

Verifiable Disease Twins:

Grounded, Auditable, and Machine-Readable Rare-Disease Knowledge at Population Scale

Dimas Timmers Raras-AI Team
Raras Health, São Paulo, Brazil • raras.org
contato@raras.org

July 2026 • Technical Report v1

Abstract

Consumer-facing rare-disease information is routinely dismissed by clinical societies as “not validated.” We describe the validation architecture behind *Disease Twin*, a production system that maintains a living, source-grounded Portuguese-language page for each of 10,468 rare diseases catalogued in Orphanet. Our central design commitment is that **every sentence a patient reads is independently verifiable**. Content is authored by an autonomous agent that may cite *only* a curated allowlist of canonical biomedical and regulatory sources; a four-stage grounding gate drops any ungrounded claim before publication, and per-page grounding is thresholded (hard-block below 60%, review between 60–75%, publish above 75%). We then contribute a **claim-level provenance layer**: each verified paragraph is exported as a standards-compliant *nanopublication* (W3C PROV-O + the Nanopublication schema), a self-contained four-named-graph RDF unit binding the assertion to its evidence sources, its generating agent, its grounding flag, and its human-review status. The provenance graph is content-negotiable and dereferenceable per disease at `/api/provenance/{orphaCode}` in TriG, JSON-LD, and N-Quads. We integrate a human-in-the-loop review layer in which medical geneticists co-sign entries as curators, with the co-signature carried into the machine-readable provenance. Finally, we argue that rareness is *jurisdictional*, not absolute, and describe how endemic-in-Brazil conditions (e.g., tuberculosis) are explicitly flagged with primary sources — turning the “rare only in Portuguese” objection into an information-equity mandate for an under-served language. The system is open (CC0 on the disease layer), auditable, and addressable by any AI assistant via the Model Context Protocol.

1 Introduction

The migration of biomedical knowledge from static artefacts (textbooks, curated databases) to *agentic* substrates that read, write, and revise themselves raises an old question in a sharper form: *who validates the claims?* For rare diseases the stakes are acute. The long tail of ~7,000 conditions is individually small, collectively enormous, and structurally under-resourced; content about it is sparse, and in languages other than English it is close to absent. When a patient-facing platform fills that void with machine-assisted text, the natural and correct reaction of a clinical society — in our case the Brazilian Society of Medical Genetics — is to ask for evidence of validation, and to point out edge cases (e.g., “tuberculosis is not rare in Brazil”).

We take both objections seriously and answer them structurally rather than rhetorically. This report describes the validation architecture of *Disease Twin*, deployed in production over the full Orphanet catalogue. Our contributions are:

1. **A “never invent” authoring pipeline (§3.2)**: a canonical-source allowlist plus a four-stage grounding gate that makes unsourced clinical text unpublishable by construction, with per-page grounding thresholds.
2. **Claim-level nanopublications (§4)**: every verified sentence exported as a four-named-graph RDF unit un-

der W3C PROV-O and the Nanopublication schema, dereferenceable per disease — the machine-readable counterpart of the on-page citations.

3. **Human-in-the-loop co-signature (§5)**: specialist curators whose review status propagates into the provenance graph, so the validation seal is auditable rather than declarative.
4. **Jurisdictional rareness (§6)**: explicit, sourced flagging of endemic-in-Brazil conditions, reframing a lexical objection as an equity mandate for Portuguese-language health information.

2 Background and Related Work

Editorial validation of open medical content. Wikipedia’s WikiProject Medicine established that credibility at scale comes not from authorship credentials alone but from *sourcing discipline* (the MEDRS guideline requiring reliable secondary sources) plus expert review, exemplified by its Cochrane partnership [1, 2]. Our grounding gate is a machine-enforced analogue of MEDRS, and our curator layer is the expert-review analogue.

Rare-disease ontologies and provenance. Orphanet, OMIM, HPO [3], and the MONDO disease ontology integrated by the Monarch Initiative [4] provide the identifier and cross-reference backbone. We anchor every entity to these authorities and never mint disease concepts.

Grounded and fact-checked medical generation. FActScore introduced atomic-fact attribution as an evaluation primitive [5]; recent biomedical systems such as MedRAGChecker perform claim-level verification of retrieval-augmented generation via decomposition and entailment scoring [6], and citation-hallucination detectors such as CiteCheck target fabricated references [7]. Knowledge-graph agents (KGAREvion [8], MedKGent [9]) verify assertions against a graph before emission. Our pipeline operationalises this line *in production* over an entire disease catalogue rather than on a benchmark.

Nanopublications. Nanopublications [10] and their decentralized, provenance-aware publishing model [11] give a minimal, self-contained RDF unit — an assertion plus its provenance and publication metadata — ideal for making individual clinical claims first-class, citable, and composable. We adopt them as our claim-level export format.

3 System Architecture

Disease Twin has three layers, each governed by one non-negotiable principle.

3.1 Layer 1 — Clinical knowledge graph

A Neo4j property graph populated from canonical open sources (Orphanet, HPO, MONDO, OMIM, ClinVar, Open Targets, PubTator3, ClinicalTrials.gov, and the Brazilian public-health system: DATASUS, CEAf, SUS clinical protocols). The schema follows the Biolink Model with a KGX-exportable dump. *Principle: no node is invented* — every entity carries a stable CURIE and provenance to its originating source. The public disease layer is CC0; individual patient data never leaves the patient’s wallet, separated by a k -anonymity boundary.

3.2 Layer 2 — The autonomous page author

An agent authors and *owns* each disease’s Portuguese page as a persistent, diffable document (a “LLM Wiki”), re-invoked only when the underlying structured facts change — a fingerprint over genes, phenotypes, medications, SUS coverage, and source URLs gates re-authoring, so the narrative does not drift for no reason. The model receives *only* facts already in the graph plus their citable source URLs, and must attach `grounded:true` and `fontes:[n]` to every paragraph. Four sequential gates protect publication:

1. **Canonical-domain filter** — sources whose host is not on an explicit allowlist (Orphanet, HPO, OMIM, MONDO, GenCC, HGNC, Open Targets, ChEMBL, PubTator3, ClinicalTrials.gov, FDA, EMA, ANVISA, CONITEC, DOU `in.gov.br`, `bvsms.saude.gov.br`, DATASUS, WHO, ...) are dropped *before* the model can cite them.
2. **Mandatory per-paragraph citation** — a paragraph without ≥ 1 source is discarded.

3. **Cited-index validation** — citations must resolve to a surviving canonical source; dangling indices are removed.
4. **Coverage decision** — per-page grounding % is recomputed from what survives; below 60% the page is hard-blocked, 60–75% publishes with a *review-pending* flag, above 75% publishes.

Every LLM-mediated write to the graph carries `createdBy` and is reversible in a single query. LLM-extracted literature relations additionally require that *both* endpoints already exist as anchored entities (gene by HGNC symbol, drug by ChEMBL name, disease by Orphanet code); unanchored triples are dropped.

3.3 Layer 3 — Public, machine-addressable surface

Disease pages render the authored content with inline `[n]` citations; a groundedness scorecard (`/api/groundedness`) reports, catalogue-wide, the share of disease nodes carrying authority cross-references, literature support, and per-page grounding. The knowledge graph is exposed as Linked Data (Turtle/JSON-LD/N-Triples via `/api/rdf`), with a VOID dataset description, a SPARQL endpoint, and dereferenceable RARAS identifiers. Any AI assistant can query a disease directly via the Model Context Protocol.

4 Claim-Level Provenance as Nanopublications

On-page `[n]` citations answer “is this sourced?” for a human reader. To answer it for a *machine* — and to let a skeptical third party audit without trusting our UI — we export each verified paragraph as a nanopublication. A nanopublication is four RDF named graphs (Figure 1):

- **head** — links the nanopublication to its three sub-graphs;
- **assertion** — the claim: its text (`rdf:value`, language-tagged), the disease it is about (Orphanet/MONDO IRIs), and the page section;
- **provenance** — `prov:wasDerivedFrom` each evidence source, `prov:wasGeneratedBy` an authoring `prov:Activity` associated with the Disease Twin `prov:SoftwareAgent`, a grounding flag, and the review status (with `prov:wasAttributedTo` the reviewing specialist when co-signed);
- **pubinfo** — generation time, attribution to Raras, CC0 license, page grounding %, and dataset membership.

The document is content-negotiable and dereferenceable per disease at `/api/provenance/{orphaCode}` (accepting a RARAS D-prefixed id or a bare Orphanet code) in TriG (native named graphs), N-Quads, and named-graph JSON-LD. Because provenance travels *with* the assertion in a

```

<.../np/D36899/c1> {
  <.../c1> a np:Nanopublication;
  np:hasAssertion <.../c1#assertion>;
  np:hasProvenance <.../c1#provenance>;
  np:hasPublicationInfo <.../c1#pubinfo> }

<.../c1#assertion> {
  <.../c1#claim> a rnc:Claim, prov:Entity;
  rnp:aboutDisease orphanet:36899, mondo:0000903;
  rdf:value "A Síndrome de Mioclonia-Distonia
    ...espasmos mioclonicos rapidos."@pt;
  rnp:section "Visao geral"@pt }

<.../c1#provenance> {
  <.../c1#claim>
    prov:wasDerivedFrom <orpha.net/...36899>;
    prov:wasGeneratedBy <.../c1#authoring>;
    rnp:isGrounded true;
    rnp:reviewStatus "expert-reviewed";
    prov:wasAttributedTo <.../especialista/...>.
  <.../c1#authoring> a prov:Activity,
    rnc:DiseaseTwinAuthoring;
    prov:wasAssociatedWith <.../agent/disease-twin>;
    prov:used <orpha.net/...36899> }

<.../c1#pubinfo> {
  <.../c1> prov:generatedAtTime "2026-..."^^xsd:dateTime;
  dcterms:license <.../publicdomain/zero/1.0/>;
  rnp:pageGroundingPct 100 }

```

Figure 1: One production nanopublication (TriG, abridged) for a paragraph of the Myoclonus–Dystonia page (ORPHA:36899). Four named graphs bind the claim to its source, its generating agent, and its review status.

standard vocabulary, third parties can ingest, requery, and cross-check our claims without our cooperation — the concrete, non-repudiable form of “validated.”

5 Human-in-the-Loop Co-Signature

Autonomy is not the terminal state. Pages whose grounding falls in the 60–75% band are published with a visible *review-pending* chip and prioritised in a review queue. Medical geneticists — drawn from a directory of ~2,000 opt-in specialists sourced from the national genetics society — co-sign entries as *curators*. Crucially, the co-signature is not a cosmetic badge: it is written into the nanopublication provenance as `prov:wasAttributedTo` the reviewer, and the page-level `rnp:reviewStatus` transitions from *autonomous-grounded* to *expert-reviewed*. This mirrors the Wikipedia–Cochrane strategy of turning domain experts from critics into contributors, but with the review state made machine-auditable. Adopting a critic’s own community as the validating authority is, we argue, the durable answer to “who validated this.”

6 Rareness Is Jurisdictional

A disease is rare *in a population*, not in the abstract; prevalence thresholds are set per jurisdiction (EU <1:2,000; Brazil’s PNGDR <1:1,300). Tuberculosis appears in Orphanet yet is endemic in Brazil, so presenting it as “rare” locally would be wrong. We therefore maintain an explicit,

curated flag (`notRareInBR`) over ~77 conditions that are endemic or common in Brazil, each carrying a category and a *primary source* (e.g., the Compulsory Notification ordinance, Ministry of Health epidemiological bulletins, INCA cancer estimates); the UI substitutes an “endemic in Brazil” badge for the rare-disease pill, and the most sensitive cases are hidden from the rare-disease catalogue entirely. Far from inflating rareness, the system actively deflates it where the local epidemiology demands.

This reframes the “rare only in Portuguese” objection. The deficit is real: local epidemiology of endemic conditions (dengue, Chagas, tuberculosis) is largely absent from English-language resources [12], and non-English Wikipedia editions are systematically under-served. Closing that gap for ~260 million Portuguese speakers is not a bug to apologise for — it is the mandate.

7 Deployment and Evaluation

The system runs in production over the full Orphanet catalogue (10,468 diseases), with a daily agentic loop on managed cloud infrastructure. Validation is reported, not asserted: the public groundedness scorecard exposes catalogue-wide coverage (authority cross-references, literature support, per-page grounding), and the transparency dashboard shows, per disease, what was verified or rejected in a rolling window. Page-level grounding is enforced at the thresholds of §3.2; only the surviving, cited paragraphs are published and exported as nanopublications. We regard these figures as conservative, self-reported transparency artefacts — structural groundedness — not external benchmark scores, and we publish the methodology alongside the numbers.

8 Discussion, Limitations, and Ethics

Grounding guarantees *traceability to a canonical source*, not clinical correctness of that source or of the synthesis; this is why the review layer and the “informative, not a substitute for medical evaluation” disclaimer are load-bearing. Coverage of the literature-mining agents is uneven across the long tail. Structural groundedness is not a substitute for prospective clinical evaluation. Hypotheses (drug-repurposing, Swanson ABC) are always labelled as candidates for investigation and never mixed with verified evidence in any API route. All content is CC0 on the disease layer; individual patient data never leaves the wallet. The architecture’s political premise is that a knowledge commons for neglected diseases must be governed by the people who live the conditions — hence public transparency, patient-founder governance, and association-only community creation.

9 Conclusion

“Your information isn’t validated” is answerable by construction. Make unsourced clinical text unpublishable; attach each surviving claim’s evidence, generating agent, and review status to the claim itself in a standard, dereferenceable form; and invite the critics’ own professional community to co-sign. Validation then stops being a promise and becomes an artefact anyone — human or machine — can audit. We offer this architecture, in production over an entire rare-disease catalogue and open by default, as a replicable recipe for any knowledge commons the market has chosen to forget.

References

- [1] Cochrane Community. *Improving health-related content on Wikipedia*. community.cochrane.org.
- [2] Wikipedia. *Identifying reliable sources (medicine) (MEDRS)*. en.wikipedia.org.
- [3] Gargano, M., Matentzoglou, N., Köhler, S., et al. (2024). The Human Phenotype Ontology in 2024. *Nucleic Acids Research* 52(D1):D1333.
- [4] Putman, T., Matentzoglou, N., et al. (2024). The Monarch Initiative in 2024. *Nucleic Acids Research* 52(D1):D938.
- [5] Min, S., Krishna, K., et al. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision. *EMNLP*. arXiv:2305.14251.
- [6] MedRAGChecker: Claim-Level Verification for Biomedical Retrieval-Augmented Generation (2026). arXiv:2601.06519.
- [7] CiteCheck: Retrieval-Grounded Detection of LLM Citation Hallucinations in Scientific Text (2026). arXiv:2605.27700.
- [8] Su, X., Zitnik, M., et al. (2024). KGAREvion: An AI Agent for Knowledge-Intensive Biomedical QA. *NeurIPS*. arXiv:2410.04660.
- [9] MedKGent: An LLM Agent Framework for Temporally Evolving Medical Knowledge Graphs (2025). arXiv:2508.12393.
- [10] Groth, P., Gibson, A., Velterop, J. (2010). The anatomy of a nanopublication. *Information Services & Use* 30(1–2):51–56.
- [11] Kuhn, T., et al. (2016). Decentralized provenance-aware publishing with nanopublications. *PeerJ Computer Science* 2:e78.
- [12] MedPT: A Massive Medical Question Answering Dataset for Brazilian-Portuguese Speakers (2025). arXiv:2511.11878.