Ontology-Grounded Semantic Routing Under Diagnostic Uncertainty: A Biolink-Federated CDSS for Emergency Care

Roberto Shimizu[1,*], Thiago Porto[1] and Roberta Garcia[1]

[1]*Intellidoctor.ai, Brazil*

Abstract

Emergency-department clinical AI is often evaluated as if its first task were to predict the final ICD-10 code. At first assessment, however, the actionable question is often which institutional care pathway the patient should enter. We report a production clinical decision-support system (CDSS) for adult emergency care in a large Brazilian private hospital. It routes uncertain disease hypotheses through a Biolink-federated ontology layer linking symptoms, diseases, evidence, CID/ICD-10 families, institutional protocols, exams, medications, and CPOE order sets. The resulting path from hypothesis to protocol family and care actions is inspectable. Among 923 initial visits with both an AI and a physician ICD-10 code, 66.7% matched on the exact code, rising to 73.1% routing-aware concordance once differing codes that map to the same institutional protocol family are included. Because physicians could view the AI output before coding, these are operational-concordance rather than independent diagnostic-accuracy measures. We describe the architecture and three production lessons: semantic routing is a useful early metric, prevalence priors improve relevance, and knowledge-graph incompleteness requires multi-evidence safeguards.

Keywords

Ontology-grounded clinical decision support, Clinical knowledge graphs, Biomedical ontologies, Biolink Model, Semantic routing, Healthcare AI, Explainable AI, Software as a Medical Device

1. Motivation: From Black-Box Diagnosis to Auditable Routing

Software that supports clinical decisions is moving into a regulatory environment in which performance alone is not enough. Brazil regulates Software as a Medical Device under RDC 657/2022 [1], while international guidance for SaMD and health AI emphasizes clinical evaluation, transparency, human oversight, and risk management [2, 3, 4]. We do not claim regulatory approval in this paper. The practical lesson from deployment is narrower but important: for an emergency-department CDSS, the output must be inspectable by physicians and, in future, defensible to regulators. A purely generative LLM answer is a poor substrate for that scrutiny. It may be fluent, but it does not by itself expose which symptoms were recognized, which disease family was selected, which institutional directive applied, or which exams and medications came from the care pathway. This is where biomedical ontologies matter in production: they provide controlled concepts, typed relations, and stable crosswalks from clinical language to auditable workflow decisions.

Our deployment therefore treats the LLM as a component, not as the decision substrate. The actionable layer is semantic routing. In emergency care, first assessment is often syndromic: the physician may not yet know whether the final code will be, for example, acute sinusitis, unspecified upper-respiratory infection, or acute pharyngitis. What matters operationally is that all three may enter the same institutional upper-airway protocol, where tests, medications, reevaluation, discharge, and admission decisions are handled. Exact CID/ICD-10 coding remains necessary, but it is not always the right primary objective at the first decision point. The CDSS should route the case to the right pathway under uncertainty, while keeping the trace from clinical signal to protocol visible.

ISWC 2026 Companion Volume, October 25–29, 2026, Bari, Italy

*Corresponding author.

rshimizubr@gmail.com (R. Shimizu); thiagoreisporto@gmail.com (T. Porto); robdallagogar@gmail.com (R. Garcia)

0009-0009-6330-4304 (R. Shimizu)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Architecture: Biolink as the Federation Layer

The system is deployed as the CDSS engine behind IntelliDoctor's emergency-care workflow, alongside transcription and documentation components. Its semantic layer has two connected parts.

First, a disease–phenotype knowledge graph maps symptoms and diseases into standard biomedical identifiers. Symptom-frequency and demographic data are aligned to HPO phenotypes [5], MONDO diseases [6], and publication identifiers, and represented as Biolink Model RDF [7]. In this model, a disease–phenotype edge is not an untyped text relation: it is a `biolink:DiseaseToPhenotypicFeatureAssociation` with subject, object, predicate, frequency, evidence, and source. The graph is served through Apache Jena Fuseki and SPARQL [8, 9]. Biolink is therefore not merely a file format; it is the cross-ontology bridge that makes disease, phenotype, evidence, prevalence, and demographic modifiers queryable in one model.

Second, the hospital CDSS layer maps disease hypotheses to institutionally curated protocol families. The local CDSS database contains 27 protocol families and 132 variations. Diagnostic variations carry MONDO, SNOMED CT, and UMLS identifiers where available; protocol care actions carry exam and medication codes including LOINC, RxNorm, ATC, and SNOMED CT. A reviewed mapping layer resolves CID/ICD-10 codes into protocol families by exact match and prefix anchors. This is the semantic route from an uncertain disease hypothesis to an actionable pathway: clinical presentation → disease hypotheses → CID family → institutional protocol → exams, medications, and CPOE order sets. A typical route is therefore inspectable as `J019` (acute sinusitis) or `J069` (unspecified upper-respiratory infection) → `IVAS_012` → upper-airway institutional protocol → suggested exams, medications, and reevaluation criteria.

The runtime query strategy is deliberately two-directional. A top-down retrieval step uses biomedical embeddings to obtain a bounded candidate set quickly. A bottom-up step verifies candidates against symptoms, disease–phenotype associations, prevalence priors, demographic modifiers, and protocol-family mappings. Concept extraction and disambiguation remain difficult because vignette terms vary and biomedical NER models are uncertain. An AI agent can explore ontology labels, synonyms, equivalent terms, and context, but actionable output remains routed through explicit identifiers and protocol records rather than through free-form LLM reasoning alone.

3. Deployment Snapshot and Routing Metric

The system is in ongoing production use in an adult emergency-department deployment in a large Brazilian private hospital setting. Because volume grows daily, we report telemetry as a submission-time snapshot rather than as the identity of the paper. To target initial routing we retained the earliest record per visit, removing the 47 repeat rows contributed by 36 revisited cases and leaving 960 initial visits from about thirty physicians; 923 of these had both AI and physician ICD-10 codes. The product also includes transcription, which reduced documentation time from approximately 13 to 8 minutes locally; we treat that number as transcription context, not as a CDSS outcome. Only aggregate results are reported; no patient-level data or internal identifiers are disclosed. De-identified validation vignettes contain only complaints, symptoms, age, and sex.

For the semantic CDSS, the evaluation question is different: when the AI and physician use different CID/ICD-10 labels, do they nevertheless route to the same institutional protocol family? Table 1 summarizes the snapshot. Exact ICD-10 string match is useful but too strict for first-assessment routing. We therefore report a routing-aware concordance metric: exact ICD-10 agreement plus cases where different codes map to the same protocol family.

The 59 "different code, same family" cases produce a modest 6.4-percentage-point lift over exact agreement: workflow-equivalent disagreement that exact-code metrics misclassify. Wilson 95% confidence intervals are 63.6–69.7% for exact agreement and 70.2–75.9% for routing-aware concordance. The two are nested rather than independent estimates: routing-aware concordance reclassifies a subset of exact-code disagreements, so the lift is a change of measurement category, not a between-group effect.

Table 1
Initial-visit telemetry. Shares use 923 visits with both codes.

Metric	Count	Share
Exact ICD-10 string match	616	66.7%
Different code, same protocol family	59	6.4%
Different code, different mapped family	98	10.6%
Different, at least one code unmapped	150	16.3%
Routing-aware concordance	675	73.1%

Because physicians saw the AI output before coding, these figures measure operational concordance rather than independent diagnostic accuracy. For example, a physician may record acute sinusitis while the AI first-pass hypothesis is an unspecified upper-respiratory infection. At the exact-code level this is a disagreement; at the workflow level both route to the same upper-airway protocol. This is not an attempt to lower the bar for diagnosis. Rather, it separates two tasks that are often conflated: early semantic routing to a safe institutional pathway, and later definitive coding after tests, medication response, reevaluation, and admission/discharge decisions. Cases outside the curated surface are not forced into a protocol; they are delegated to the broader differential-diagnosis engine.

4. Production Lessons for Semantic Clinical AI

Lesson 1: Exact coding is the wrong first metric when the decision is routing. Clinical coding and clinical action are related but not identical. Early emergency-care support needs a route to the right protocol family. Semantic technologies make that abstraction explicit: related CIDs can be grouped into audited protocol families, and the route can be traced from presentation to disease family to care actions. This is a more useful first-assessment metric than treating every ICD-10 difference as an error.

Lesson 2: Relevance requires prevalence priors, not only symptom matching. A graph that ranks only by symptom overlap can surface rare diseases that are semantically similar but clinically implausible in the local ER population. The deployed scorer therefore includes disease prevalence priors and demographic modifiers. Priors are curated from epidemiological sources relevant to the Brazilian setting where possible and combined with symptom evidence in a Bayesian score. This is not just a modeling detail; it is a trust requirement. Physicians quickly discount a CDSS that repeatedly returns rare but irrelevant hypotheses.

Lesson 3: Knowledge-graph incompleteness requires multi-evidence safeguards. In validation, a rheumatic-fever case exposed an incomplete disease–symptom profile: semantic evidence and the MONDO definition supported the candidate while one association source did not. This motivated the deployed heuristic composition of semantic retrieval and bidirectional KG evidence, with ontology definitions and hierarchy information used to expose conflicts rather than silently treat the graph as complete. Production telemetry evaluates the resulting end-to-end routing behavior; the contribution and optimal calibration of individual signals have not been established through ablation and remain open research questions.

5. Positioning and Conclusion

EHR-native engines such as Tasy [10] execute a protocol once the pathway is known. Point-of-care references such as Afya Whitebook [11] advise without routing into local order sets. Neither publicly performs inspectable semantic routing from an uncertain first-assessment presentation into an institutional protocol family. That is the gap this deployment fills.

The routing architecture—presentation to hypothesis to protocol family—is transferable to other emergency departments. Protocol mappings, order sets, formulary, and prevalence priors are local and must be recoded. Later diagnostic reassessment is a different task: it requires a richer longitudinal

knowledge graph of laboratory, imaging, and treatment-response evidence. Semantic technologies here are governance infrastructure: they make black-box hypotheses actionable and auditable at the first decision point.

Author Contributions

Roberto Shimizu led the conceptualization, methodology, software development, production integration, analysis, and writing. Thiago Porto and Roberta Garcia contributed to the investigation and validation by comparatively evaluating clinical named-entity recognition and biomedical embedding models used to inform model selection. All authors reviewed and approved the final manuscript.

Declaration on Generative AI

During manuscript preparation, the authors used GPT-5.6-sol through the pi coding-agent interface for **Improve writing style** (copy-editing) and **Formatting assistance** (LaTeX formatting), using the contribution categories of the CEUR-WS GenAI Usage Taxonomy. The same tool also assisted with compiling and presenting the aggregate deployment telemetry reported in Section 3. Generative AI was not used to conduct the research, nor to derive or interpret the reported findings. The authors reviewed and edited all outputs and take full responsibility for the manuscript.

References

[1] Agência Nacional de Vigilância Sanitária (ANVISA), RDC No. 657/2022: Regulation of Software as a Medical Device, 2022. URL: https://www.gov.br/anvisa/pt-br/assuntos/produtosparasaude/temas-em-destaque/arquivos/2024/rdc-657-2022-en.pdf, accessed 2026-07-03.

[2] International Medical Device Regulators Forum, Software as a Medical Device (SaMD): Clinical Evaluation, 2017. URL: https://www.imdrf.org/documents/software-medical-device-samd-clinical-evaluation, iMDRF/SaMD WG/N41FINAL:2017.

[3] World Health Organization, Ethics and Governance of Artificial Intelligence for Health, 2021. URL: https://www.who.int/publications/i/item/9789240029200, accessed 2026-07-03.

[4] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence, 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj, artificial Intelligence Act.

[5] S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower, et al., The Human Phenotype Ontology in 2021, Nucleic Acids Research 49 (2021) D1207–D1217. doi:10.1093/nar/gkaa1043.

[6] N. A. Vasilevsky, N. Matentzoglu, S. Toro, J. E. Flack, H. Hegde, D. Unni, G. Alyea, J. S. Amberger, L. Babb, J. P. Balhoff, et al., Mondo: Unifying diseases for the world, by the world, medRxiv (2022). doi:10.1101/2022.04.13.22273750.

[7] D. R. Unni, S. A. T. Moxon, M. Bada, M. Brush, R. Bruskiewich, J. H. Caufield, N. A. Vasilevsky, U. B. K. G. W. Group, M. A. Haendel, P. N. Robinson, C. J. Mungall, Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science, Clinical and Translational Science 15 (2022) 1848–1855. doi:10.1111/cts.13302.

[8] S. Harris, A. Seaborne, SPARQL 1.1 Query Language, 2013. URL: https://www.w3.org/TR/sparql11-query/, w3C Recommendation.

[9] Apache Software Foundation, Apache Jena Fuseki, 2026. URL: https://jena.apache.org/documentation/fuseki2/.

[10] Bionexo, Tasy: Sistema Integrado de Gestão em Saúde, 2026. URL: https://bionexo.com/tasy/.

[11] Afya, Whitebook Clinical Decision, 2026. URL: https://whitebook.pebmed.com.br/planos.