Encoding Guideline Disagreement as First-Class Knowledge: A Neo4j CDSS for Multi-Authority Reconciliation and Dual-Layer Disease Inference

Chenyang Ding[1], Ahmad Taha[1] and João Ponciano[1]

[1]James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, Scotland, United Kingdom

Abstract

Clinical decision support systems (CDSS) that follow the threshold of a single guideline authority cannot show where authorities disagree or detect subclinical disease, and the explanations they give rarely trace back to citable evidence. We present a CDSS built on a knowledge graph that treats these as problems of knowledge engineering and handles all three in one architecture. Each laboratory result is classified against nine independent guideline authorities, every vote is kept as a separate element of the schema, and the disagreement between authorities is reported for each patient as a 95% confidence interval. Disease inference runs in two layers: an overt layer, anchored on CRITICAL classifications, for confirmed disease, and a hidden layer, anchored on ATTENTION classifications, for subclinical states, kept apart by a formally proven exclusion clause. The system is a Neo4j property graph in which the evidence behind every output can be recovered in three traversal steps, so an explanation is read from the graph rather than rebuilt afterwards. On 54,467 classified results from 3,869 patients, the overt layer did not change with the choice of authorities (Cohen's $\kappa = 1.000$); the hidden layer found 130 patients (3.4%) with subclinical conditions that a single threshold misses; and widening the panel from one to nine authorities narrowed the mean confidence interval by 34.2% with no change to any classification.

Keywords

knowledge graph, clinical decision support, guideline reconciliation, symbolic inference, provenance, explainability

1. Introduction and Motivation

A well-built CDSS can change clinical practice [1, 2], yet routine adoption stays low [3, 4]. Three problems of knowledge representation sit behind this gap. The first is interpretability. Regulation such as the EU AI Act and the FDA guidance on AI/ML software as a medical device [5] asks for outputs that can be audited, but a 2025 meta-analysis found that only 17.7% of explainable AI CDSS formally tested how faithful their explanations are [6], and the two methods usually applied after the fact, SHAP [7] and LIME [8], give approximations that cannot be traced to citable clinical evidence. The second is threshold reconciliation. Authorities publish different decision limits for the same test [9], so when a system follows one authority the chosen threshold is hidden inside the choice of authority, with no way to measure the disagreement. The third is subclinical inference. Conditions such as subclinical hypothyroidism [10] and prediabetes [11] sit in states that a single threshold cannot represent: set it at the treatment level and it hides subclinical patients, lower it and it labels patients who should not carry a diagnosis.

We show that one knowledge graph handles all three and gives a reusable component of computable knowledge for learning health systems. This poster reports the architecture and the main findings; the full evaluation is under review elsewhere.

The 25th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2026) - Satellite Events (Workshops, Tutorials, Demos, Posters)

2612869d@student.gla.ac.uk (C. Ding); ahmad.taha@glasgow.ac.uk (A. Taha); joao.ponciano@glasgow.ac.uk (J. Ponciano)
0009-0001-1753-0548 (C. Ding); 0000-0003-1246-8981 (A. Taha)



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

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-26

CEUR-WS.org/Vol-4340/paper1.pdf

2. Related Work

Guideline formalisms such as GLIF [12], Asbru [13], and SAGE [14] model how a single guideline is executed. Knowledge graphs have been used in CDSS for explainable decision support [15], hypertension medication [16], and gastric cancer treatment [17]. None of these run several independent guidelines over the same record at once and report the disagreement between them as a quantity that can be computed for each patient. Our work does this, and it keeps every authority's vote and the resulting evidence inside the graph, so an explanation is read from the graph rather than rebuilt afterwards. A conventional CDSS collapses all of this into one authority and one alerting tier. That configuration is baseline B3 in Section 4, and it is what we measure against.

3. Knowledge Graph Architecture

Each patient record is a set of (test, value) pairs, processed in two stages (Figure 1).

In the first stage every authority that covers a test casts a vote of NORMAL, ATTENTION, or CRITICAL, and each vote is kept as a separate element of the schema rather than collapsed into one label. A conservative rule governs escalation: a CRITICAL consensus needs support from at least two authorities, or a split between CRITICAL and ATTENTION. The consensus confidence, together with a 95% confidence interval for each patient, records how far the authorities disagree.

The second stage maps those classified results to disease risk through two rule sets: an overt layer (anchored on CRITICAL, eight rules) for confirmed disease, and a hidden layer (anchored on ATTENTION, seven rules) for subclinical states. Each rule carries an exclusion specification with two parts. An *overt counterpart* suppresses the matching hidden rule whenever it fires; a set of exclusion test conditions suppresses it regardless. The layers stay apart by construction, not by convention.

Theorem 1 (Non-overlapping guarantee). *Let r_H be a hidden rule whose exclusion specification lists an overt counterpart $r_O \in O(r_H)$ inferring the same disease label ρ. Then for any patient P,* $\neg\big(\mathrm{Overt}(P,\rho) \wedge \mathrm{Hidden}(P,\rho)\big)$.

Proof. Suppose both hold. From $\mathrm{Overt}(P,\rho)$ the rule r_O fires, so $\mathrm{match}(r_O,P) = 1$ and $\mathrm{excluded}_O(r_O,P) = 0$. Since $r_O \in O(r_H)$, the overt-counterpart clause sets $\mathrm{excluded}_H(r_H,P) = 1$. But $\mathrm{Hidden}(P,\rho)$ requires $\mathrm{excluded}_H(r_H,P) = 0$ for r_H to fire, a contradiction. □

The condition holds for all fifteen rules of the implemented system, so the labels inferred for a patient partition into an overt-only and a hidden-only subset. Because this stage works on the classifications rather than on raw values, disagreement is reported openly instead of being carried downstream unnoticed.

Both stages run on a Neo4j property graph (Figure 2). Each disease risk node links to the classified result that triggered it, and each classified result links to the guideline authority nodes that voted on it; every authority node stores an evidence grade and a DOI. The full chain, from patient through test and voting authorities to the guideline publication, is recovered in three traversal steps, so an explanation is part of the graph rather than a later addition.

4. Preliminary Results

We evaluated the framework on 54,467 results from 3,869 de-identified patients at a secondary care referral laboratory (Synlab, Belgium), using an authority ablation with four configurations (from one authority to nine) and three baselines.

The baselines isolate what each design element contributes. B1 keeps both layers but draws every threshold from NICE alone. B2 replaces the guideline thresholds with the source laboratory's own reference ranges; because those carry no CRITICAL tier it fires no overt rule at all, which is the degenerate case we expected. The baseline that matters is B3, a conventional single-threshold system:

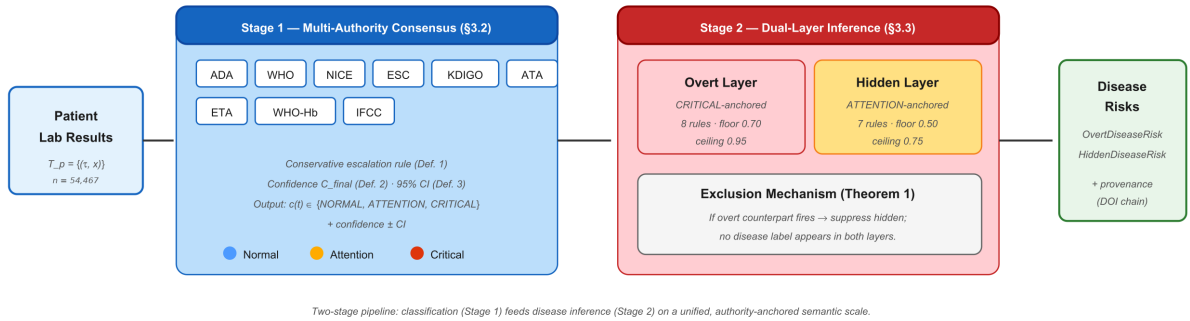


Figure 1: End-to-end pipeline. Patient laboratory results pass through Stage 1, a consensus over nine guideline authorities that returns a classification with a 95% confidence interval, and then through Stage 2, the inference over two layers that produces overt and hidden disease risk. The exclusion clause (Theorem 1) keeps any disease out of both layers.

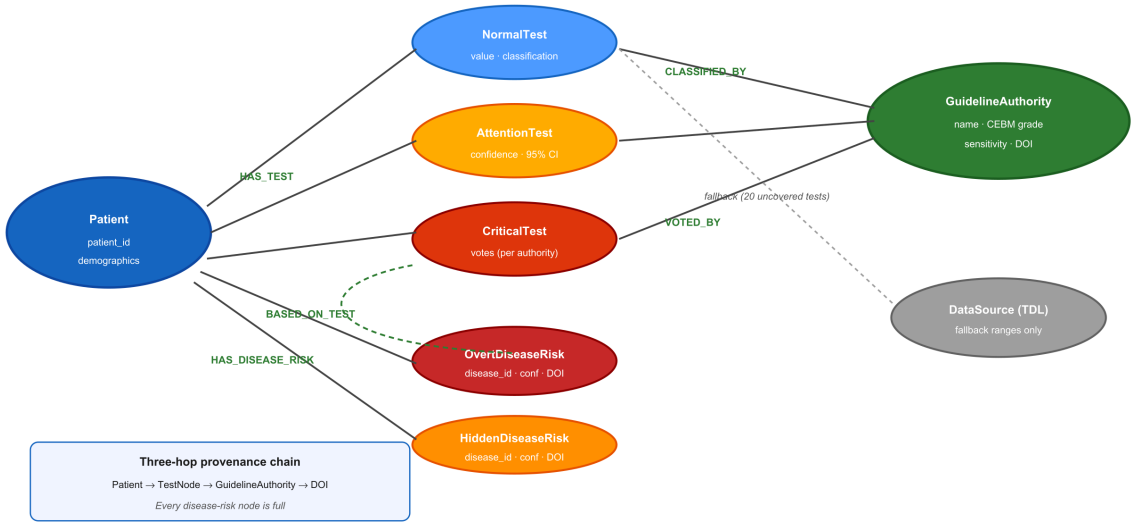


Figure 2: Neo4j property graph schema. Patient, test classification, guideline authority, and the two disease risk layers are linked so that every disease risk node exposes a three-step chain from the patient to the guideline publication. Tests that no authority covers fall back to the laboratory's own reference ranges.

NICE thresholds, overt layer only. Against B1 the full nine-authority system returns an overt cohort of the same size (3,226 against 3,228 patients, both 83.4%), so the added authorities do not inflate the confirmed-disease count. The separation shows up against B3 instead. There, 130 patients (3.4%) invisible to the baseline carry hidden-layer risk under the full system; 2 more (0.1%) are recovered in the overt layer, and 513 (13.3%) carry no detected risk under either configuration.

Two divergences changed classifications. For C-reactive protein ESC sets a stricter threshold than NICE, 3 mg/L against 10 mg/L [18], which moved 499 results from ATTENTION to CRITICAL. For thyroid-stimulating hormone ATA and ETA set 2.5 mIU/L against the NICE value of 4.0 mIU/L [19], which moved 223 results from NORMAL to ATTENTION (Figure 3). A third divergence changed nothing at all: KDIGO's creatinine threshold reclassified no result in this cohort, because every value in the range where the authorities differ was already CRITICAL under the lower bound. Where the panel does disagree, the disagreement is reported for each patient as a 95% confidence interval rather than hidden in the choice of authority.

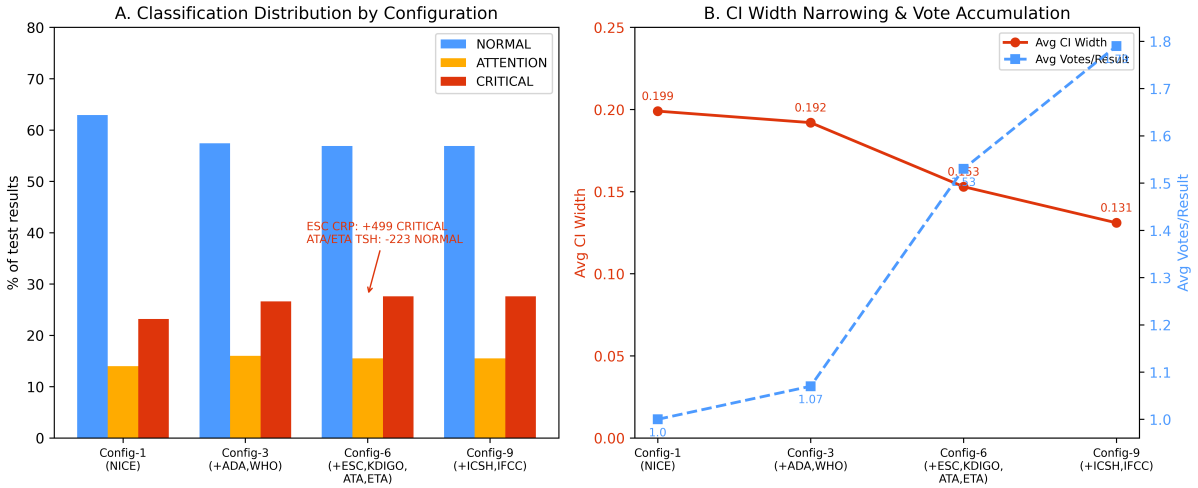


Figure 3: Authority ablation across four configurations. (A) Distribution of NORMAL, ATTENTION, and CRITICAL classifications; two divergences change classifications between the three-authority and six-authority configurations. (B) The mean width of the 95% confidence interval falls from 0.199 to 0.131 as the average number of votes per result rises from 1.00 to 1.79.

Table 1
Subclinical findings surfaced by the two-layer architecture. The counts by condition add up to more than 130 because some patients carry more than one subclinical condition.

Subclinical condition	**Patients**	**Guideline / protocol**
Subclinical hypothyroidism	72	ETA 2013 [19]
Early CKD risk	110	KDIGO 2024 [20]
Prediabetes	35	ADA 2025 [21]
Hidden only	**130 (3.4%)**	missed by a single threshold

Across every configuration of authorities the overt classifications came out identical (Cohen's $\kappa = 1.000$ for all pairwise comparisons); 3,228 patients (83.4%) carried at least one confirmed disease, most often chronic kidney disease (74.4%) and liver disease (66.3%), in line with a secondary care referral cohort.

What the added authorities did change was precision. Widening the panel from one to nine narrowed the mean 95% confidence interval by 34.2%, from 0.199 to 0.131, with no change to any consensus classification. The choice of authorities therefore affects only the precision of the disagreement signal, not the classification itself.

The hidden layer found 130 patients (3.4%) who trigger no overt rule yet carry active subclinical risk, a group that a single threshold misses, and each finding maps to an established clinical action (Table 1). Because the data carry no reference standard diagnoses, these counts denote results that meet guideline thresholds rather than confirmed cases, and prospective validation is part of ongoing work.

5. Conclusion and Ongoing Work

Consensus across authorities, inference in two layers, and provenance held in the graph work together in one architecture that makes the disagreement between guidelines computable and brings subclinical risk to the surface, with an explanation that is part of the graph rather than a later addition. Ongoing work widens the set of authorities and adds patient trajectories over time. The step we regard as most urgent is validation in clinical workflow, followed by Delphi calibration of the confidence parameters, which are currently fixed at design time. Code for the consensus classification, the dual-layer inference, and the evaluations reported here is available at https://github.com/ethan-cn/kg-cdss-dual-layer, archived

at https://doi.org/10.5281/zenodo.20630234.

Declaration on Generative AI

During the preparation of this work the authors used Claude (Anthropic) and Gemini (Google) for drafting content, paraphrasing and rewording, improving the writing style, and formatting assistance. The tools were not used in the study design, in the development of the framework or its rules, in the data analysis, or in the generation of any result. All figures report the authors' own experimental results and were not generated by AI. The authors reviewed and edited the content and take full responsibility for it.

References

[1] Kawamoto, K., Houlihan, C.A., Balas, E.A., Lobach, D.F.: Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 330, 765 (2005).

[2] Roshanov, P.S., et al.: Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. BMJ 346, f657 (2013).

[3] Meunier, P.Y., et al.: Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review. Ann. Fam. Med. 21(1), 57–69 (2023).

[4] Derksen, C., et al.: The implementation challenge of computerized CDSS for the detection of disease in primary care: systematic review and recommendations. Implement. Sci. 20, 33 (2025).

[5] U.S. Food and Drug Administration: Artificial Intelligence/Machine Learning-Based Software as a Medical Device Action Plan (2021).

[6] Abbas, Q., Jeong, W., Lee, S.W.: Explainable AI in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges. Healthcare 13(17), 2154 (2025).

[7] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: NeurIPS 30, pp. 4765–4774 (2017).

[8] Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": explaining the predictions of any classifier. In: KDD 2016, pp. 1135–1144 (2016).

[9] Ozarda, Y., Sikaris, K., Streichert, T., Macri, J.: Distinguishing reference intervals and clinical decision limits: a review by the IFCC Committee on Reference Intervals and Decision Limits. Crit. Rev. Clin. Lab. Sci. 55(6), 420–431 (2018).

[10] Urgatz, B., Razvi, S.: Subclinical hypothyroidism, outcomes and management guidelines: a narrative review. Curr. Med. Res. Opin. 39(3), 351–365 (2023).

[11] Tabák, A.G., Herder, C., Rathmann, W., Brunner, E.J., Kivimäki, M.: Prediabetes: a high-risk state for diabetes development. Lancet 379(9833), 2279–2290 (2012).

[12] Ohno-Machado, L., et al.: The guideline interchange format: a model for representing guidelines. J. Am. Med. Inform. Assoc. 5(4), 357–372 (1998).

[13] Shahar, Y., Miksch, S., Johnson, P.: The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. Artif. Intell. Med. 14(1–2), 29–51 (1998).

[14] Tu, S.W., et al.: The SAGE guideline model: achievements and overview. J. Am. Med. Inform. Assoc. 14(5), 589–598 (2007).

[15] Rajabi, E., Etminani, K.: Towards a knowledge graph-based explainable decision support system in healthcare. Stud. Health Technol. Inform. 281, 502–503 (2021).

[16] Zhou, G., et al.: Clinical decision support system for hypertension medication based on knowledge graph. Comput. Methods Programs Biomed. 227, 107220 (2022).

[17] Li, S., et al.: GC-CDSS: personalized gastric cancer treatment recommendations system based on knowledge graph. Int. J. Med. Inform. 185, 105402 (2024).

[18] Visseren, F.L.J., et al.: 2021 ESC guidelines on cardiovascular disease prevention in clinical practice. Eur. Heart J. 42(34), 3227–3337 (2021).
[19] Pearce, S.H.S., et al.: 2013 ETA guideline: management of subclinical hypothyroidism. Eur. Thyroid J. 2(4), 215–228 (2013).
[20] KDIGO CKD Work Group: KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney Int. 105(4S), S117–S314 (2024).
[21] American Diabetes Association: Diagnosis and classification of diabetes: standards of care in diabetes 2025. Diabetes Care 48(Suppl 1), S27–S49 (2025).