Journal of Health Informatics

J Health Inform. 2026;18:e1765
DOI: https://doi.org/10.59681/2175-4411.v18.2026.1765

ORIGINAL ARTICLE

Design and evaluation of an ontology-based knowledge graph integrating births, climate, and socioeconomic data in Brazil

Desenho e avaliação de grafo de conhecimento baseado em ontologia integrando nascimentos, clima e dados socioeconômicos no Brasil

Diseño y evaluación de grafo de conocimiento basado en ontología que integra nacimientos, clima y datos socioeconómicos en Brasil

Guilherme Coelho[1], Danielle Satie Kassada[1], Charles M'poca Charles[1], Rodolfo de Carvalho Pacagnella[1], Clarimar José Coelho[2] ✉

ORCID: 0000-0003-3045-4057 · 0000-0002-6960-6444 · 0000-0001-7133-0574 · 0000-0002-5739-0009 · 0000-0002-5163-2986

[1] Universidade Estadual de Campinas (UNICAMP), Campinas (SP), Brazil.

[2] Pontifícia Universidade Católica de Goiás (PUC Goiás), Goiânia (GO), Brazil.

ARTICLE INFORMATION

TYPE
Original Article

IDENTIFIER
e1765

DOI
10.59681/2175-4411.v18.2026.1765

RECEIVED
15 September 2026

ACCEPTED
23 November 2026

REVIEW PROCESS
Double Blind Review

HANDLING EDITOR
Juliano de Souza Gaspar

ACCESS
Open, CC BY 4.0

CORRESPONDING AUTHOR
Clarimar José Coelho
clarimarc@gmail.com

LICENSE

This is an open access article published under the Creative Commons CC BY 4.0 license, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited.

CC BY 4.0

ABSTRACT

Objective: To design and evaluate a reproducible, ontology-based knowledge graph that integrates Brazilian live birth records with climate and socioeconomic data. **Method:** We specified an ontology in Web Ontology Language 2, aligned to the Basic Formal Ontology and mapped to international clinical terminologies, and a parameterized extract-transform-load pipeline, deployed as a proof of concept for Goiás, Brazil (2015–2024). We evaluated logical consistency through automated reasoning, structural conformance through shape constraints, the recall of record linkage between the national live birth (SINASC) and mortality (SIM) systems, and query performance. **Results:** The ontology comprised 34 classes and 36 terminology mappings. The pipeline integrated 934,806 births into 13,334,157 triples, with no unsatisfiable classes or constraint violations. Record linkage recovered 85.2% of the 667 neonatal deaths of 2020. Median query times ranged from 0.026 to 2.263 s. As a demonstration of query capability, a single query returned municipal counts of preterm births to mothers exposed to extreme heat in the third trimester for 245 of 246 municipalities in 0.430 s. **Conclusion:** The architecture delivers verifiable consistency and declarative data-quality checks at low cost, and its jurisdiction-independent design supports national scale-up as reusable infrastructure for climate-aware perinatal research.

Keywords: Biological Ontologies; Semantic Web; Vital Statistics.

RESUMO

Objetivo: Desenhar e avaliar um grafo de conhecimento reprodutível, baseado em ontologia, integrando nascidos vivos brasileiros a dados climáticos e socioeconômicos. **Método:** Especificamos uma ontologia em Web Ontology Language 2, alinhada à Basic Formal Ontology e mapeada a terminologias clínicas internacionais, e um processo parametrizado de extração, transformação e carga, implantado como prova de conceito em Goiás, Brasil (2015–2024). Avaliamos a consistência lógica por raciocínio automatizado, a conformidade estrutural por restrições de forma, a sensibilidade do relacionamento entre os sistemas nacionais de nascimento (SINASC) e óbito (SIM) e o desempenho das consultas. **Resultados:** A ontologia reuniu 34 classes e 36 mapeamentos terminológicos. Foram integrados 934.806 nascimentos em 13.334.157 triplas, sem classes insatisfatíveis nem violações de restrições. O relacionamento recuperou 85,2% dos 667 óbitos neonatais de 2020. As medianas das consultas variaram de 0,026 a 2,263 s. Como demonstração de capacidade de consulta, uma consulta devolveu contagens municipais de nascimentos prematuros sob calor extremo no terceiro trimestre em 245 dos 246 municípios em 0,430 s. **Conclusão:** A arquitetura oferece consistência verificável e checagem declarativa da qualidade dos dados a baixo custo, e seu desenho independente de jurisdição favorece a expansão nacional como infraestrutura reutilizável para pesquisa perinatal sensível ao clima.

Descritores: Ontologias Biológicas; Web Semântica; Estatísticas Vitais.

RESUMEN

Objetivo: Diseñar y evaluar un grafo de conocimiento reproducible, basado en ontología, integrando nacidos vivos brasileños con datos climáticos y socioeconómicos. **Método:** Especificamos una ontología en Web Ontology Language 2, alineada con la Basic Formal Ontology y mapeada a terminologías clínicas internacionales, y un proceso parametrizado de extracción, transformación y carga, implementado como prueba de concepto en Goiás, Brasil (2015–2024). Evaluamos la consistencia lógica mediante razonamiento automatizado, la conformidad estructural mediante restricciones de forma, la sensibilidad del enlace entre los sistemas nacionales de nacimiento (SINASC) y defunción (SIM) y el rendimiento de las consultas. **Resultados:** La ontología incluyó 34 clases y 36 mapeos terminológicos. Se integraron 934.806 nacimientos en 13.334.157 tripletas, sin clases insatisfacibles ni violaciones de restricciones. El enlace recuperó 85,2% de las 667 muertes neonatales de 2020. Las consultas tuvieron medianas de 0,026 a 2,263 s. Como demostración, una única consulta devolvió recuentos municipales de nacimientos prematuros bajo calor extremo en el tercer trimestre en 245 de 246 municipios en 0,430 s. **Conclusión:** La arquitectura ofrece consistencia verificable y control declarativo de calidad de datos a bajo costo, y su diseño independiente de jurisdicción favorece la expansión nacional como infraestructura reutilizable para investigación perinatal sensible al clima.

Descriptores: Ontologías Biológicas; Web Semántica; Estadísticas Vitales.

HOW TO CITE THIS ARTICLE: Coelho G, Kassada DS, Charles CM, Pacagnella RC, Coelho CJ. Design and evaluation of an ontology-based knowledge graph integrating births, climate, and socioeconomic data in Brazil. J Health Inform 2026;18:e1765. doi: https://doi.org/10.59681/2175-4411.v18.2026.1765

1 INTRODUCTION

Brazil maintains some of the largest population-wide health information systems in the world. The Live Birth Information System (*Sistema de Informações sobre Nascidos Vivos*, SINASC) and the Mortality Information System (*Sistema de Informações sobre Mortalidade*, SIM), both operated by the Ministry of Health, are designed to register every live birth and death in the country; SINASC alone records approximately 2.5 million live births annually, a figure that has declined over the past decade [1]. SINASC has high and increasingly homogeneous national coverage [2], and SIM provides the complementary mortality component. Yet a question of growing public health importance, namely how environmental and social conditions shape perinatal outcomes, cannot be answered by these systems alone, because it requires situating each birth within the climatic and socioeconomic context of its place and time. Systematic reviews consistently associate high ambient temperature during pregnancy with preterm birth, reduced birth weight, and stillbirth [3,4], and station-based gridded climate datasets now make daily temperature available at fine spatial resolution for the whole of Brazil [5].

The limiting factor is therefore less the availability of data than the difficulty of integrating them. SINASC, SIM, gridded climate data, and municipal socioeconomic indicators are distributed in heterogeneous formats, under distinct identifiers, spatial units, and temporal resolutions, and none carries explicit, machine-interpretable semantics. Linking a birth to the climate exposure or human development index of the mother's municipality is usually done through bespoke, study-specific scripts. Because SINASC and SIM share no common key, linking births to deaths likewise relies on record-linkage procedures that each study re-implements [6]; even the 100 Million Brazilian Cohort, which links national systems for tens of millions of individuals, delivers its linked result as a cohort dataset built for analysis rather than as a semantically described resource [7,8]. Such pipelines are expensive to build, hard to reproduce, and seldom reusable, because the meaning of the resulting variables stays implicit in code.

Formal knowledge graphs offer a principled alternative. By describing entities and their relations with an ontology expressed in the Web Ontology Language (OWL 2) [9], data from disparate sources can be brought together under shared, explicit semantics, queried with the standard SPARQL Protocol and Resource Description Framework (RDF) Query Language (SPARQL), and checked for logical consistency by automated reasoners [10]. In biomedicine, knowledge graphs have become a mature instrument for integrating fragmented data [11], with large resources such as PrimeKG for precision medicine [12] and RNA-KG for molecular biology [13]. Interoperability is obtained by aligning domain ontologies to a common upper-level ontology, the Basic Formal Ontology (BFO), following the principles of the Open Biological and Biomedical Ontologies (OBO) Foundry [14,15]. The OBO Foundry library, however, contains no accepted perinatal ontology, and the existing biomedical graphs focus on molecular and clinical-genomic scales rather than on population-level vital statistics combined with environmental and socioeconomic context.

Two technological families dominate current practice: the stack of World Wide Web Consortium standards formed by RDF, OWL 2, and SPARQL, designed for formal semantics, inference, and interoperability; and labeled property graphs, optimized for traversal but without native logical semantics or comparable standardization [10,16]. For a resource whose value depends on semantically consistent integration and alignment with the biomedical ontology ecosystem, the guarantees of the RDF/OWL 2 stack are the more relevant criterion. The semantic web components adopted here are deliberately well established; their maturity favors reproducibility, and the novelty lies in how they are combined and evaluated for a domain not yet served.

This study aimed to design and evaluate PerinatalKG, a formal knowledge graph that integrates SINASC, gridded daily climate data, and municipal socioeconomic indicators under an OWL 2 ontology aligned to BFO and mapped to Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), International Classification of Diseases, 10th revision (ICD-10), and Logical Observation Identifiers Names and Codes (LOINC), with SIM used to validate record linkage. We report it as a proof of concept for the state of Goiás, evaluate it for logical consistency, structural conformance, linkage recall, and query performance, and demonstrate its analytical capability through descriptive queries.

2 METHODS

2.1 Data sources

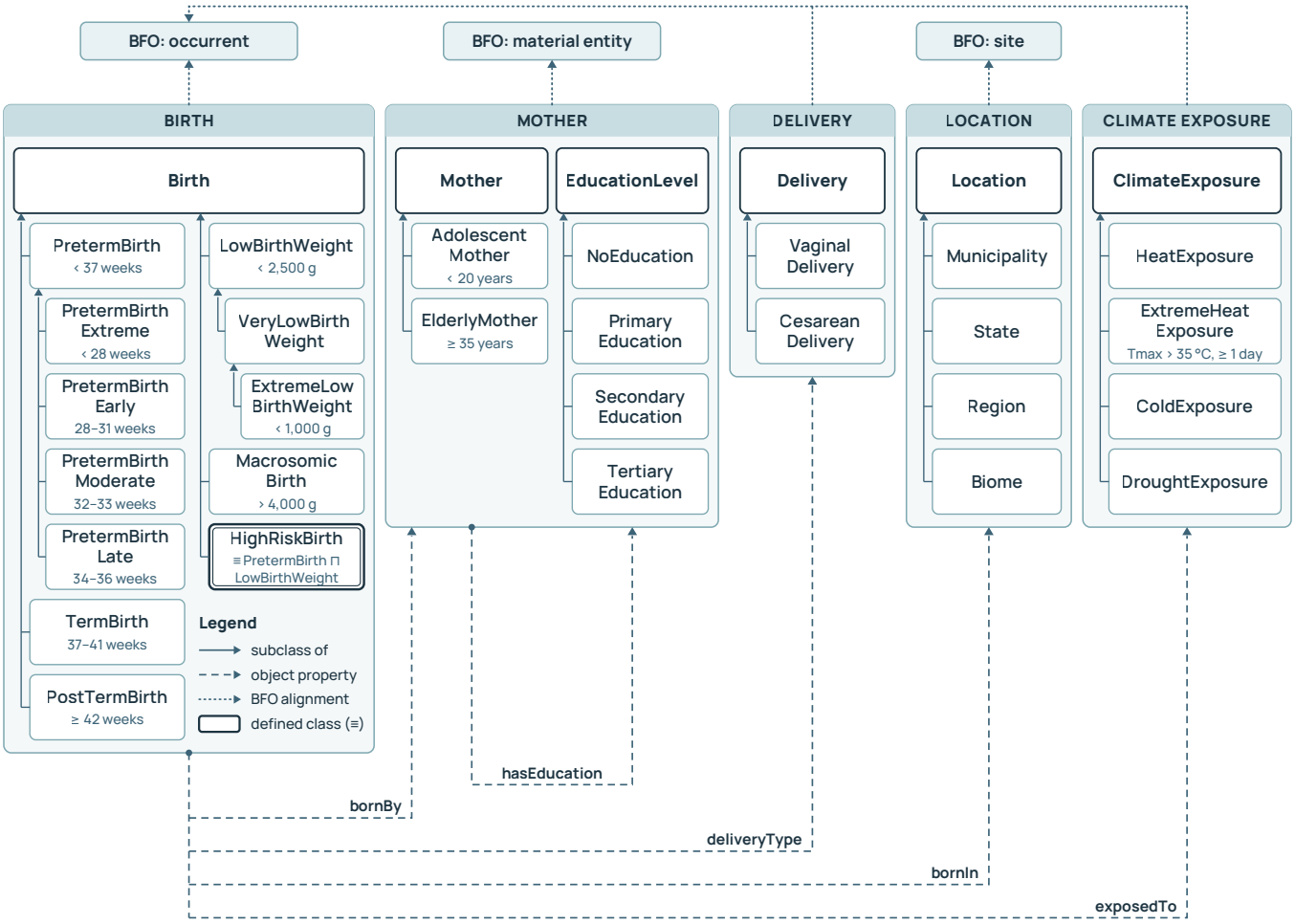
PerinatalKG integrates four public, secondary, and de-identified data sources. Live birth records were obtained from SINASC for the state of Goiás, a Central-West state of approximately 7 million inhabitants, for 2015–2024, covering its 246 municipalities; the files were downloaded from the Department of Informatics of the Brazilian Unified Health System (DATASUS) and harmonized on 9 February 2026, when the 2024 data were already final (consolidated). Mortality records for Goiás in 2020 came from SIM (667 neonatal deaths) and were used exclusively to validate the linkage, not loaded into the graph. Daily climate data were drawn from the Brazilian Daily Weather Gridded Dataset (BR-DWGD) [5], which provides maximum and mean

temperature, precipitation, and related variables at 0.1° × 0.1° resolution. We used version 3.2.4, released in April 2026, which extends the 1961–2020 series of the original publication to 31 December 2025 [17], so that daily data cover the gestation of every birth in 2015–2024. Municipal socioeconomic context comprised the 2010 Municipal Human Development Index, from the Atlas of Human Development in Brazil of the United Nations Development Programme (UNDP), the Institute for Applied Economic Research (Ipea), and the João Pinheiro Foundation (FJP), calculated from census data of the Brazilian Institute of Geography and Statistics (IBGE) [18], and resident population from the 2022 IBGE Census, both modeled as attributes of municipalities.

2.2 Ontology design

The ontology was specified in the description logic profile of OWL 2 (OWL 2 DL) [9] and is reported following the Minimum Information for Reporting an Ontology (MIRO) guidelines [19]. It was developed top-down from the data dictionaries of the source systems and refined iteratively against the source records; the four-grade subclassification of preterm birth, for instance, emerged from the granularity of the SINASC gestational-age field. Following OBO Foundry principles [14], terms from established terminologies were reused before new ones were created. Version 0.3.1 comprises 34 classes, 8 object properties, and 19 data properties in five modules: births, mothers, deliveries, locations, and climate exposures (Table 1, Figure 1). The birth module implements the World Health Organization gestational-age classification (preterm, term, and post-term), with preterm birth subdivided into four severity grades [20], alongside birth-weight categories. A defined-class axiom states that a high-risk birth is equivalent to the intersection of preterm birth and low birth weight (HighRiskBirth ≡ PretermBirth ⊓ LowBirthWeight), so that such births are entailed rather than asserted; instance-level classification under this axiom requires reasoning that the deployment does not perform at query time.

Figure 1. The PerinatalKG ontology, version 0.3.1



All 34 classes of version 0.3.1 are shown, grouped into five modules. Solid arrows denote subclass relations, dashed labeled arrows object properties, and dotted arrows the five alignment axioms linking module roots to Basic Formal Ontology categories; EducationLevel carries no such axiom. HighRiskBirth (double border) is a defined class. Object properties (8 in total): bornIn, bornBy, deliveryType, exposedTo, locatedIn, partOf, hasEducation, hasBiome. Data properties (19 in total) include birthWeight, gestationalAge, and apgar1min. Counts and terminology mappings: Table 1.

Source: Prepared by the authors (2026).

Table 1. PerinatalKG ontology statistics (version 0.3.1)

Component	Count	Notes
OWL 2 classes	34	Five conceptual modules
Object properties	8	e.g., bornIn, exposedTo, bornBy
Data properties	19	e.g., birthWeight, gestationalAge, apgar1min
Ontology RDF triples	414	Before reasoning
Inferred triples (owlrl)	+739	0.24 s; no class inferred unsatisfiable
OWL 2 DL check (HermiT)	Consistent	0.92 s; coherent, with BFO core axioms
SNOMED CT exactMatch	16	SKOS mappings
ICD-10 closeMatch	13	SKOS mappings
LOINC exactMatch	7	Data property mappings
SHACL node shapes	6	Clinical and data-quality constraints
IAO:0000115 definitions	6	Formal textual definitions
BFO alignment axioms	5	OBO Foundry interoperability

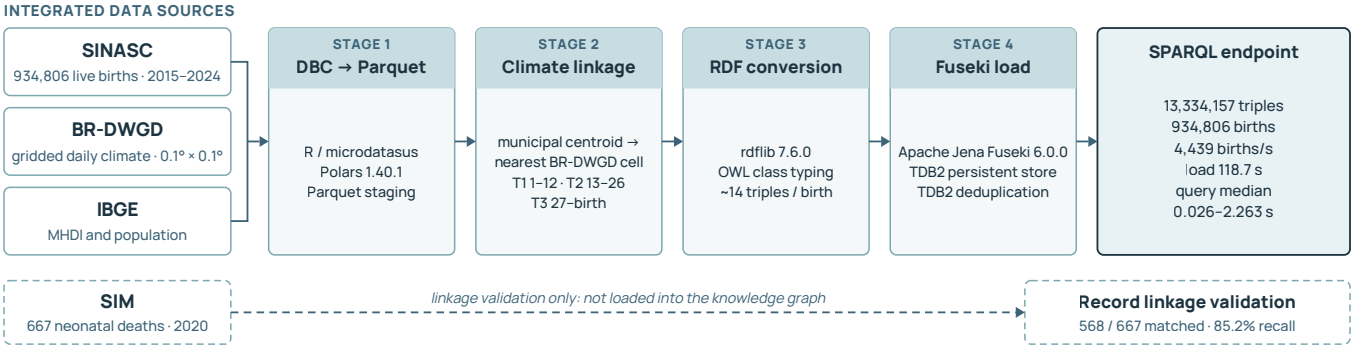
Source: Prepared by the authors (2026).

Five alignment axioms anchor the ontology in BFO [15], one per module root: Birth, Delivery, and ClimateExposure under occurrent categories, Mother under material entity, and Location under site. Six core classes carry formal definitions annotated with the Information Artifact Ontology definition property (IAO: 0000115), and all classes carry English and Portuguese labels. Using Simple Knowledge Organization System (SKOS) mapping properties [21], we added 36 terminology mappings: 29 on classes, comprising 16 exact matches to SNOMED CT [22] and 13 close matches to ICD-10 (2019 edition) [23], and 7 exact matches to LOINC [24] on data properties.

2.3 Extract-transform-load pipeline

An extensible extract-transform-load (ETL) pipeline converts the heterogeneous source files into a single integrated graph (Figure 2). It accepts federative unit and year as parameters, so a different state or period requires only the corresponding input files. SINASC and SIM files, distributed by the Ministry of Health in the compressed DBC format of DATASUS, were converted to tabular form in R with the microdatasus package [25] and stored as Parquet; subsequent processing used Python 3.13.9, Polars 1.40.1, and rdflib 7.6.0. Each birth was instantiated as an OWL individual linked to its mother, municipality, climate exposure, and perinatal outcomes. The object property bornIn links a birth to the mother's municipality of residence (SINASC field CODMUNRES), the same unit used for the climate join; municipality of occurrence is not represented in the graph. Each birth receives both its specific gestational-age subtype and the parent class as asserted types, so that queries on PretermBirth retrieve all preterm births without property-path expressions and without relying on entailment at query time.

Figure 2. The PerinatalKG extract-transform-load pipeline



Three sources feed four processing stages ending in the SPARQL endpoint, with SIM used only for linkage validation and not loaded into the graph; federative unit and year enter as parameters at the climate linkage stage. SHACL validation is applied to the loaded graph; owlrl materializes the OWL 2 RL deductive closure over the ontology (Section 2.5).
Source: Prepared by the authors (2026).

Trimesters followed the classic obstetric convention (weeks 1–12, 13–26, and 27 to birth). Gestational age was used as recorded, without imputation; a birth with missing gestational age or a value outside 20–45 weeks would still be instantiated but would receive no gestational-period classification or trimester exposure, and no such case occurred in the Goiás extraction. Climate exposure was assigned through a spatial join between the centroid of the mother's municipality of residence, derived from the IBGE 2022 boundary dataset, and the nearest BR-DWGD grid cell, a deliberate simplification consistent with the municipal granularity of the records. The graph was loaded into Apache Jena Fuseki 6.0.0 [26] with TDB2 storage, which deduplicates identical triples at load time; a SINASC record appearing in two annual files therefore yields a single Birth individual. All processing ran on one workstation (AMD Ryzen 9 3900X, 24 threads, 64 GB RAM, Ubuntu Linux).

2.4 Record linkage between SINASC and SIM

We implemented a deterministic agreement-scoring approach informed by the Fellegi-Sunter framework [27] and validated it against the SIM records for Goiás in 2020. Validation was restricted to 2020, the proof-of-concept year, because linkage was evaluated as a component of the pipeline rather than used to build a linked 2015–2024 cohort. Candidate pairs were restricted by blocking on exact date of birth and the six-digit municipality code, a strict design that bounds the comparison space but excludes any true pair disagreeing on these fields. Within blocks, agreement was scored on birth weight (difference ≤ 50 g: 3 points; ≤ 100 g: 2; ≤ 200 g: 1), gestational age (exact: 2; ≤ 1 week: 1), and maternal age (exact: 2; ≤ 1 year: 1), to a maximum of 7. Pairs scoring at least 5 were classified as confirmed matches and those scoring 3 or 4 as probable. When a death had several candidates, the highest-scoring one was retained, with ties resolved by birth-weight proximity.

2.5 Evaluation

Because we identified no comparable Brazilian perinatal knowledge graph to serve as a benchmark, the evaluation was internal and descriptive, along four dimensions. Logical consistency was assessed with two reasoners. The owlrl reasoner [28] materializes the deductive closure under the OWL 2 rule-language profile (OWL 2 RL), of polynomial complexity, in pure Python; because OWL 2 RL reasoning is sound but not complete with respect to OWL 2 DL, this check alone does not prove consistency under OWL 2 DL semantics. The ontology was therefore also checked with HermiT [29], a sound and complete reasoner for OWL 2 DL, run through Owlready2 0.51 with the axioms of the BFO 2020 core ontology loaded alongside the PerinatalKG axioms; the ontology was considered logically sound when HermiT found it consistent and coherent (no unsatisfiable named class) and owlrl derived no inconsistency. Structural conformance was assessed with the Shapes Constraint Language (SHACL) [30] against six node shapes (BirthShape, MotherShape, ClimateExposureShape, ApgarShape, PretermBirthShape, and LowBirthWeightShape), which constrain birth weight, gestational age, maternal age, prenatal visits, Apgar scores, and climate variables to physiological ranges and require preterm instances to have gestational age below 37 weeks and low-birth-weight instances a weight below 2,500 g. Linkage performance was measured as recall against SIM and as concordance of auxiliary variables; precision was not estimated because no gold-standard linkage was available. Query performance was assessed with nine SPARQL queries of increasing complexity, each also written as a semantically equivalent Structured Query Language (SQL) statement over the source Parquet files in DuckDB 1.5.2, with ten repetitions after three warm-up runs and medians reported.

Finally, descriptive perinatal queries demonstrated the analytical capability of the graph, and the use case below exercised the integrated dimensions jointly. For the query using the climate dimension, extreme heat exposure was defined as at least one day with maximum temperature above 35 °C during the third trimester, an absolute threshold reflecting maternal thermal stress rather than a percentile-based definition. This query is reported as a demonstration of query capability, not as an epidemiological estimate.

To test whether the integrated graph supports substantive inquiry and not only structural validation, we specified one use case that crosses the three source systems in a single declarative query: the municipal distribution of preterm births among mothers exposed to extreme heat during the third trimester. In the source files this question has no direct answer, because a birth record carries a municipality code but no climate attribute, and the climate product is a grid indexed by coordinates and dates. Because exposure is assigned at municipal resolution and no covariate adjustment or lag structure is modeled, the use case is reported as a descriptive characterization of the analytical reach of the resource, not as an estimate of effect.

2.6 Ethical aspects

The study used exclusively public, secondary data made openly available by the Ministry of Health and IBGE in de-identified form, without direct identifiers. The record linkage between SINASC and SIM likewise used only the public, de-identified microdata files of the two systems, comparing non-nominal fields (date of birth, municipality code, birth weight, gestational age, and maternal age); no nominal or identified database was accessed at any stage. This exclusive use of public, de-identified data is the basis for the waiver of review

by a Research Ethics Committee, in accordance with Brazilian National Health Council Resolution No. 510/2016. No procedure was performed on human participants, no attempt was made to re-identify individuals, and results are reported only in aggregate, in compliance with the Brazilian General Data Protection Law (Law No. 13,709/2018).

3 RESULTS

3.1 Ontology and knowledge graph

The ontology comprises 414 RDF triples (Table 1). OWL 2 RL reasoning materialized 739 additional triples in 0.24 seconds and inferred no unsatisfiable class; HermiT found the ontology consistent and coherent under OWL 2 DL semantics, with the BFO core axioms loaded, in 0.92 seconds. The six classes with formal definitions are Birth, PretermBirth, LowBirthWeight, Mother, ClimateExposure, and ExtremeHeatExposure. SHACL validation reported no violations. Table 2 summarizes the integrated datasets. All 934,806 births were instantiated, with no duplicate birth identifier across the annual files, yielding 13,334,157 triples. Conversion throughput was 4,439 births per second, and loading into TDB2 took 118.7 seconds.

3.2 Record linkage validation

Of the 667 neonatal deaths registered in SIM in 2020, 568 (85.2%) were linked to a SINASC record: 514 confirmed and 54 probable matches (Table 3). Of all linked records, 414 (72.9%) scored 6 or 7, and the mean score was 6.2. Among the 559 linked records with birth weight recorded in both systems, concordance was 88.0% for exact agreement and 96.6% within 100 g (mean absolute difference, 27.5 g), consistent with known inter-system transcription variability. Among confirmed matches, 384 (74.7%) were early neonatal deaths (0–6 days).

3.3 Query performance and descriptive characterization

Table 4 presents the benchmark. Class-based counts (Q2–Q6) returned in 0.026 to 0.824 seconds. The multidimensional join (Q8), linking preterm birth, third-trimester heat exposure, and municipality, returned in 0.430 seconds and is reported in full in Section 3.4. The columnar engine was faster on every comparable query (0.0038 to 0.0095 seconds), but the ratio was not constant: widest for the heat-exposure count (531× for Q7) and for trivial class counts (217× for Q2), and narrower for the join (45× for Q8). Q7 showed

Table 2. Datasets integrated in the PerinatalKG proof of concept (Goiás state)

Source	Records/units	Scope	Use in the graph
SINASC	934,806 births	Goiás, 2015–2024	Birth nodes and attributes
SIM	667 neonatal deaths	Goiás, 2020	Linkage validation only
BR-DWGD	Nearest grid cell per municipal centroid	0.1° × 0.1°, daily; version 3.2.4	Climate exposure variables
IBGE; Atlas (UNDP, Ipea, FJP)	246 municipalities	Census 2022 (IBGE); MHDI 2010 (Atlas)	Location and socioeconomic context

MHDI: Municipal Human Development Index. Forty births (0.004%) carried an invalid municipality code and therefore received no climate linkage; they remain in the denominators, and all 246 municipalities of Goiás are represented.

Source: Prepared by the authors (2026), from SINASC, SIM, BR-DWGD [5,17], IBGE, and the Atlas of Human Development in Brazil (UNDP, Ipea, and FJP) [18].

Table 3. SINASC–SIM record linkage results (Goiás, 2020)

Metric	Value	Notes
SIM neonatal deaths	667	Ages 0–27 days
Linked (recall)	568 (85.2%)	Confirmed plus probable
Confirmed matches (score ≥ 5)	514 (77.1%)	High certainty
Probable matches (score 3–4)	54 (8.1%)	Lower certainty
High-confidence links (score 6–7)	414/568 (72.9%)	Among linked records
Mean linkage score (all linked records)	6.2/7	Among the 568 linked records
Birth-weight concordance, exact	88.0% (492/559)	Linked records with weight in both systems
Birth-weight concordance, ±100 g	96.6% (540/559)	Linked records with weight in both systems
Early neonatal deaths (0–6 days)	384/514 (74.7%)	Among confirmed matches
Precision	Not estimated	No gold standard available

Percentages for linked, confirmed, and probable matches use the 667 SIM deaths as denominator; high-confidence links use the 568 linked records; early neonatal deaths use the 514 confirmed matches; the mean score uses the 568 linked records; birth-weight concordance uses the 559 linked records with weight recorded in both systems.

Source: Prepared by the authors (2026).

Table 4. SPARQL queries over the knowledge graph versus equivalent SQL over the source Parquet files (medians of ten repetitions after three warm-up runs)

Query	Description	Result	SPARQL (s)	SQL (s)	Ratio
Q1	Total count (triples vs rows)	13,334,157	2.263	0.0037	n.c.
Q2	Birth individuals	934,806	0.824	0.0038	217×
Q3	Preterm births (< 37 weeks)	101,287 (10.8%)	0.104	0.0068	15×
Q4	Term births (37–41 weeks)	815,094 (87.2%)	0.802	0.0083	97×
Q5	Post-term births (≥ 42 weeks)	18,425 (2.0%)	0.026	0.0072	4×
Q6	Low birth weight (< 2,500 g)	83,660 (8.9%)	0.093	0.0074	13×
Q7	Extreme heat exposure in third trimester	387,012 (41.4%)	2.176	0.0041	531×
Q8	Join: preterm × heat × municipality	245/246 municipalities	0.430	0.0095	45×
Q9	Mean birth weight	3,141 g	0.336	0.0061	55×

Apache Jena Fuseki 6.0.0 with TDB2 (13,334,157 triples) and DuckDB 1.5.2 in memory (934,806 rows); AMD Ryzen 9 3900X, 64 GB RAM. Ratio: median SPARQL time divided by median SQL time, both from the same ten repetitions; SQL times are shown to four decimals because they fall below 0.01 s. n.c.: not comparable, because Q1 counts graph triples against table rows. Percentages use 934,806 births as denominator.

Source: Prepared by the authors (2026).

Figure 3. SPARQL query answering the use case

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX pkg: <http://perinatalkg.org/ontology/>

SELECT ?municipality (COUNT(?b) AS ?n_preterm_heat)
WHERE {
  ?b   rdf:type     pkg:PretermBirth .
  ?b   pkg:exposedTo ?exp .
  ?exp rdf:type     pkg:ExtremeHeatExposure .
  ?b   pkg:bornIn   ?municipality .
}
GROUP BY ?municipality
ORDER BY DESC(?n_preterm_heat)
```

The query joins the birth dimension (SINASC), the climate dimension (BR-DWGD), and the geographic dimension (IBGE), which share no common key in the source files. Extreme heat is defined as at least one day with a maximum temperature above 35 °C during the third trimester; that restriction is encoded in the ExtremeHeatExposure class itself, so the query needs no trimester filter. Exposure is modeled on the Birth individual as a proxy for maternal exposure, and results are grouped by the municipality individual, whose identifier carries the IBGE municipal code. Query text reproduced from the project repository.

Source: Prepared by the authors (2026).

the widest spread across repetitions (0.511 to 2.649 seconds). Descriptively, the preterm birth rate was 10.8%, the term rate 87.2%, the post-term rate 2.0%, the low-birth-weight rate 8.9%, and the mean birth weight 3,141 g; 41.4% of births had at least one day above 35 °C in the third trimester, a figure that illustrates query capability rather than a validated prevalence. The full query text is available in the project repository.

3.4 Use case: one question across three source systems

We asked how preterm births among mothers exposed to extreme heat in the third trimester are distributed across the municipalities of Goiás. In the graph the question is a single query (Figure 3). It returned 245 of the 246 municipalities of the state in 0.430 seconds; the remaining municipality recorded no birth meeting both conditions. Grouped by the mother's municipality of residence, the largest count was in Goiânia, the state capital, with 11,035 preterm births whose mothers experienced at least one day above 35 °C during the third trimester. For context, the same graph places the overall preterm birth rate at 10.8% (101,287 of 934,806 births) and third-trimester extreme-heat exposure at 41.4% (387,012 of 934,806).

Three features of this result concern the resource rather than the epidemiology. The join is expressed declaratively, with no procedural code reconciling identifiers, spatial units, or temporal resolutions, because the three dimensions share an explicit semantics. It executes in well under a second over 13.3 million triples, so the question can be posed interactively and refined. And substituting a single class turns it into a different question: replacing PretermBirth with LowBirthWeight, or ExtremeHeatExposure with ColdExposure, requires no change to the data, the pipeline, or the query structure. The counts themselves are crude and unadjusted, exposure is attributed through the municipal centroid rather than to the individual, and the distribution across municipalities reflects population size as much as any exposure gradient, so no association between heat and preterm birth should be read from them.

4 DISCUSSION

PerinatalKG integrates Brazilian live birth, climate, and socioeconomic data under an OWL 2 ontology aligned to BFO and mapped to international terminologies, with record linkage to the national mortality system evaluated for recall as a pipeline component. The evaluation showed a logically consistent ontology with no unsatisfiable classes under OWL 2 RL and OWL 2 DL reasoning, no SHACL violations, 85.2% linkage recall, and sub-2.3-second median responses for all nine queries over more than 13 million triples. To our knowledge, and based on searches of PubMed and Scopus combining knowledge graph and ontology terms with perinatal concepts, no comparable formally specified resource exists for the Brazilian perinatal domain; the contribution is thus both a domain

resource and a reproducible, empirically evaluated architecture.

The use case reported in Section 3.4 shows the practical consequence of that formal semantic layer. Asking how preterm births among heat-exposed mothers are distributed across 246 municipalities is, in the source files, a data-engineering task that must be completed before any question can be asked; in the graph the link is already part of the representation, so the question is one query and the answer returns in under half a second. The same pattern answers the low-birth-weight or cold-exposure variants without touching the pipeline. This is the capability on which the planned applied analyses depend, and it is what distinguishes a reusable resource from a dataset assembled for one study.

Although deployed for a single state, three properties support national scalability. The ontology encodes no state-specific terms. The pipeline is parameterized by federative unit and period, so a new state requires only new input files. The measured throughput makes national scale tractable: at 4,439 births per second, the approximately 27 million births recorded in SINASC over 2015–2024 would be converted in about 100 minutes of sequential processing, and parallelization across the 27 federative units would reduce this further. Population-scale integration of Brazilian administrative data is already proven by the 100 Million Brazilian Cohort [7]; what PerinatalKG adds is a formal semantic layer. Because nothing in the design is specific to Brazil, the same ontology and pipeline could be instantiated wherever vital-statistics systems coexist with gridded climate data.

Extending the graph to another federative unit is an operational rather than a design task, and it is worth stating precisely what it requires. Four inputs are needed: the SINASC file for the target unit and period, which DATASUS distributes with identical structure for all 27 federative units; the BR-DWGD grid cells intersecting the unit, available at the same resolution for the whole national territory; the IBGE municipal boundary layer, from which the centroids are derived; and the municipal socioeconomic indicators, published for every Brazilian municipality. With those files in place the pipeline is invoked with the federative-unit code and the year range as parameters. No term is added to the ontology, no transformation rule is rewritten, and the SHACL shapes apply unchanged, because every constraint they encode is clinical rather than jurisdictional. What changes with scale is operational: the graph holds approximately 14.3 triples per birth, so a national consolidation of about 27 million births would reach the order of 385 million triples. That volume is within what TDB2 is engineered to store, but it exceeds what we have tested for OWL materialization. We therefore expect extension to individual states to be immediate, and national consolidation to require a decision on reasoning strategy, most plausibly materialization partitioned by federative unit.

One component of this argument requires caution. Conversion throughput is the inexpensive part; viability at national scale depends on whether a graph of several hundred million triples remains tractable to reason over and query. Loading and routine retrieval scale predictably under TDB2, but OWL reasoning at that volume, and complex joins, were not tested; incremental or partition-wise materialization by federative unit are plausible routes whose adequacy remains an open empirical question.

Health data harmonization is currently dominated by relational common data models such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model maintained by the Observational Health Data Sciences and Informatics (OHDSI) community, whose standardized vocabularies are far larger in scope than the 36 mappings reported here [31]. The distinction is not that such models lack semantics, but that they use relational schemas without OWL axioms, decidable reasoning, or upper-level alignment. PerinatalKG makes the complementary choice, trading tooling maturity for logical expressivity, verifiable consistency, and federability with the biomedical ontology ecosystem, in line with the Findable, Accessible, Interoperable, and Reusable (FAIR) principles [32]. Compared with Brazilian linkage efforts [6-8], it treats linkage as one component of a reusable, formally specified pipeline rather than as a study-specific end product. The benefit of the formal layer is already visible: the join in Q8 is a single declarative pattern because its dimensions share explicit semantics, and the SHACL shapes turn clinical-consistency rules into machine-checkable constraints rather than bespoke validation code. The measured overhead, although large in relative terms, remained below 2.3 seconds in absolute terms.

This work has limitations. First, scope: the deployment covers one state, so national generalization is argued from architectural properties and measured throughput rather than demonstrated. Second, the benchmark compares retrieval against a columnar engine over the same files and does not measure PerinatalKG against common data models, relative to which its positioning remains argued. Third, SIM was used only for linkage validation, so mortality is not yet a first-class dimension of the graph. Fourth, linkage recall was estimated for 2020 only and may differ in other years of the series; precision was not estimated for lack of a gold standard, the deterministic design is simpler than full probabilistic approaches [8,27], and strict blocking bounds recall to pairs agreeing on date of birth and municipality. Fifth, the climate join operates at municipal resolution and captures neither intra-municipal variation nor maternal relocation; the

socioeconomic attributes are static, with the development index from 2010 and population from 2022, and do not vary across the 2015–2024 series; and the resource inherits the known data-quality limitations of SINASC [2]. Sixth, SINASC records sex assigned at birth, not gender, and no sex-stratified analysis was performed. Seventh, consistency under OWL 2 DL was verified for the ontology, not for the instance data, whose conformance was checked with SHACL rather than with a complete reasoner. Finally, the use case in Section 3.4 is descriptive: the counts are crude, exposure is assigned at municipal resolution rather than individually, no covariate adjustment or lag structure is modeled, and the municipal distribution reflects population size as much as any exposure gradient. Neither it nor the extreme-heat figure supports a causal or prevalence inference; both show what the resource makes answerable.

Next steps are to extend the graph to all federative units, to load SIM as a first-class dimension, and to submit the ontology to BioPortal. The resource was designed as the infrastructure for a planned national study of preterm birth phenotyping, in which the climate–birth integration shown here will support association analyses of the kind recently conducted for ambient heat and neonatal outcomes in Brazil [33], replacing pipelines that each study currently rebuilds.

5 CONCLUSION

PerinatalKG replaces bespoke, single-use integration of Brazilian perinatal records with a formally specified alternative in which birth, climate, and socioeconomic data share explicit semantics, automated consistency checking, and standardized querying, with record linkage to the national mortality system whose recall was validated against one year of neonatal deaths. In the Goiás proof of concept, the ontology proved logically consistent and the graph structurally valid and fast to query in absolute terms. Because the ontology carries no jurisdiction-specific terms and the pipeline requires only new input files, the architecture extends to national scale, pending confirmation of reasoning at full volume, and to other settings that pair vital statistics with gridded climate data. In the use case reported here it answered, in a single declarative query, a question spanning birth, climate, and geography that the source files cannot answer without bespoke integration. Released openly with a versioned ontology, it is offered as shared infrastructure for climate-aware perinatal research.

Author contributions. GC: conceptualization, data curation, formal analysis, methodology, visualization, writing – original draft, writing – review & editing. DSK: conceptualization, investigation, methodology, project administration, writing – review & editing. CMC: conceptualization, data curation, writing – original draft, writing – review & editing. RCP: conceptualization, data curation, formal analysis, supervision, visualization, writing – original draft, writing – review & editing. CJC: data curation, formal analysis, methodology, software, validation, writing – review & editing. All authors read and approved the final version.

Acknowledgments. We thank the members of the Climate and Maternal-Child Health Research Group (CliMaterna) at the Universidade Estadual de Campinas (UNICAMP) for their discussions and methodological support, and Breno Bernard Nicolau de França (Instituto de Computação, UNICAMP) for data management support in the project repository.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data and materials availability. The ontology, extract-transform-load pipeline, SHACL shapes, and SPARQL queries are publicly available under the Apache 2.0 license in an open repository with a permanent archival DOI, tagged to ontology version 0.3.1 (https://github.com/clarimar/perinatalkg; DOI: https://doi.org/10.5281/zenodo.24476096). The source data are public: SINASC and SIM (Ministry of Health/DATASUS), BR-DWGD (version 3.2.4), IBGE, and the Atlas of Human Development in Brazil.

Conflicts of interest. The authors declare that they have no conflicts of interest.

Use of artificial intelligence. Artificial intelligence tools were used to support the writing, translation of the titles and abstracts, language revision, and textual organization of the manuscript. Tool: Claude (Anthropic). The tool was not used to generate, analyze, or interpret data, did not replace human authorship, and all content was reviewed and validated by the authors, who take full responsibility for the work presented. A generative image model (Gemini, Google) was additionally used to produce the abstract illustrative image submitted as a supplementary file at the request of the editorial office, from a prompt written by the authors; it was not used to produce any figure that reports data, and Figures 1 to 3 were produced by the authors.

REFERENCES

1. Instituto Brasileiro de Geografia e Estatística. Estatísticas do Registro Civil 2023 [Internet]. Rio de Janeiro: IBGE; 2025 [cited 2026 Sep 16]. Available from: https://www.ibge.gov.br/estatisticas/sociais/populacao/9110-estatisticas-do-registro-civil.html

2. Szwarcwald CL, Leal MC, Esteves-Pereira AP, Almeida WS, Frias PG, Damacena GN, et al. Avaliação das informações do Sistema de Informações sobre Nascidos Vivos (SINASC) no Brasil. Cad Saude Publica. 2019;35(10):e00214918. doi: https://doi.org/10.1590/0102-311X00214918

3. Lakhoo DP, Brink N, Radebe L, Craig MH, Pham MD, Haghighi MM, et al. A systematic review and meta-analysis of heat exposure impacts on maternal, fetal and neonatal health. Nat Med. 2025;31(2):684-94. doi: https://doi.org/10.1038/s41591-024-03395-8

4. Chersich MF, Pham MD, Areal A, Haghighi MM, Manyuchi A, Swift CP, et al. Associations between high temperatures in pregnancy and risk of preterm birth, low birth weight, and stillbirths: systematic review and meta-analysis. BMJ. 2020;371:m3811. doi: https://doi.org/10.1136/bmj.m3811

5. Xavier AC, Scanlon BR, King CW, Alves AI. New improved Brazilian daily weather gridded data (1961-2020). Int J Climatol. 2022;42(16):8390-404. doi: https://doi.org/10.1002/joc.7731

6. Maia LTS, Souza WV, Mendes ACG, Silva AGS. Uso do linkage para a melhoria da completude do SIM e do SINASC nas capitais brasileiras. Rev Saude Publica. 2017;51:112. doi: https://doi.org/10.11606/S1518-8787.2017051000431

7. Barreto ML, Ichihara MY, Pescarini JM, Ali MS, Borges GL, Fiaccone RL, et al. Cohort profile: the 100 Million Brazilian Cohort. Int J Epidemiol. 2022;51(2):e27-38. doi: https://doi.org/10.1093/ije/dyab213

8. Pita R, Pinto C, Sena S, Fiaccone R, Amorim L, Reis S, et al. On the accuracy and scalability of probabilistic data linkage over the Brazilian 114 million cohort. IEEE J Biomed Health Inform. 2018;22(2):346-53. doi: https://doi.org/10.1109/JBHI.2018.2796941

9. W3C OWL Working Group. OWL 2 Web Ontology Language document overview (second edition): W3C Recommendation 11 December 2012 [Internet]. 2012 [cited 2026 Sep 16]. Available from: https://www.w3.org/TR/owl2-overview/

10. Hogan A, Blomqvist E, Cochez M, d'Amato C, de Melo G, Gutierrez C, et al. Knowledge graphs. ACM Comput Surv. 2022;54(4):71. doi: https://doi.org/10.1145/3447772

11. Callahan TJ, Tripodi IJ, Pielke-Lombardo H, Hunter LE. Knowledge-based biomedical data science. Annu Rev Biomed Data Sci. 2020;3:23-41. doi: https://doi.org/10.1146/annurev-biodatasci-010820-091627

12. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023;10(1):67. doi: https://doi.org/10.1038/s41597-023-01960-3

13. Cavalleri E, Cabri A, Soto-Gomez M, Bonfitto S, Perlasca P, Gliozzo J, et al. An ontology-based knowledge graph for representing interactions involving RNA molecules. Sci Data. 2024;11(1):906. doi: https://doi.org/10.1038/s41597-024-03673-7

14. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25(11):1251-5. doi: https://doi.org/10.1038/nbt1346

15. Arp R, Smith B, Spear AD. Building ontologies with Basic Formal Ontology. Cambridge (MA): MIT Press; 2015. doi: https://doi.org/10.7551/mitpress/9780262527811.001.0001

16. Alocci D, Mariethoz J, Horlacher O, Bolleman JT, Campbell MP, Lisacek F. Property graph vs RDF triple store: a comparison on glycan substructure search. PLoS One. 2015;10(12):e0144578. doi: https://doi.org/10.1371/journal.pone.0144578

17. Xavier AC. Brazilian Daily Weather Gridded Data (BR-DWGD): release notes, version 3.2.4 [Internet]. 2026 [cited 2026 Nov 23]. Available from: https://sites.google.com/site/alexandrecandidoxavierufes/brazilian-daily-weather-gridded-data

18. Programa das Nações Unidas para o Desenvolvimento, Instituto de Pesquisa Econômica Aplicada, Fundação João Pinheiro. O Índice de Desenvolvimento Humano Municipal Brasileiro (Série Atlas do Desenvolvimento Humano no Brasil 2013) [Internet]. Brasília: PNUD, Ipea, FJP; 2013 [cited 2026 Nov 23]. Available from: https://www.undp.org/sites/g/files/zskgke326/files/2023-07/indice_de_desenvolvimento_humano_municipal_brasileiro_-_2013.pdf

19. Matentzoglu N, Malone J, Mungall C, Stevens R. MIRO: guidelines for minimum information for the reporting of an ontology. J Biomed Semantics. 2018;9(1):6. doi: https://doi.org/10.1186/s13326-017-0172-7

20. Ohuma EO, Moller AB, Bradley E, Chakwera S, Hussain-Alkhateeb L, Lewin A, et al. National, regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis. Lancet. 2023;402(10409):1261-71. doi: https://doi.org/10.1016/S0140-6736(23)00878-4

21. Miles A, Bechhofer S, editors. SKOS Simple Knowledge Organization System reference: W3C Recommendation 18 August 2009 [Internet]. 2009 [cited 2026 Sep 16]. Available from: https://www.w3.org/TR/skos-reference/

22. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. Stud Health Technol Inform. 2006;121:279-90. Available from: https://pubmed.ncbi.nlm.nih.gov/17095826/

23. World Health Organization. International statistical classification of diseases and related health problems, 10th revision: 2019 version [Internet]. Geneva: WHO; 2019 [cited 2026 Sep 16]. Available from: https://icd.who.int/browse10/2019/en

24. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem. 2003;49(4):624-33. doi: https://doi.org/10.1373/49.4.624

25. Saldanha RF, Bastos RR, Barcellos C. Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). Cad Saude Publica. 2019;35(9):e00032419. doi: https://doi.org/10.1590/0102-311X00032419

26. Apache Software Foundation. Apache Jena Fuseki [software], version 6.0.0 [Internet]. 2026 [cited 2026 Sep 16]. Available from: https://jena.apache.org/

27. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc. 1969;64(328):1183-210. doi: https://doi.org/10.1080/01621459.1969.10501049

28. Herman I. OWL-RL: a simple OWL 2 RL reasoner on top of RDFLib [software]. Geneva: Zenodo; 2014. doi: https://doi.org/10.5281/zenodo.14543

29. Glimm B, Horrocks I, Motik B, Stoilos G, Wang Z. HermiT: an OWL 2 reasoner. J Autom Reason. 2014;53(3):245-69. doi: https://doi.org/10.1007/s10817-014-9305-1

30. Knublauch H, Kontokostas D, editors. Shapes Constraint Language (SHACL): W3C Recommendation 20 July 2017 [Internet]. 2017 [cited 2026 Sep 16]. Available from: https://www.w3.org/TR/shacl/

31. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574-8. doi: https://doi.org/10.3233/978-1-61499-564-7-574

32. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi: https://doi.org/10.1038/sdata.2016.18

33. Del Carretto M, Godin A, Neves D, Paixão ES, Wan K, Pescarini J, et al. Short-term ambient heat exposure and low APGAR score in newborns: a time-stratified case-crossover analysis in São Paulo state, Brazil (2013-2019). PLOS Glob Public Health. 2025;5(9):e0004557. doi: https://doi.org/10.1371/journal.pgph.0004557

EDITORIAL NOTES

The Journal of Health Informatics is a scientific journal maintained by the Brazilian Society of Health Informatics (Sociedade Brasileira de Informática em Saúde), a non-profit scientific association dedicated to the development of Health Informatics and Digital Health in Brazil. The editorial process is conducted with scientific independence, free from institutional, commercial, or sponsor interference, and follows international best practices in publication ethics.

This article was double-blind peer reviewed by specialists in the field. The editorial team used artificial intelligence tools, under human supervision, in the initial screening, in checking the text, references and numbers and, when needed, in redrawing figures, charts and tables, which the authors approved in the proof. The tools and safeguards are described in the journal's Policy on the use of artificial intelligence. Editorial decisions were made by people.

SBIS Sociedade Brasileira de Informática em Saúde