans Cell Histiocytosis (StatPearls / NCBI Bookshelf), Lateral Wall Myocardial Infarction (StatPearls / NCBI Bookshelf), Left Bundle Branch Block (StatPearls / NCBI Bookshelf), Leprosy (StatPearls / NCBI Bookshelf), Leukemia Cutis (StatPearls / NCBI Bookshelf), Leukocytoclastic Vasculitis (StatPearls / NCBI Bookshelf), Lichen Planus (StatPearls / NCBI Bookshelf), Lipodermatosclerosis (StatPearls / NCBI Bookshelf), Lipoid Pneumonia (StatPearls / NCBI Bookshelf), Liposarcoma (StatPearls / NCBI Bookshelf), Lisinopril (StatPearls / NCBI Bookshelf), Listeria monocytogenes Infection (Listeriosis) (StatPearls / NCBI Bookshelf), Liver Abscess (StatPearls / NCBI Bookshelf), Liver Metastasis (StatPearls / NCBI Bookshelf), Liver Toxicity (StatPearls / NCBI Bookshelf), Long QT Syndrome (StatPearls / NCBI Bookshelf), Lung Adenocarcinoma (StatPearls / NCBI Bookshelf), Lung Cancer (StatPearls / NCBI Bookshelf), Lymphoma (StatPearls / NCBI Bookshelf), Lymphoplasmacytic Lymphoma (StatPearls / NCBI Bookshelf), Macrocephaly (StatPearls / NCBI Bookshelf), Macroglossia (StatPearls / NCBI Bookshelf), Macular Edema (StatPearls / NCBI Bookshelf), Malakoplakia (StatPearls / NCBI Bookshelf), Malignancy-Related Hypercalcemia (StatPearls / NCBI Bookshelf), Malignant Melanoma (StatPearls / NCBI Bookshelf), Medial Medullary Syndrome (StatPearls / NCBI Bookshelf), Melanocytic Nevi (StatPearls / NCBI Bookshelf), Melanoma Pathology (StatPearls / NCBI Bookshelf), Meningocele (StatPearls / NCBI Bookshelf), Mesenteric Artery Ischemia (Archived) (StatPearls / NCBI Bookshelf), Metastatic Melanoma (StatPearls / NCBI Bookshelf), Microcephaly-Capillary Malformation Syndrome (GeneReviews / NCBI Bookshelf), Middle Ear Cholesteatoma (StatPearls / NCBI Bookshelf), Monoclonal Gammopathy of Undetermined Significance (StatPearls / NCBI Bookshelf), Morning Glory Syn-

drome (StatPearls / NCBI Bookshelf), Mucormycosis (StatPearls / NCBI Bookshelf), Mycobacterium kansasii Infection (Archived) (StatPearls / NCBI Bookshelf), Mycobacterium marinum Infection (Archived) (StatPearls / NCBI Bookshelf), Mycosis Fungoides (StatPearls / NCBI Bookshelf), Myelodysplastic Syndrome (StatPearls / NCBI Bookshelf), Myopathy (StatPearls / NCBI Bookshelf), Myopericarditis (StatPearls / NCBI Bookshelf), Myopic Foveoschisis (StatPearls / NCBI Bookshelf), Neck Abscess (StatPearls / NCBI Bookshelf), Necrolytic Migratory Erythema (StatPearls / NCBI Bookshelf), Necrotizing Fasciitis (StatPearls / NCBI Bookshelf), Neonatal Brain Tumors (StatPearls / NCBI Bookshelf), Neonatal MIS-C (StatPearls / NCBI Bookshelf), Neuroanatomy, Hypothalamus (StatPearls / NCBI Bookshelf), Neurofibroma (StatPearls / NCBI Bookshelf), Neurosarcoidosis (StatPearls / NCBI Bookshelf), Neurosyphilis (StatPearls / NCBI Bookshelf), Neutropenia (StatPearls / NCBI Bookshelf), Nevus Comedonicus (StatPearls / NCBI Bookshelf), Non-Hodgkin Lymphoma (StatPearls / NCBI Bookshelf), Nonseminomatous Testicular Tumors (StatPearls / NCBI Bookshelf), Nonviral Myocarditis (StatPearls / NCBI Bookshelf), Noonan Syndrome (GeneReviews / NCBI Bookshelf), Ocular Hypotony (StatPearls / NCBI Bookshelf), Ocular Melanoma (StatPearls / NCBI Bookshelf), Oliguria (Archived) (StatPearls / NCBI Bookshelf), Omphalitis (StatPearls / NCBI Bookshelf), Onychomycosis (StatPearls / NCBI Bookshelf), Ophthalmic Manifestations of Coronavirus (COVID-19) (StatPearls / NCBI Bookshelf), Optic Disc Drusen (StatPearls / NCBI Bookshelf), Optic Nerve Coloboma (StatPearls / NCBI Bookshelf), Optic Neuritis (StatPearls / NCBI Bookshelf), Oral Lichen Planus (StatPearls / NCBI Bookshelf), Pacemaker Malfunction (StatPearls / NCBI Bookshelf), Pachyonychia Congenita (GeneReviews / NCBI Bookshelf), Pancreatic Fistula (StatPearls / NCBI Bookshelf), Paraganglioma (StatPearls / NCBI Bookshelf), Parathyroid Carcinoma (StatPearls / NCBI Bookshelf), Parkinson Disease (StatPearls / NCBI Bookshelf), Parkinsonism (StatPearls / NCBI Bookshelf), Paroxysmal Supraventricular Tachycardia (StatPearls / NCBI Bookshelf), Parvovirus B19 Infection (StatPearls / NCBI Bookshelf), Patent Foramen Ovale (StatPearls / NCBI Bookshelf), Patent Urachus (StatPearls / NCBI Bookshelf), Pediatric Dyslipidemia (StatPearls / NCBI Bookshelf), Pemphigus Foliaceus (StatPearls / NCBI Bookshelf), Pemphigus Herpetiformis (StatPearls / NCBI Bookshelf), Pemphigus Vegetans (StatPearls / NCBI Bookshelf), Pemphigus Vulgaris (StatPearls / NCBI Bookshelf), Pericardial Effusion (StatPearls / NCBI Bookshelf), Pericarditis (StatPearls / NCBI Bookshelf), Perilymphatic Fistula (StatPearls / NCBI Bookshelf), Peripheral T-Cell Lymphoma (StatPearls / NCBI Bookshelf), Pheochromocytoma (StatPearls / NCBI Bookshelf), Physiology, Adrenocorticotropic Hormone (ACTH) (StatPearls / NCBI Bookshelf), Physiology, Liver (StatPearls / NCBI Bookshelf), Physiology, Tyrosine Kinase Receptors (StatPearls / NCBI Bookshelf), Pigmented Purpuric Dermatosis (StatPearls / NCBI Bookshelf), Pituitary Adenoma (StatPearls / NCBI Bookshelf), Pityriasis Lichenoides Et Varioliformis Acuta (PLEVA) (StatPearls / NCBI Bookshelf), Pityriasis Rubra Pilaris (StatPearls / NCBI Bookshelf), Plaque Psoriasis (StatPearls / NCBI Bookshelf), Plasmablastic Lymphoma (StatPearls / NCBI Bookshelf), Pleural Effusion (StatPearls / NCBI Bookshelf), Polycythemia (StatPearls / NCBI Bookshelf), Porokeratosis (StatPearls / NCBI Bookshelf), Portal Hypertension (StatPearls / NCBI Bookshelf), Premature Ventricular Complex (StatPearls / NCBI Bookshelf), Prinzmetal Angina (Archived) (StatPearls / NCBI Bookshelf), Prostate Cancer (StatPearls / NCBI Bookshelf), Prurigo Pigmentosa (StatPearls / NCBI Bookshelf), Pseudohypoaldosteronism Type II (GeneReviews / NCBI Bookshelf), Pseudomembranous Colitis (Archived) (StatPearls / NCBI Bookshelf), Pulmonary Hemorrhage (StatPearls / NCBI Bookshelf), Pulmonary Sequestration (StatPearls / NCBI Bookshelf), Pulseless Ventricular Tachycardia (Archived) (StatPearls / NCBI Bookshelf), Pustular Psoriasis (Archived) (StatPearls / NCBI Bookshelf), Q Fever (StatPearls / NCBI Bookshelf), Radiation Proctitis (StatPearls / NCBI Bookshelf), Rapidly Progressive Glomerulonephritis (StatPearls / NCBI Bookshelf), Rectal Bleeding (StatPearls / NCBI Bookshelf), Recurrent Acute Rhinosinusitis (StatPearls / NCBI Bookshelf), Relapsing Malaria (StatPearls / NCBI Bookshelf), REM Sleep Behavior Disorder (Archived) (StatPearls / NCBI Bookshelf), Renal Angiomyolipoma (StatPearls / NCBI Bookshelf), Renal Mass (StatPearls / NCBI Bookshelf), Renal Tubular Acidosis (StatPearls / NCBI Bookshelf), Respiratory Acidosis (StatPearls / NCBI Bookshelf), Restless Legs Syndrome (StatPearls / NCBI Bookshelf), Rhabdomyolysis (StatPearls / NCBI Bookshelf), Rhabdomyosarcoma (StatPearls / NCBI Bookshelf), Rhino-Orbital Cerebral Mucormycosis (Archived) (StatPearls / NCBI Bookshelf), Right Bundle Branch Block (StatPearls / NCBI Bookshelf), Schamberg Disease (StatPearls / NCBI Bookshelf), Schistosomiasis (StatPearls / NCBI Bookshelf), Schizophrenia (StatPearls / NCBI Bookshelf), Scleroderma (StatPearls / NCBI Bookshelf), Seborrheic Keratosis (StatPearls / NCBI Bookshelf), Seizure (StatPearls / NCBI Bookshelf), Sensorineural Hearing Loss (StatPearls / NCBI Bookshelf), Sinus Arrhythmia (StatPearls / NCBI Bookshelf), Sinus Node Dysfunction (StatPearls / NCBI Bookshelf), Small Bowel Obstruction (StatPearls / NCBI Bookshelf), Small Cell Lung Cancer (StatPearls / NCBI Bookshelf), Soft Tissue Clear Cell Sarcoma (StatPearls / NCBI Bookshelf), Spinal Cord Compression (StatPearls / NCBI Bookshelf), Spinal Cord Infarction (StatPearls / NCBI Bookshelf), Spontaneous Coronary Artery Dissection (StatPearls / NCBI Bookshelf), Spontaneous Intracranial Hypotension (Archived) (StatPearls / NCBI Bookshelf), Steatocystoma (StatPearls / NCBI Bookshelf), Strongyloidiasis (StatPearls / NCBI Bookshelf), Strongyloidiasis (Archived) (StatPearls / NCBI Bookshelf), Subarachnoid Hemorrhage (StatPearls / NCBI Bookshelf), Subcutaneous Emphysema (Archived) (StatPearls / NCBI Bookshelf), Subdural Hematoma (StatPearls / NCBI Bookshelf), Subungual Melanoma (StatPearls / NCBI Bookshelf), Suction Drains (StatPearls / NCBI Bookshelf), Superior Vena Cava Syndrome (StatPearls / NCBI Bookshelf), Supraventricular Tachycardia (StatPearls / NCBI Bookshelf), Sweet Syndrome (StatPearls / NCBI Bookshelf), Syncope (StatPearls / NCBI Bookshelf), Syringoma (StatPearls / NCBI Bookshelf), Systemic Lupus Erythematosus (StatPearls / NCBI Bookshelf), Systemic Sclerosis (Scleroderma) (StatPearls / NCBI Bookshelf), T-Cell Prolymphocytic Leukemia (StatPearls / NCBI Bookshelf), Temporal Artery Biopsy

(StatPearls / NCBI Bookshelf), Testicular Cancer (StatPearls / NCBI Bookshelf), Third-Degree Atrioventricular Block (StatPearls / NCBI Bookshelf), Thoracic Aorta Aneurysm (StatPearls / NCBI Bookshelf), Thrombotic Thrombocytopenic Purpura (StatPearls / NCBI Bookshelf), Thymoma (StatPearls / NCBI Bookshelf), Tilt Table (StatPearls / NCBI Bookshelf), Tinea Corporis (StatPearls / NCBI Bookshelf), Tinea Manuum (StatPearls / NCBI Bookshelf), Tinea Versicolor (StatPearls / NCBI Bookshelf), Tourette Syndrome and Other Tic Disorders (StatPearls / NCBI Bookshelf), Tracheoinnominate Fistula (StatPearls / NCBI Bookshelf), Tracheostomy Tube Change (StatPearls / NCBI Bookshelf), Trench Fever (StatPearls / NCBI Bookshelf), Trichosporonosis (StatPearls / NCBI Bookshelf), Tuberous Sclerosis Complex (GeneReviews / NCBI Bookshelf), Tympanic Membrane Perforation (StatPearls / NCBI Bookshelf), Ulcerative Colitis (StatPearls / NCBI Bookshelf), Unstable Angina (StatPearls / NCBI Bookshelf), Uveitic Macular Edema (StatPearls / NCBI Bookshelf), Uveitis (StatPearls / NCBI Bookshelf), Vasculitis (StatPearls / NCBI Bookshelf), Venous Insufficiency (StatPearls / NCBI Bookshelf), Venous Leg Ulcer (StatPearls / NCBI Bookshelf), Ventricular Fibrillation (StatPearls / NCBI Bookshelf), Ventricular Tachycardia (StatPearls / NCBI Bookshelf), Vertebral Osteomyelitis (StatPearls / NCBI Bookshelf), VEXAS Syndrome (GeneReviews / NCBI Bookshelf), Viral Encephalitis (StatPearls / NCBI Bookshelf), Viral Hepatitis (StatPearls / NCBI Bookshelf), Viral Myocarditis (StatPearls / NCBI Bookshelf), Viral Pneumonia (StatPearls / NCBI Bookshelf), Vitamin B12 Deficiency (StatPearls / NCBI Bookshelf), Vitamin C Deficiency (StatPearls / NCBI Bookshelf), Vitiligo (StatPearls / NCBI Bookshelf), Wells Syndrome (StatPearls / NCBI Bookshelf), Wernicke-Korsakoff Syndrome (Archived) (StatPearls / NCBI Bookshelf), Whipple Disease (StatPearls / NCBI Bookshelf), Wound Dehiscence (StatPearls / NCBI Bookshelf), Aicardi Syndrome (NCBI Bookshelf), EANO guideline on rational molecular testing of gliomas, glioneuronal, and neuronal tumors in adults for targeted therapy selection (Neuro-oncology), EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood (Nature reviews. Clinical oncology), EAU-EANM-ESTRO-ESUR-ISUP-SIOG Prostate Cancer Guidelines 2025 (European Association of Urology), ESMO Clinical Practice Guideline interim update on the treatment of locally advanced oesophageal and oesophagogastric junction adenocarcinoma and metastatic squamous-cell carcinoma (ESMO open), FBN1-Related Marfan Syndrome (NCBI Bookshelf), Hypermobile Ehlers-Danlos Syndrome (NCBI Bookshelf), Large B-cell lymphoma (LBCL): EHA Clinical Practice Guidelines for diagnosis, treatment, and follow-up (HemaSphere), Multiple Myeloma: EHA-ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-up (HemaSphere), Rotor Syndrome (NCBI Bookshelf), Wilson Disease (NCBI Bookshelf).