

SMART Guidelines, Smarter AI

Grounding Generative Clinical AI in WHO SMART Guidelines — and Accelerating the Guideline Lifecycle in Return

On why generative AI and computable clinical guidelines are markedly stronger together than apart — across the full knowledge lifecycle, from authoring to the point of care.

A practitioner perspective, written from the vantage of building and operating clinical knowledge systems at the systems level — with generative AI, knowledge graphs, and knowledge acquisition.

Discussion draft · June 2026

Contents

Executive Summary	1
1. The Setting: SMART Guidelines in the World of AI	2
2. Guidelines as the Deterministic Backbone for Generative Clinical Reasoning	3
3. The Layer Above the Standard: A Conceptual Domain-Knowledge Architecture.....	7
4. Generative AI to Accelerate the Guideline Content Lifecycle	9
5. Better Together: Synergy Across the Full Lifecycle.....	12
6. Recommendations for WHO	12
7. Conclusion	14
Selected References and Further Reading	14

Executive Summary

WHO's SMART Guidelines initiative arrives at a decisive moment. Generative AI is moving rapidly into clinical settings — often as a shadow technology (via personal subscriptions to AI), including in the low-resource, frontline settings SMART Guidelines are built to serve — yet a generative model, on its own, cannot guarantee that its clinical output is correct, evidence-based, or safe. This raises a strategic question: what is the right foundation for trustworthy clinical AI — standards-based computable guidelines, a purpose-built knowledge graph, a deliberate composition of the two, or something more, up to a multi-agent system with access to the full range of medical knowledge and patient data?

This paper argues that generative AI and computable clinical guidelines are not competing answers but complementary ones, and that they are markedly stronger together than apart — across the entire knowledge lifecycle, from authoring to the point of care. It also argues that the deepest leverage lies in a layer most discussions overlook.

In one direction, a computable guideline is among the most effective ways to make a generative system trustworthy. By keeping the model at the conversational surface and routing the actual clinical inference through deterministic, validated logic — the guideline's own decision logic, patient-state inference, evidence, and terminology, executed by a clinical reasoning engine — a guideline confines

an otherwise open-ended model and supplies much of the grounding it lacks. Because that engine is already callable as a software service (including over the Model Context Protocol, MCP), the same capability that powers formal decision support can ground an AI assistant, with bounded, checkable translation steps in between.

There is a further point, and it is the one most easily missed. The computable guideline is the most aligned and grounded layer for execution and interoperability — and it should be retained and invested in — but, by its own design, it is the computable representation of the guideline’s final work product: it carries the information and defers the deeper conceptualization, particularly of clinical workflow and of the way clinicians actually reason. The highest-leverage layer sits just above it: a conceptual domain-knowledge architecture that bridges formal semantics, an expression logic reaching beyond what the executable standards are designed to carry, and a human-readable surface approaching controlled natural language. That layer is a primary source of fidelity, and it is the anchor that lets a computable guideline and a knowledge graph reinforce one another rather than compete.

In the other direction, generative AI is among the most effective ways to substantially accelerate the production of computable guidelines themselves — the long-standing bottleneck between SMART Guidelines as an architecture and SMART Guidelines as shelf-ready capability in every country that needs it. Used as a design-time accelerator under a strict discipline — AI proposes, standards and determinism judge, experts decide — generative AI can take far more guidelines from narrative form (L1) to machine-readable and executable form (L3–L4), faster, with local adaptation built in and with fidelity and trust preserved.

These directions form a single, compounding loop: AI accelerates the authoring of computable guidelines; deployed guidelines then ground AI at the point of care; and real-world use and expert correction feed back to improve both. A computable guideline, anchored to the conceptual model, is one investment that returns three times over — it runs deterministically, it grounds generative AI, and it generates the structured signal that improves the authoring pipeline.

Beneath all of this is a choice about where clinical knowledge — and the trust placed in it — should live. Representing clinical guideline knowledge faithfully is a clinical-knowledge-representation problem, owned by an established, evidence-based, multidisciplinary discipline. Crucially, authority does not live in any datastore: it lives in the science, the evidence, and the expertise that produced the guidance, and the architecture’s job is to preserve that authority through every translation downstream. The most capable graph and machine-learning methods can be powerful instruments in the service of that discipline. On that foundation, generative AI takes its proper place: a true workhorse in expert hands, with the knowledge graph on the team — and even more to harness.

1. The Setting: SMART Guidelines in the World of AI

WHO’s SMART Guidelines — Standards-based, Machine-readable, Adaptive, Requirements-based, and Testable — express clinical and public-health recommendations as a sequence of knowledge layers, each with associated products that countries adapt and implement: narrative guidelines (L1); operational Digital Adaptation Kits, or DAKs (L2); machine-readable recommendations (L3); executable reference applications and services (L4); and dynamic, locally optimized models (L5). The purpose of the framework is to carry the meaning of a recommendation faithfully from prose to software — and to do so in a form each country can localize.

Two developments now press on this framework at once.

First, generative AI has entered clinical practice faster than any prior technology, and it is reaching the very settings SMART Guidelines are meant to serve. Frontline and primary-care workers in low-resource systems are already consulting large-language-model assistants — often through personal subscriptions, outside any governance — and AI-enabled records and decision support are being

piloted in exactly these contexts. The opportunity is real; so is the hazard. WHO's own guidance on large multi-modal models is candid about the risks — inaccuracy, bias, and unclear accountability — and the wider regulatory discussion increasingly distinguishes confined clinical AI, anchored to a defined indication and to validated logic, from unconfined systems that generate open-ended clinical advice and behave, in effect, like unregulated medical devices. A bare generative model used for clinical reasoning sits on the wrong side of that line.

Second, and partly in response, there is growing interest in giving generative AI a deterministic, governed foundation to reason against. A specific architectural question follows, and it is the one this paper addresses: should that foundation be WHO's own standards-based computable guidelines — the L3 and L4 layers of SMART Guidelines — a separately constructed knowledge graph (whether of medical concepts, of clinical content and literature, or of patients' longitudinal histories), or a deliberate composition of the two? Or, as this paper argues, something more? The framing matters, because it shapes where the responsibility for clinical reasoning sits and how the authority of the underlying evidence is preserved.

The answer this paper develops is that *a computable guideline and a knowledge graph are neither equivalent nor exclusive* — and that both rest on a deeper conceptual layer that is a primary source of fidelity. A **computable guideline** expressed through the HL7 FHIR Clinical Practice Guidelines Implementation Guide (CPG-IG) — the HL7 standard for computable guideline content, on which L3 implementation guides build — offers a different vantage than a concept graph: executable decision and orchestration logic, patient-state inference, a formal connection to the evidence base, and rigorous terminology, all under explicit lifecycle governance. A **knowledge graph**, in turn, can provide a broad, navigable web for grounding and discovery — whether of medical concepts, of clinical literature and content, or of patients' longitudinal histories and care — and both, it should be said, are complementary and can serve AI agents directly, as tools and as knowledge bases. The sections that follow make the case that the computable guideline should be the governed, executable home for clinical decision logic; that it and a knowledge graph are best unified rather than chosen between; that both are enriched by a conceptual domain-knowledge architecture above them — an explicit, formal representation that can itself be reasoned over; and that this is the path most faithful to how global health has always developed and trusted clinical knowledge.

2. Guidelines as the Deterministic Backbone for Generative Clinical Reasoning

The central difficulty with generative AI in clinical reasoning is not capability; it is trust. A large language model produces fluent, plausible output, but it cannot itself guarantee that the output is correct, evidence-based, or conformant — it asserts where it ought to defer. And even the strongest reasoning models, with an unbounded thinking budget and the best guardrails, depend on the harness that exposes the patient's data and the surrounding context to them: incomplete, outdated, or biased inputs can only compound the risk. The gap is now measurable: a recent multilingual benchmark spanning 95 models and 87 tasks drawn from real-world clinical text found a leading reasoning model that scores 92 on medical licensing-exam questions falling to 44.2 out of 100 on real clinical documentation — with no model, under any prompting strategy, exceeding 55.5 (Wu et al., 2025). WHO's own guidance on large multi-modal models names exactly these hazards: accuracy, bias, and accountability. The emerging regulatory conversation sharpens the point with a useful distinction between confined clinical AI — anchored to a specific indication and to validated logic — and unconfined systems that behave like open-ended, unregulated medical devices. Seen in that light, a computable guideline is not merely a content asset. It is one instrument that confines an otherwise unconfined model: it supplies validated, indication-specific, evidence-based reasoning to which the model can defer, and it defines the applicability boundaries and the preconditions for deploying the model safely.

The trust boundary: AI in the conversation, determinism in the reasoning path

The pattern that works keeps the non-deterministic model at the conversational surface — understanding the question, helping translate the data, explaining the result in plain language — and routes the actual clinical inference through deterministic, validated logic. The model proposes and narrates; the computable guideline judges; the human decides. What runs against a patient's data is deterministic, validated logic evaluated by a clinical reasoning engine, not the model's own guess. This is the design-time-determinism principle carried into the live conversation: AI proposes, standards and determinism judge, and — certainly in the care of patients — the cognitively augmented expert-in-context decides and ultimately provides the care.

The architecture is already within reach

The components exist today. In a guideline-authoring pipeline, the clinical reasoning engine is already invocable as a service and made available to agents over MCP. The exposure can take more than one shape, and the distinction matters: a guideline can be wrapped as a specific, goal-oriented tool — a named risk calculator, a particular guideline's recommendation service, with the knowledge baked in — or offered through a generic apply interface that takes the relevant knowledge artifact as a parameter, leaving the agent to discover which artifact is needed and run it. That discovery is itself aided by knowledge management: the standard envisions its knowledge assets — and the data they bind to — linked and navigable as a graph, so that an agent, or an author, can find which artifact applies, what it depends on, and what relates to or conflicts with it, at author time and at run time. Either way, if a pipeline agent can call the engine, so can a chat assistant. The engine itself is more than a rule evaluator: it performs derivations and calculations, diagnostic and risk assessments, outcome predictions, and decision recommendations over deterministic algorithms — rules, graphs, and validated predictive models alike — and it can orchestrate the same kinds of tools an agent would use, to supplement, ground, and verify the generative model. Notably, the reference engine WHO relies on is available as open source.

The connective tissue is translating the data a user supplies into the resources the engine consumes, and translating the result back — bounded, checkable tasks in both directions. Here a readable case representation does real work: the patient situation can be captured as a small set of knowledge-anchored clinical facts (a human-readable, machine-actionable rendering of the guideline's own case-feature definitions) that compiles deterministically to validated structured data, with each fact marked by the role it plays — asserted (an observed input), inferred (computed by the logic), or recommended (a proposed action). The engine then returns its result in that same readable, structured form: the inferred patient state, the recommendations, and — critically — the rationale tying them together, carried by the representation rather than generated by the model. The runtime loop is therefore deterministic and traceable end to end: the user describes a situation; it is structured into validated facts; the relevant guideline logic is applied and the case features — including the higher-value inferred features that depict the patient's clinical and care-process state — are resolved; the engine returns recommendations with provenance back to the originating recommendation and its patient-specific evidence; and the assistant renders them in plain language. The clinical inference and its justification are formal; the model's contribution is comprehension and phrasing at the edges. For WHO this is decisive in a practical way: the same computable guideline, deployed once, can serve formal decision-support integrations (EHRs, the L4 reference applications) and the fast-growing reality of frontline workers consulting an AI assistant. Rather than resist that reality, a computable guideline lets WHO lead and govern it.

Where AI earns its keep — and where it must be watched

This boundary deserves more honesty than it usually receives, because it cuts in two directions. At the input boundary, turning a real patient situation into the structured features the engine consumes is itself a hard problem — and when the inputs are unstructured (clinical notes, images, voice, device streams, time series), it requires AI models of its own, which deserve at least as much scrutiny as the

decision layer. An error here is the classic ‘garbage in, garbage out’: the determinism downstream faithfully executes on whatever it is handed, and the safety of the reasoning engine sits downstream of a fallible extraction step. The fragility is measurable: the same real-world benchmark found model accuracy on clinical normalization and coding — mapping narrative text to standard terminologies — stalling near 15 percent even with worked examples in the prompt (Wu et al., 2025). One robust pattern keeps the human clinician as the source of the values, with the assistant merely helping to acquire and structure them. At the output boundary the opposite is true: here, agents can do more than deliver recommendations. Because they hold the context, they can soften, gate, and verify what the deterministic layer returns, and assist with the individualization — patient preferences, exceptions, setting-specific factors — that no rule set can fully enumerate in advance, with a human in the loop and across multiple turns. The deterministic core is what makes a recommendation trustworthy; the agentic edges are where context, judgment, and individualization re-enter, under supervision. A computable guideline also has limits worth naming plainly: it depends on reasonably structured inputs, and it is only ever as complete as the knowledge that has been encoded. Acknowledging these is not a concession — it is what tells you where to put the scrutiny.

Grounding the conversation: retrieval as a runtime execution surface

Invoking the engine is not enough; the conversation must also be grounded in governed concepts, formal knowledge-organization systems, and validated reasoning patterns rather than in the model’s parametric memory. This is retrieval-augmented generation (RAG) — but over a formal, domain-specific knowledge architecture that sits above and informs well-curated, instance-data-informed domain models, and its shape is neuro-symbolic and, in practice, multi-spine. A structural spine traverses the governed knowledge graph for multi-hop, lineage-aware reasoning (graph-based retrieval). A semantic spine (dense vector retrieval) provides recall over narrative and descriptions. A pragmatic spine carries provenance, jurisdiction, and temporal governance — which guidance applies, where, and when. And a cross-representational spine aligns the same knowledge across its readable and computable forms, and across the alternative representations and reasoning capabilities each supports. The neural and symbolic legs are coupled at the most consequential point, generation itself: the model proposes within a typed grammar (grammar-constrained decoding), so its output is fluent yet type-safe by construction, not merely checked after the fact. The empirical record points the same way: on real clinical text, unaided step-by-step reasoning did not improve — and for many models degraded — performance, while supplying curated worked examples improved nearly every model tested; and medically fine-tuned models often lagged current general-purpose ones (Wu et al., 2025). Reliability comes from what the model is grounded in and held to — not from unaided introspection, and not from domain-specific tuning alone. The deeper consequence is a shift in what retrieval is for: run it over a mature formal architecture — ontologies, typed expressions, executable logic — and it ceases to be document lookup and becomes a runtime execution and usage surface of that architecture, with the formal logic acting as compile-time guardrails on what the model is permitted to say. Retrieval narrows the model to governed truth; the engine adjudicates the reasoning; provenance makes both auditable.

Guideline and graph: a complementary architecture

It is reasonable — and increasingly common — to build a knowledge graph for generative-AI grounding, and the case is strong. Typed entities and relationships genuinely outperform vector similarity for multi-hop reasoning and lineage; and in capable hands, a well-curated, ontology-backed graph is a powerful asset for navigating *what relates to what* — concepts, mechanisms, and the surrounding literature. None of what follows disputes that. The argument here is not graph versus guideline — a comparison that would pit a generic graph against a highly curated artifact and miss the point. It is about which form of knowledge is for which purpose.

Faithful clinical reasoning draws on several distinct forms of knowledge, and — this is worth stating plainly — most of them are themselves graphs. Patient data is a graph: resolve the references in a FHIR

record and a graph is exactly what remains. A care plan is a graph. Provenance is a graph; the W3C has standardized it as one. Terminology is the readable surface of a concept graph. The useful question is therefore not ‘graph or not,’ but which forms of knowledge a clinical system must represent, in which formalism each is best expressed, and how they are composed under a clean *separation of concerns*. Addressing exactly that — formalizing each form of knowledge in the representation suited to it, under one separation of concerns — is the work of a conceptual domain-knowledge architecture (Section 3); a computable guideline is then *one* governed, executable composition projected from it, bringing together forms that a single, flat graph of asserted facts does not, on its own, carry:

- **Decision and orchestration logic (the Plan) — itself a graph, and an executable one.** Recommendation and decision logic and longitudinal care pathways are encoded so that, when applied to a patient, they execute — producing concrete proposed actions for that specific patient, at that point in time, and in that context, not merely retrievable assertions. Cognitive patterns for best-practice, evidence-based, context-aware care — including temporal and distributed reasoning — are encoded in the Plan and then realized in the CarePlan.
- **Patient-state inference (the Case) — computed over the patient’s own data graph.** The guideline helps define the case features observed or inferred from the patient’s data, and the engine computes the clinically meaningful state the Plan reasons on. A graph can store that diabetes relates to nephropathy; the computable guideline computes whether *this* patient meets a criterion — the relative severity and risk, the state the patient has been in, is in now, and may be heading toward — at whatever resolution the use case demands and the architecture supports.
- **Evidence and provenance — related, but distinct.** Each recommendation carries its strength and its link to the evidence: PICOTS structures (population, intervention, comparator, outcome, timing, setting) tied to eligibility, the GRADE framework for the quality of evidence and the strength of recommendation, and formal Evidence and Evidence-Variable representations (EBM-on-FHIR) wired to the computable recommendation. Separately, it carries its provenance — where each element came from and how it was derived, itself a graph. And the guideline’s instances carry data as evidence in turn: an eCaseReport conveys knowledge-enriched information that feeds the learning health system and the ‘systems’ level of the evidence pyramid.
- **Terminology — the surface of concept graphs.** Value sets and well-formed post-coordinated expressions define the exact computable boundary of each concept, and code fidelity is reasoning fidelity. The computable guideline carries this natively — and it can reach well beyond traditional healthcare terminologies into numerous deeply interrelated ontologies for broader or orthogonal domains, scoped to the appropriate context. That reach is exactly what a formalized conceptual domain-knowledge architecture provides.

What a single graph struggles with is not relationships — it is everything that is not a flat, asserted fact. The knowledge-representation literature makes the point carefully: formalisms optimized for asserted, truth-conditional statements — whether RDF/OWL or labeled property graphs — do not natively carry the dependencies between statements, the epistemic status of a claim (asserted, reported, hypothesized, contested), directive and conditional logic, or structured argument from evidence to conclusion. Mechanisms to add these exist — reification, RDF-star, named graphs, property-graph edge attributes — but they are often verbose, cognitively opaque, or insufficiently standardized for the purpose (Vogt, 2026). The clinically critical case is the conflation of representing a claim with asserting it true: a grounding store that treats every stored statement as true cannot, by itself, reliably distinguish an established fact from a weak, contested, contraindicated, superseded, or inherently stateful one — exactly the distinctions clinical safety turns on. A computable guideline carries strength, evidence, provenance, and state natively. This is a representational-maturity point, not a verdict on

graphs — and indeed the same literature is converging on richer, graph-adjacent representations to close these very gaps.

With the forms of knowledge distinguished, the apparent dilemma dissolves — and the resolution is not that the guideline beats the graph. A shared conceptual domain-knowledge architecture (Section 3) defines the concepts the guideline, the patient data, and the broader literature graph are all about; each projects from it, and none is canonical ‘over’ the others. It is worth being precise about which graph one means: the graph of governed clinical knowledge, the graph of medical concepts and literature, and the patient’s own data graph (the longitudinal history of a person and their care) are related but distinct, and their union is itself a graph. The guideline’s principled domain modeling becomes the organizing backbone for the broader graph; the graph becomes the navigable evidence-and-concept web around the guideline; the patient’s data is the instance the whole apparatus reasons over. The right move is to represent the knowledge once, anchor it to a common semantic spine, and let each form do what it is for — composed, governed, and lowered to a standard so that the result is still deterministic at the point of care, simply by other means.

And the question of authority is worth restating, because it is easily placed wrong. Neither a graph nor a computable guideline is authoritative in itself. Authority lives in the science, the evidence, and the expertise that produced the guidance — and, increasingly, in the AI-accelerated work upstream of L1 as well. The role of the architecture, and of its provenance, is to preserve that authority through every downstream translation, so that the graph and the guideline can be trusted for the tasks they were designed for; the role of the deterministic engine is to execute it faithfully at the point of care. Provenance and architecture do not generate authority. They preserve it.

The question beneath the question: techno-centric, or health-centric?

There is a deeper reason to favor this composition, and it is the one that should weigh most heavily for WHO. The decision is often presented as a technology choice — a graph database plus a language model — but that framing miscategorizes the problem. Representing clinical guideline knowledge faithfully is, at bottom, a *domain-specific knowledge-representation* problem, and one that an established, multidisciplinary, evidence-based discipline already owns: guideline developers, subject-matter experts, clinical informaticians, terminologists, and knowledge engineers, working within methods and standards — HL7’s, and WHO’s own SMART framework — that are themselves the product of decades of deliberate collaboration among experts. The standards-based computable guideline is the artifact in which that expertise is deposited and made computable. The most advanced graph and machine-learning methods have an important role to play — but as powerful instruments in the service of that discipline, not as a substitute for it. So the question WHO should ask is not ‘which datastore,’ but this: *is this fundamentally a graph-and-generative-AI concern, or a clinical-knowledge-representation concern — and which must come first?* Posed that way the answer is clear, and it keeps the locus of trust, accountability, and evidence where global health has always placed it — with the experts and the evidence. Generative AI and graph technology then earn their place as workhorses in expert hands as care is delivered.

3. The Layer Above the Standard: A Conceptual Domain-Knowledge Architecture

The discussion so far has treated the computable guideline as the foundation. For execution and interoperability it is exactly that, and WHO has good reasons to build on it. But it is worth being precise about what that layer is and what it is not, because the single most consequential design decision lies just above it — and it is the one most easily skipped.

What the standard carries, and what it defers

The CPG-IG is candid about its own scope. It describes its task as realizing a guideline “through explicit formalisms and ultimately as information” — a computable representation of the guideline’s final work product. By design it defers the dimensions that are hardest and most local. It explicitly does not address the details of clinical workflow (though it can certainly carry that information), given their variation and complexity, handling workflow only through a thin “common pathway” abstraction. Its separation of concerns names three process axes — Case (the patient’s clinicopathological state), Plan (decision logic abstracted from setting), and Workflow (care delivery in a specific setting) — and then largely sets the third aside, at least from a domain-knowledge-architecture perspective. And the artifacts meant to align information to the way clinicians actually think — case-and-plan summary views shaped to clinician mental models and cognitive processes — are specified in the standard but, in practice, barely realized. None of this is a deficiency; it is a sensible scope boundary for an interoperability standard — one of remarkable sophistication and reach even so. But these intentional choices have an intentional consequence: the standard is an information-and-execution carrier, and the deeper intelligence — the conceptualization of clinical and care-process knowledge, and especially of the work domain and how clinicians reason and work within it — has to live somewhere above it.

The layer above: a conceptual domain-knowledge architecture

That “somewhere” is a conceptual domain-knowledge architecture: a layer that binds together three things the standard is intended to carry information about, but not to fully define. It carries formal semantics — an ontology of clinical, care-process, and work-domain concepts, spanning the very Case / Plan / Workflow axes, including the workflow and cognitive dimensions the standard defers (yet can still exchange for interoperability). It carries an expression logic that reaches beyond what the executable standards are designed to hold — not because CQL is weak or incomplete (it is neither), but because this layer expresses the domain knowledge, the work-domain context, and the reasoning rationale that the executable standards intentionally leave to other layers. This is in keeping with what FHIR’s originators intended: carry and exchange the information, but defer the full definition of the nature and behavior of each domain concept to the experts who own that concern. And it projects a human-readable surface that approaches a controlled natural language, so that the same knowledge a machine can act on is knowledge an expert can read and validate. It is, in effect, a hybrid representation — part ontology, part logic, part language — in which clinical knowledge is authored and reasoned about at a fidelity the work-product layer does not attempt. In this picture the computable guideline is a faithful projection of a richer model, not its origin. This is the formalization the forms-of-knowledge question of Section 2 was pointing toward — a single governed place where facts, decision and plan logic, computed patient state, evidence, provenance, terminology, workflow, and clinician cognition are each represented in the form suited to them, under one separation of concerns — and from which the same validated knowledge can be cast into a computable guideline (the CPG-IG), one or more knowledge graphs, or other representations, as each use case requires, without re-authoring the knowledge for each.

This layer is not a new idea; it predates the CPG-IG. Its approach was pioneered and implemented at a leading academic medical center, where a conceptual domain-knowledge architecture projected and governed computable artifacts across multiple run-time environments for a set of adaptive, reactive, and interdependent clinical pathways. Those pathways addressed some of the most demanding scenarios in the surgical care continuum — the perioperative and acute post-operative phases, with their complications, comorbidities, and cross-pathway dependencies — and, by combining knowledge-enriched patient data with explicit models of clinician reasoning and cognition, delivered advanced cognitive support for genuinely complex decisions, across care teams and the full quality-improvement and learning lifecycle. Grounded in sound scientific and informatics principles, it existed ultimately to serve clinicians working in high-stakes settings where excellence in patient care is paramount. Years later, a distillation of that conceptual model was described and realized in HL7’s CPG-IG (STU 1) as a formal guide to implementing computable clinical and care processes — with the deliberate intent of deferring the more sophisticated, formalized domain expression to the knowledge-engineering

lifecycle or, better still, to a future in which the conceptual architecture itself could be brought to bear within a more formal process. That time may now have come — and bringing it to bear is the opportunity this paper is written to name.

Why this is the locus of value

Three consequences follow, and together they make this layer the highest-leverage part of the architecture. First, it is a primary source of fidelity: meaning survives the crossing from narrative to executable code precisely because there is a rich conceptual model above the code; translate a guideline straight into an information carrier and the work-domain and cognitive fidelity — and precision — is exactly what is lost. Second, it is the anchor introduced in the previous section, but far more than a schema: because both a computable guideline and a knowledge graph can project from this shared model, it is what lets them coexist without either being canonical over the other, and it enriches both — the graph gains a principled organizing backbone, and the guideline gains conceptual depth it cannot express alone — even as the conceptual model itself is enriched, in turn, as both are used and as the system learns. Third, it is where the work-domain and cognitive alignment live — the summary views shaped to clinician reasoning that the standard specifies but does not yet realize. The deepest leverage in the whole architecture, then, is neither the graph nor the guideline, but the conceptual model that enriches both and is enriched by both.

A readable surface, and grounding in the literature

This is also where the human-readable, machine-actionable representation the paper has referred to becomes concrete — and where it pays to be exact, because there are two depths of such representation and they can coexist. One mirrors the standards-based computable guideline: a readable form of the work-product layer, close to FHIR and CQL. The other expresses the richer conceptual architecture above it, carrying the semantics, the broader expression logic, and the work-domain and cognitive dimensions. The two reinforce one another, the richer layer deepening the content of the guideline-level representation and of the computable guideline itself; and the same readable, knowledge-anchored treatment extends down to patient instances and their reasoning traces, so that a case — and the inference over it — can be carried across many interaction modalities while remaining as computable as the knowledge that produced it.

None of this is idiosyncratic; it is where the knowledge-representation field is heading. Recent work argues for treating statements as first-class semantic units with explicit identity and epistemic status — asserted, reported, hypothesized, contested — and for representations that are at once machine-actionable (FAIR) and cognitively interoperable and human-readable (the CLEAR principle), realized across existing technologies rather than as a new silo (Vogt, 2026). A conceptual domain-knowledge architecture of the kind described here is a clinical realization of exactly that direction, going a step beyond a general knowledge graph by carrying an expression logic and a near-natural-language surface. And it is not merely theoretical: representations of this kind have been built and operated in real clinical settings, and the computable guidelines they yield — including those expressed through the CPG-IG — are richer for it.

4. Generative AI to Accelerate the Guideline Content Lifecycle

If Section 2 shows the guideline grounding the AI and Section 3 the conceptual layer beneath both, this section shows the AI accelerating the guideline. It is the same investment viewed from the authoring end of the lifecycle.

The rate-limiting step

Authoring computable knowledge sets the pace of the entire enterprise. Moving from a narrative recommendation to, for example, executable FHIR and CQL is a crossing between two deep domains — clinical medicine and standards-conformant reasoning logic, L1 to L3 — and, as Section 3 argued,

the fidelity of that crossing depends on the conceptual model that sits above the code. Anyone who has carried a set of well-orchestrated, interrelated recommendations from narrative prose to running, validated logic knows the hard part is not writing the logic; it is preserving clinical meaning — faithfully, at scale, repeatably, and with trust. This bottleneck is precisely what has limited how many guidelines reach L3 and L4, how quickly countries can adapt them, and — just as much — the specificity, scope, and ultimate impact of those that are realized. It is the gap between SMART Guidelines as an architecture and many high-impact SMART Guidelines as shelf-ready capability for every country, and every person, that needs them.

The pattern: AI proposes, standards and determinism judge (constrain), experts decide

Generative AI is genuinely transformative here — but as an accelerator and workhorse at design time, not as an oracle. The discipline is trust by construction. Agents propose candidate and alternative representations — structured logic, terminology bindings, mappings — anchored to the conceptual model. Deterministic gates (augmented as needed by agentic adjudication) then judge those candidates for conformance and quality before a human ever looks; the model is explicitly forbidden from asserting conformance, semantics, or type facts, which are deferred to deterministic verifiers. Domain experts then review high-quality, pre-validated drafts — not raw model output — and hold the acceptance decision. Every artifact carries provenance on each node: model, version, confidence, evidence link, and any open ambiguities for an expert to resolve and for the system to learn from — so review is informed, auditable, and used to improve each future acquisition. The model proposes; it does not assert facts. Expressed as a repeating cycle: generate, judge, repair, anchor, route. And every agent works against scoped sub-graphs of the multi-spine neuro-symbolic retrieval layer, anchored in the conceptual domain-knowledge architecture. This is what the conceptual architecture buys the acquisition workflow: experienced experts work once, in a representation they can read and validate, and the same validated knowledge is then cast — by deterministic tooling, with the expert(s) holding the decision — into the CPG-IG, the knowledge graph(s), or whatever representation a given use case demands.

The constraint, moreover, is layered by design — multiple forms of determinism, applied at multiple levels. Typed grammars constrain generation itself, so candidates are well-formed by construction. Deterministic compilers and verifiers check syntax, types, terminology bindings, and standards conformance. Worked examples, settled with experts up front, are executed deterministically to validate behavior against expected results. And the reasoning engine remains the sole executor of the knowledge at the point of care. Such deterministic conformance gates, across all of these dimensions, are not incidental to agent-assisted acquisition — whether from narrative sources or from real-world evidence; for highly formalized, evidence-based, expert-informed expressions of clinical and care-process knowledge, they are what makes the acceleration safe at all. The empirical record underscores the point in both registers: the same benchmark evidence that counsels confinement at run time — normalization and coding near 15 percent, no model above 55.5 on real clinical text, with performance degrading further outside English — describes precisely the class of extraction, structuring, and coding work the authoring pipeline performs at design time. Generative AI is a vital participant at both ends, and a bounded one at both ends. It is no accident that the benchmark's strongest lever — supplying worked examples — is this pipeline's native method, drawn from governed knowledge rather than ad hoc prompting.

The supporting machinery

These are not aspirational pieces. In working systems, the machinery takes a recognizable shape, layer by layer: the conceptual domain-knowledge architecture of Section 3 as the semantic and logical substrate; readable representations — at the two depths described — as the surface experts and models author against; a deterministic compiler and verifier suite that lowers those representations faithfully to conformant FHIR and CQL through an intermediate representation; a clinical reasoning engine that

executes them; a governed content store — the standard’s own Knowledge Content Management System — that manages the resulting artifacts, their versions, and their dependencies; as well as a broad knowledge graph that serves at once as governed memory, as a grounding substrate for the agents, and as a learner across the whole ecosystem. The standard itself anticipates that last piece: it describes an advanced knowledge-management capability in which knowledge assets and the data they bind to are linked as a graph and mined — including with machine learning — to surface related, similar, or conflicting assets and even candidate case features, at author time and at run time. The particular names matter less than the shape — and the shape is what survives repeated implementation and refinement against real guidelines and real clinical data. What it buys is the discipline above made concrete: AI accelerates the work, while determinism and experts remain the deciding authorities.

Knowledge arrives from several directions

The same machinery accommodates knowledge entering from several distinct directions — and here, too, the frame is the standard’s own. The CPG-IG names the acquisition methods of its knowledge-engineering lifecycle — elicitation from experts, de novo authoring, synthesis from narrative and existing artifacts, discovery from data, and hybrids of these — and, critically, requires that knowledge derived by discovery be vetted through the evidence ecosystem and the guideline development group before it is trusted. In practice these become distinct acquisition workflows over one shared substrate, with a single control plane routing a given input across whichever apply. Narrative acquisition — the standard’s synthesis — takes a guideline from prose to computable form. Discovery turns real-world data and the patterns mined from it into candidate knowledge. Runtime refinement turns feedback from deployed knowledge — overrides, outcome discordances, exceptions, events — into corrections. Legacy lift takes existing computable assets back up into the readable conceptual layer so they can be understood and improved. And substrate evolution coordinates updates to the shared model itself. Each workflow runs the same disciplined cycle — generate, judge deterministically, repair, anchor to the conceptual model, route — and every correction, and every discovered incompleteness or inexactness, becomes structured, labeled signal (wrong concept, wrong intent, unsafe logic, missing exception, unsupported expression) that improves the pipeline rather than vanishing into a single edit. This is also where machine learning has its proper place: not generating clinical assertions for patients, but mining real-world data for candidate patterns that experts and standards then adjudicate — exactly the vetting the standard prescribes.

An Agile, compounding lifecycle — and it is the standard’s own method

Critically, this is not a process bolted onto guideline work; it is the development approach the CPG-IG itself prescribes — an integrated, cross-functional, rapid-cycle (Agile) process, inclusive of local implementation, with closed feedback and feed-forward loops across the knowledge, practice, and discovery ecosystems. The standard goes further still: it frames the computable guideline as the means of reaching the “systems” level of the evidence pyramid — where evidence is integrated directly with patient information and delivered into the clinical workflow (in many forms) — and it provides a formal case-report instrument that captures the data requirements, the inferred patient state, the recommendations issued, and the resulting clinical activity, from conception to delivery, expressly to feed outcomes research and continuous improvement. Generative AI is what finally makes that prescribed method practical at scale. Two-tier validation separates meaning from execution, so that when results diverge an expert can localize the fault — is it the meaning, the expression, the translation, or the runtime evaluation? Specification by worked example — the standard’s own test-driven knowledge engineering — lets domain experts settle “given this patient, the expected result is X” up front, so disagreements surface early and cheaply rather than in late-stage quality assurance that frustrates implementers and countries alike. Learning compounds across guidelines and across country adaptations: each correction makes the next guideline, and often the next action, better.

The surprising payoff

The highest-leverage use of AI in this work is not authoring the content; it is building the deterministic machinery itself — the language toolchain, the editors, the verifiers, the knowledge services. AI helps construct the very guardrails that then constrain and check AI's authoring. That layered, almost recursive use of the technology is what turns “at scale” from aspiration into practice — and it is what stands behind the claim that more guidelines can reach L3 and L4, faster, with country adaptation built into the method and with fidelity and trust preserved, because experts and standards remain in the deciding seat throughout.

5. Better Together: Synergy Across the Full Lifecycle

The directions described above are not separate initiatives; they are arcs of one loop — the closed knowledge-to-delivery-to-discovery cycle that the CPG-IG itself is built to support, turning on the conceptual model at its center.

Generative AI accelerates the authoring of computable guidelines (Section 4). Those guidelines, once deployed, ground generative AI at the point of care (Section 2). The clinical use that follows — and, above all, the corrections and observations that expert reviewers and frontline practice generate — feed back as structured signal that improves both the guidelines and the pipeline that produces them. Each turn of the loop raises quality and lowers the cost of the next. Generative AI and deterministic knowledge representation each create value inside the other's context: the AI is made trustworthy by the guideline; the guideline is made affordable and current by the AI.

One investment, three returns

This is where the choice of foundation pays off most visibly. A computable guideline, anchored to the conceptual model, is a single investment that returns three distinct ways:

- **It runs.** The same artifact executes deterministically against patient data through a clinical reasoning engine — in formal decision support and in AI-grounded conversation alike.
- **It grounds.** It is a rich retrieval substrate for generative AI, carrying executable logic, inference, evidence, and terminology rather than mere associations, and — anchored to the shared conceptual model — it helps organize the broader knowledge graph as well.
- **It teaches.** Its structure makes expert corrections capturable as labeled signal, so every clinical judgment improves the authoring pipeline and compounds across guidelines and country adaptations. And it captures rich information on patient state and the state of care — longitudinally, at scoped resolution, including indicators, interventions, and outcomes, all aligned to the conceptual domain-knowledge architecture. Both the case-and-plan summary view and the eCaseReport provide an unusually rich substrate from which to learn.

Used only as a grounding store, a knowledge graph contributes the second of these returns; the computable guideline, the broader graph, and the conceptual model together deliver all three. The synergy is not incidental; it is a direct consequence of representing the knowledge once, well, anchored to a common semantic spine, and lowered to a standard for interoperable, deterministic computability.

6. Recommendations for WHO

For WHO and the SMART Guidelines community, several practical directions follow. They are offered in support of, and to strengthen, the direction WHO is already taking: governed, computable, evidence-based clinical knowledge, evaluated before deployment.

1. **Anchor clinical AI on computable guidelines, intentionally unified with — not replaced by — a knowledge graph(s).** Prioritize L3 and L4 computable guidelines as the deterministic substrate that both formal decision support and AI assistants reason against, and

treat a clinical knowledge graph as a complementary evidence-and-retrieval layer rather than a separate source of truth. Govern the two through a shared domain ontology (or formal conceptual domain-knowledge architecture) so that the guideline carries the decision logic, the graph carries the broader web, and neither is canonical over the other — with all of them learning and improving together, and appropriate experts in the loop across the continuum.

2. **Invest in the conceptual layer above the standard, not only in the standard.** The computable guideline is the most aligned execution and interoperability target and should be retained; but the fidelity of clinical and care-process knowledge — especially the dimensions the standard defers — depends on a conceptual domain-knowledge architecture above it. Treat that layer, and the readable representations that surface it, as first-class assets: they are a primary source of fidelity and the anchor that unifies guideline and graph. The CPG-IG supports asserted and inferred Case Features, the conceptual domain-knowledge architecture establishes the rules for the formalizing the clinical pathophysiological processes that actually define these Case Features (their nature and behavior).
3. **Adopt the design-time / runtime trust boundary explicitly, and verify in both registers.** Use generative AI as a design-time accelerator and as a conversational interface, but keep clinical inference on deterministic, validated logic evaluated by a reasoning engine — while attending honestly to the input boundary, where extracting structured features from unstructured data is itself a fallible, scrutiny-worthy AI task. Verify by construction — the model proposes and explains, the computable guideline formalisms judge, and the clinician decides — and verify by evaluation, benchmarking AI outputs against WHO-governed reference standards. Together these keep AI-assisted care on the confined side of the regulatory line, and lead toward trusted clinical intelligence.
4. **Make the authoring method ‘AI proposes, standards judge, experts decide.’** Where WHO, countries and partners use AI to accelerate the translation from narrative to computable form, require deterministic conformance checks and expert acceptance as gates, with provenance recorded on every artifact. This is the CPG-IG’s own Agile, closed-loop development approach, now made practical at scale and enabling speed-to-value.
5. **Strengthen L2 with a formal, readable, AI-ready representation — without imposing a new language.** Extend the Digital Adaptation Kit with a human-readable, machine-actionable representation — cognitively interoperable for experts, information-dense for generative tooling, and deterministically lowerable to FHIR and CQL. The aim is to enrich the layer countries already work in, not to ask them to write CQL. This gives clinical informaticians the tools they badly need to bridge the gap between the subject-matter expert and the computable representation of their knowledge — and the same intermediate representation bridges generative AI, the standards, the human, and the conceptual domain-knowledge architecture itself.
6. **Expose reasoning as a service, and instrument the feedback loop.** Where guidelines are deployed, make the reasoning engine callable as a tool (for example, over MCP) so that a single computable guideline can serve EHR-integrated decision support, standalone cognitive support apps, and AI assistants alike; and capture expert corrections, field observations, and the guideline’s own case-report data as structured signal so the knowledge base and the authoring pipeline improve over time (systems-level evidence in action).
7. **Keep the locus of trust health-centric.** In governance, procurement, and capability-building, treat faithful guideline representation as a clinical-knowledge discipline led by guideline developers, subject-matter experts, clinical informaticians, terminologists, and knowledge engineers — with AI and graph technology as powerful tools in their hands. This is

both the safer choice and the one most consistent with WHO's existing guideline-development apparatus.

7. Conclusion

Generative AI is not magic, and it is not a clinician. It is a very powerful tool — and, in the right hands, it can build clinical knowledge that is trustworthy, scalable, and genuinely valuable. The “right hands” are not incidental to that sentence; they are its subject. The approach succeeds because it keeps clinical knowledge where it belongs — with the guideline developers, subject-matter experts, clinical informaticians, terminologists, and knowledge engineers who have spent decades building the evidence-based methods and standards that SMART Guidelines themselves embody — and it puts AI to work in service of that expertise rather than in place of it, ideally directed by experts in both.

The alternative framing — treating faithful clinical knowledge as primarily a database-and-model problem — answers a narrower question than the one in front of WHO. A knowledge graph and a language model are genuinely valuable, but they are most valuable in service of the standards-based computable guideline and the conceptual model above it, not in place of either. The whole is much greater than the sum of the parts. SMART Guidelines already make the better choice institutionally; the argument of this paper is that the same choice is well founded technically — that the computable guideline should be kept and built upon, that the deepest leverage is the conceptual layer that enriches it, and that generative AI and knowledge graphs are strongest when unified on that foundation.

Value first, then speed-to-value. Computable guidelines and generative AI, taken together and in that order — and resting on a conceptual domain-knowledge architecture built and refined by people who do this work at the systems level — can deliver both: a genuinely new capability to author high-value computable clinical knowledge at scale, and to place it, safely and deterministically, at the point of care. SMART Guidelines, made smarter by AI; AI, made trustworthy by SMART Guidelines.

Matthew M. Burton, MD, FAMIA — Physician Informatician, Health Systems Engineer, Clinical Knowledge Architect, Advisor to numerous Guideline Development organizations; former Medical Informatics faculty at Regenstrief Institute/Indiana University, Mayo Clinic, and Arizona State University. Co-Editor and implementer of the CPG-IG; Founding Lead for the Covid-19 Digital Guideline Working Group; co-developer of AI-assisted Knowledge Acquisition for CPG-IG and derived content.

Sivaram Arabandi, MD, General Surgery, MS (Computing) — Physician Informatician and Ontologist, Software Engineer for Clinical Information Systems, former Chief Knowledge Architect for numerous Terminology/Ontology vendors; bridging the gap between evidence-based science and computable systems to drive pragmatic, resilient AI implementation. Contributor and advisor to numerous ontology standards and HL7 FHIR CPG-IG.

Davide Sottara, PhD — Software Engineer, AI Computer Scientist (Hybrid Reasoning), Knowledge Engineer and Architect, AI/ML Engineer/ Enterprise Architect. Former assistant professor of Biomedical Informatics at Arizona State University; Co-author of Health eDecisions (FHIR Clinical Reasoning and CQL predecessor); Contributor to the HL7 FHIR CPG-IG; Co-author of the OMG API4KP standard for knowledge-based platforms; Contributor to the BPM4Health authoring guide, and former co-chair of the BPM4Health authoring group.

Selected References and Further Reading

The sources below inform the framing and arguments above and are offered as further reading rather than as exhaustive citation.

- World Health Organization. SMART Guidelines: Standards-based, Machine-readable, Adaptive, Requirements-based, Testable. <https://smart.who.int>

- WHO SMART guidelines: optimising country-level use of guideline recommendations in the digital age. The Lancet Digital Health, 2021.
- World Health Organization. Digital Adaptation Kits (SMART Guidelines L2) and the SMART Guidelines Starter Kit. <https://smart.who.int/ig-starter-kit>
- HL7 International. FHIR Clinical Practice Guidelines Implementation Guide (CPG-on-FHIR), v2.0.0 – including the guideline-development, conceptual knowledge architecture, and evidence-ecosystem documentation. <https://hl7.org/fhir/uv/cpg>

Key sections:

- Approach — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/approach.html>
- Methodology — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/methodology.html>
- Knowledge Engineering Overview (the acquisition-through-implementation lifecycle; roles, including the unifying knowledge architect; knowledge management) — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/documentation-approach-05-knowledge-engineering-overview.html>
 - Knowledge Acquisition (elicitation, authoring, synthesis, discovery, and hybrid methods; vetting of discovered knowledge) — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/documentation-approach-05-01-knowledge-acquisition.html>
 - Agile CPG Development Approach (test-driven knowledge engineering; specification by example; the Knowledge Content Management System) — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/documentation-approach-05-06-agile-cpg-development-approach.html>
- Guideline Development & the Evidence Ecosystem (PICOTS, GRADE, EBM-on-FHIR, the 6S pyramid / Learning Health System, the CPGeCaseReport feedback instrument) — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/documentation-approach-04-guideline-development.html>
- Knowledge Architecture — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/documentation-approach-06-03-knowledge-architecture.html>
- Tiers of Functionality — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/documentation-approach-07-tiers-of-functionality.html>
- Conceptual Knowledge Architecture (separation of concerns; the Case / Plan / Workflow model) — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/documentation-approach-12-conceptual-knowledge-architecture.html>
 - Separation of Concerns — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/documentation-approach-12-01-separations-of-concerns-in-the-cpg.html>
 - Separating & Defining Case, Plan, and Workflow — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/documentation-approach-12-02-separating-and-defining-case-plan-and-workflow.html>
 - Workflow & the Common Pathway — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/documentation-approach-12-06-cpg-common-pathway.html>
 - Derivative & Related Assets (CPGMetrics, CPGMeasures, CaseSummary, CasePlanSummaryView, CPGeCaseReport) — <https://build.fhir.org/ig/HL7/cqf-recommendations/en/documentation-approach-12-07-cpg-derivative-and-related-assets.html>
- Embedded references: EBM-on-FHIR IG — <https://hl7.org/fhir/uv/ebm> · GRADE Working Group — <https://www.gradeworkinggroup.org> · FHIR Quality Measures and electronic case-reporting IGs (linked from the Evidence and Derivative-Assets sections).
- HL7 International. FHIR Clinical Reasoning Module and the Using CQL with FHIR Implementation Guide.
- HL7 International. Evidence-Based Medicine on FHIR (EBM-on-FHIR) Implementation Guide — Evidence and EvidenceVariable resources, PICOTS.

- GRADE Working Group. Grading of Recommendations Assessment, Development and Evaluation (quality of evidence and strength of recommendation).
- Vogt, L. The Semantic Units Framework: a technology-agnostic representational approach to FAIR and CLEAR knowledge infrastructures. *Scientific Data*, 2026; 13:936.
<https://doi.org/10.1038/s41597-026-07588-3>
- Wu J, Gu B, Zhou R, et al. BRIDGE: Benchmarking large language models for understanding real-world clinical practice text. 2025 (preprint; continuously updated public leaderboard). 95 LLMs evaluated across 87 tasks from 59 real-world clinical data sources in nine languages.
- World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. 2024 (updated 2026).
- Weissman GE, Mankowitz T, Kanter GP. Unregulated large language models produce medical device-like output. *npj Digital Medicine*, 2025.
- Regulation of clinical AI in the age of agents: unconfined non-deterministic clinical software (UNDCS) systems for healthcare. *npj Digital Medicine*, 2026.
- Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. *arXiv:2401.11120*.
- Retrieval-augmented large language model for clinical decision support with a medical knowledge graph. *Electronics*, 2026.
- Retrospective evaluation of a generative-AI-enabled electronic medical record system in primary health-care facilities in Kenya. *medRxiv*, 2025.