EpiCorpus-BR: A Named Entity Recognition Corpus for Epidemiological Documents in Portuguese

Christian Freitas, Lilian Berton

[1]Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brazil

christian.freitas@unifesp.br, lberton@unifesp.br

Abstract. *This paper presents* ***EpiCorpus-BR****, the first annotated corpus for Named Entity Recognition (NER) in Brazilian Portuguese epidemiological surveillance documents, covering the annual SIREVA-SUS reports (2013–2024) and complementary documents from the Brazilian Ministry of Health. The corpus comprises* ***8,314 textual units*** *across 22 documents (6,358 narrative text sentences and 1,956 table rows), with* ***4,159 entity mentions*** *in the silver set. The tagset consists of eight specialized categories (*PATOGENO*,* SOROTIPO*,* FAIXA_ETARIA*,* LOCAL*,* PERIODO_TEMPORAL*,* METODO_LAB*,* METRICA_EPI*,* MANIFESTACAO_CLINICA*), annotated via few-shot prompting with GPT-4o-mini and manually reviewed in a stratified sample of* ***280 units*** *(180 text sentences and 100 table rows) by two independent annotators (κ = 0.76). Strict IOB2 evaluation with seqeval yields a combined micro-F_1 of 0.77.* METRICA_EPI *is absent from narrative text but reaches F_1 = 0.89 on table rows, which shows why the tabular sub-corpus must be included in the evaluation. A BERTimbau baseline fine-tuned on the silver corpus, with the gold held out, reaches micro-F_1 = 0.73, showing that the corpus supports training a dedicated NER model. The corpus and code are publicly available.*

Keywords: Named Entity Recognition, Corpus, Epidemiological Surveillance, Brazilian Portuguese, Semi-automatic Annotation, SIREVA-SUS

1. Introduction

In Brazil, epidemiological surveillance of invasive diseases is carried out by the Health Surveillance Secretariat (SVS) of the Ministry of Health [Secretaria de Vigilância em Saúde 2024]. It produces a large volume of technical documents: annual reports, bulletins, and advisories. These texts describe, in specialized prose, etiological agents, circulating serotypes, age groups, laboratory methods, and morbidity and mortality metrics. Despite this richness, the content is still out of reach for automatic information extraction. English already has well-established corpora for biomedical Named Entity Recognition (NER), such as i2b2/n2c2 [Uzuner et al. 2011] and NCBI Disease [Doğan et al. 2014], but no comparable resource exists for the epidemiological surveillance domain in Portuguese [Albuquerque et al. 2023].

This work is part of a line of NLP research applied to SIREVA-SUS documents [Freitas et al. 2026, Freitas et al. 2025], introducing NER [Nadeau and Sekine 2007] as a semantic annotation layer. EpiCorpus-BR proposes a tagset of eight specialized categories (PATOGENO, SOROTIPO,

FAIXA_ETARIA, LOCAL, PERIODO_TEMPORAL, METODO_LAB, METRICA_EPI, and MANIFESTACAO_CLINICA), absent from traditional biomedical ontologies [Uzuner et al. 2011, Doğan et al. 2014] and nonexistent in Portuguese-language resources [Albuquerque et al. 2023].

The construction pipeline combines four steps: (i) narrative text extraction via Docling (scenario B2 from [Freitas et al. 2026]) and table row extraction via pdfplumber, (ii) silver annotation via few-shot prompting with GPT-4o-mini [OpenAI 2024], (iii) manual review of a stratified sample to build the gold standard, and (iv) evaluation with seqeval in the strict IOB2 (*Inside–Outside–Beginning*) labelling scheme. The corpus covers 12 annual SIREVA-SUS reports (2013–2024) and 10 complementary documents, totalling 22 documents, 8,314 textual units (6,358 sentences and 1,956 table rows), and 4,159 entity mentions in the silver set.

The contributions of this work are: 1) An NER tagset with eight specialized epidemiological categories for the invasive disease surveillance domain in Brazilian Portuguese (Section 3). 2) A semi-automatic annotation pipeline combining structured extraction via Docling, silver annotation with a large language model (LLM), and manual review, replicable for other specialized domains without available training data (Section 4). 3) To the best of our knowledge, the first public NER corpus for epidemiological surveillance documents in Portuguese, covering 12 years of SIREVA-SUS data and complementary documents from the Ministry of Health (Section 5). 4) A per-category evaluation using seqeval metrics comparing the silver annotation against the gold standard, a second-LLM annotator comparison (GPT-4o-mini vs. Gemini 2.5 Flash), and a supervised BERTimbau baseline trained on the silver corpus that demonstrates the corpus's downstream utility (Section 6).

2. Related Work

Biomedical and clinical NER in Portuguese. The survey by [Albuquerque et al. 2023] maps NER resources for Portuguese, showing that most available systems and corpora focus on journalistic and general-domain text, with sparse coverage for specialized biomedical domains. SemClinBr [Oliveira et al. 2022] is the main semantically annotated clinical corpus for Brazilian Portuguese, covering entities according to Unified Medical Language System (UMLS) semantic types (diseases, medications, and procedures) across 1,000 hospital clinical notes. Although a reference for clinical NLP in Portuguese, the UMLS semantic types do not encompass epidemiological surveillance categories such as serotypes, surveillance laboratory methods, population-level metrics, or temporal surveillance periods. On the modelling side, BioBERTpt [Schneider et al. 2020] adapted BERT to the clinical and biomedical domain in Portuguese and became a reference for clinical NER in the language; however, models like this depend on domain-specific annotated corpora for fine-tuning, a resource still nonexistent for population-level epidemiological surveillance. In English, NCBI Disease [Doğan et al. 2014] and i2b2/n2c2 [Uzuner et al. 2011] set standards for biomedical NER, but both focus on individual patient records in hospital settings, a text type fundamentally different from population-scale surveillance reports, where the central entity is not the patient but the etiological agent and its epidemiological distribution; comparable resources for epidemiological surveillance in Portuguese are nonexistent [Albuquerque et al. 2023].

Domain-refined NER corpora in Portuguese. A recent line of work in the Portuguese NLP community has invested in NER corpora with refined tagsets tailored to specialized domains, showing that generic categories (Person, Organization, Location) are insufficient to capture the relevant entities of each area. Examples include PetroGeoNER, in the oil and gas domain [Moreira et al. 2025]; the linguistically motivated annotation of entities in stock-market tweets [Piai et al. 2025]; CDJUR-BR, in the legal domain [Brito et al. 2023]; and approaches for NER in clinical case reports [Ramos et al. 2026] and for medical entities with compact BERT models [Pontes et al. 2024]. EpiCorpus-BR belongs to this tradition of specialized tagsets, but addresses a domain not yet covered: population-scale surveillance of invasive diseases. Its categories (pathogen, serotype, surveillance laboratory method, epidemiological metric) have no counterpart in the previous resources.

Semi-automatic annotation with language models. The use of large language models (LLMs) for generating silver annotations has gained traction as a cost-effective alternative to full manual annotation [Ding et al. 2024], particularly in scenarios where annotated training data are scarce or unavailable. [Ding et al. 2024] organize LLM-based data augmentation strategies into three learning paradigms (training, fine-tuning, and in-context learning) and identify that annotation quality depends critically on prompt design.

3. Named Entity Tagset

The tagset comprises eight closed categories, designed specifically for the epidemiological surveillance domain for invasive diseases in Brazilian Portuguese. The categories were defined through qualitative analysis of SIREVA-SUS reports and iteration over a prior prototype with a different tagset. Table 1 presents the eight categories with definitions and examples extracted from the reports.

Table 1. SIREVA-SUS-NER tagset categories, with definitions and examples from the documents.

Category	Definition	Examples
PATOGENO	Organism causing invasive disease (bacteria, viruses, fungi).	*S. pneumoniae*, *N. meningitidis*, pneumococo
SOROTIPO	Serological identification: serotype, serogroup, or type.	sorotipo 19A, sorogrupo B, tipo b
FAIXA_ETARIA	Age range used as epidemiological stratification.	menor que 12 meses, 60 anos ou mais
LOCAL	Named geographical location (country, state, region).	Brasil, São Paulo, região Sudeste
PERIODO_ TEMPORAL	Explicit temporal span: years, weeks, or months.	2024, 2013 a 2024, semana epidemiológica 35
METODO_LAB	Named laboratory or diagnostic technique.	cultura, PCR em tempo real, MALDI-TOF
METRICA_EPI	Quantitative measure: the named concept in text, each numeric value in tables.	taxa de incidência (text); 42, 97.4 (table)
MANIFESTACAO_ CLINICA	Named disease, syndrome, or clinical condition.	meningite, pneumonia, doença invasiva

3.1. Design Decisions

Separating PATOGENO and SOROTIPO; removing VACINA and ORG. In a prior prototype, serotype and pathogen were fused into a single category, generating ambiguity in expressions such as *"pneumococo sorotipo 19A"*. Separating them allows independent tracking of agent and serotype distributions across the historical series. Categories for vaccines and institutional organizations were removed: vaccines appear primarily as context for vaccination coverage (captured by METRICA_EPI), and organization names introduce high ambiguity without contributing to the central epidemiological analyses.

Distinguishing METRICA_EPI in text vs. tables. In narrative text, the named concept is annotated (*"taxa de incidência"*, *"cobertura vacinal"*); in table rows, each individual numeric value is annotated, since domain metrics are concentrated in tables rather than prose (see Section 6).

4. Corpus Construction Pipeline

The pipeline comprises four sequential steps, illustrated in Figure 1: (i) extraction of narrative text and table rows from PDFs, (ii) automatic silver annotation via few-shot prompting, (iii) manual review to build the gold standard, and (iv) automatic evaluation with seqeval.

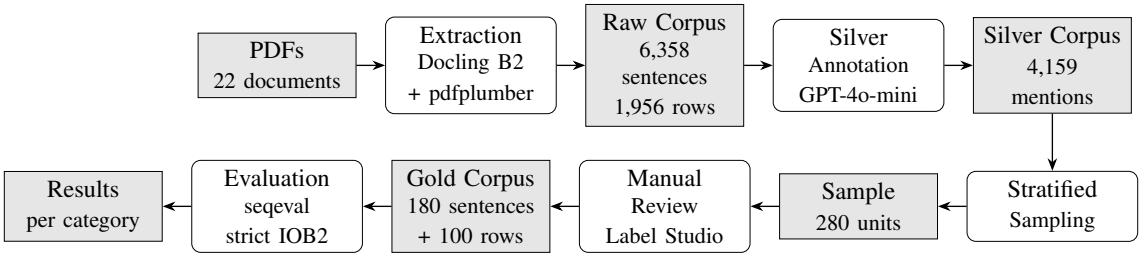

Figure 1. Overview of the EpiCorpus-BR construction pipeline. Grey rectangles indicate data artifacts; white rectangles indicate processing steps.

4.1. Text Extraction from PDF Documents

Extraction applies the **B2 scenario** from [Freitas et al. 2026]: text-only conversion via **Docling** [DS4SD Team 2024] with explicit removal of Markdown tables and preservation of captions. Since the PDFs are natively text-based, OCR is not required. The extracted text passes through three filters: removal of residual tabular lines, conservative sentence segmentation (boundaries detected only before uppercase letters or digits), and a narrative prose filter that discards non-alphabetic fragments and summary lines. The result is **6,358 sentences** in 22 JSONL files (3,343 from SIREVA-SUS and 3,015 from complementary documents).

4.2. Table Row Extraction

The epidemiological tables in SIREVA-SUS concentrate most occurrences of METRICA_EPI and SOROTIPO, categories under-represented in narrative text. A **pdfplumber**-based module extracts these rows in four steps: (1) discarding tables with rotated text (*garbling*); (2) header/data separation by proportion of numeric cells; (3) synthesis of each data row in the format `[caption] | column: value | ...`, making it

self-contained out of context; (4) discarding subtotal and total rows. The module produced **1,956 table rows** in `data/extracted_tables/`, following the same JSONL schema as narrative text with an additional `type: "table"` field.

4.3. Silver Annotation via Few-Shot Prompting

The pipeline adopts **silver annotation** with **GPT-4o-mini** [OpenAI 2024] (temperature 0, output restricted to `json_object`).[1] The system prompt describes the eight categories with a definition and three examples each, plus explicit boundary rules. The few-shot history includes three (sentence, JSON) pairs covering the most frequent patterns: pathogen+serotype, age stratification+laboratory method, and location+temporal period. The model returns spans in the format `{"texto": "...", "tipo": "CATEGORY"}`, which we convert to the IOB2 scheme by character-level alignment. For example, in *"taxa de incidência de meningite por Neisseria meningitidis sorogrupo B"*, the alignment recovers the spans *taxa de incidência* (METRICA_EPI), *meningite* (MANIFESTACAO_CLINICA), *Neisseria meningitidis* (PATOGENO), and *sorogrupo B* (SOROTIPO). Silver annotation covered all 22 documents, producing **4,159 identified entity mentions**.

4.4. Gold Standard Construction

To evaluate the quality of the silver annotation and establish a reliable reference set, the pipeline selects a stratified corpus sample for manual review.

Sampling: narrative text. Sampling is stratified by document, totalling **180 sentences** distributed across the 12 SIREVA-SUS reports. Stratification ensures that all temporal periods and epidemiological patterns of the 2013–2024 historical series are represented in the gold set.

Sampling: table rows. A second gold set was built from silver-annotated table rows, following the same stratified sampling procedure by report, with **100 rows** distributed proportionally across the 12 series years. This set is essential to cover METRICA_EPI, METODO_LAB, and FAIXA_ETARIA in tabular context, categories underrepresented or absent in narrative text.

Manual review. Both sets were exported to **Label Studio** [Tkachenko et al. 2020] in JSON format with silver pre-annotations and reviewed against the tagset guidelines (Section 3). Ambiguous cases were resolved by consulting the guidelines. The reference gold set corresponds to the first annotator's review; where the second annotator's independent review (see the agreement paragraph below) diverged, disagreements were reconciled through discussion in light of the guidelines, and the agreed decision was recorded as the final gold label. For table rows, each numeric value associated with an epidemiological metric was individually annotated as METRICA_EPI (see Section 3.1).

Final format. The text gold set is stored in `data/gold/gold_sample.jsonl` and the table set in

[1]Reproducibility: silver annotation used the `gpt-4o-mini` model; the LLM comparison used `gemini-2.5-flash` (Google AI Studio, accessed July 2026) with the same prompt and temperature 0; the supervised baseline fine-tunes BERTimbau base (110M parameters) for 4 epochs (learning rate 3×10^{-5}, batch 16, max sequence length 192, seed 42).

`data/gold/gold_tables.jsonl`, both following the same field schema as the silver records (`tokens`, `bio_tags`, and `spans`), ensuring direct compatibility with the evaluation module.

Inter-annotator agreement. Guideline consistency was verified by a second domain expert annotator who independently reviewed the same 180 text sentences. Cohen's coefficient [Landis and Koch 1977] over IOB2 labels reached κ = **0.76** (*substantial agreement*), with 90.5% token-level agreement and a symmetric span F_1 of 0.78. Table 2 details the results per category. PATOGENO achieved the highest κ (0.88), benefiting from the precision of Latin scientific nomenclature. MANIFESTACAO_CLINICA (κ = 0.36), LOCAL (κ = 0.25), and METODO_LAB (κ = 0.21) showed lower values: the second annotator identified, respectively, 48%, 11%, and 20% more mentions in these categories, indicating annotation boundary ambiguity that warrants guideline refinement.

Table 2. Inter-annotator agreement (IAA) per category: Cohen's κ (IOB2 tokens) and symmetric span F_1.

Category	κ	F_1
PATOGENO	0.88	0.94
METRICA_EPI	0.73	0.84
SOROTIPO	0.67	0.57
FAIXA_ETARIA	0.65	0.85
PERIODO_TEMPORAL	0.45	0.71
MANIFESTACAO_CLINICA	0.36	0.69
LOCAL	0.25	0.71
METODO_LAB	0.21	0.55
Overall	**0.76**	**0.78**

Boundary cases and guideline decisions. The three lowest-agreement categories share one source of disagreement: the line between a named entity and a nearby but different concept. For LOCAL (κ = 0.25), the split is place versus institution. The guidelines annotate explicit toponyms and exclude institution names, but acronyms such as *IAL* (Instituto Adolfo Lutz) fall in between when they stand for the geographic node of the surveillance laboratory network. We annotate them as LOCAL only in that locative use, and leave them out when they name the institution as an organization. For METODO_LAB (κ = 0.21), the line is a named technique versus a generic reference: *cultura* and *sorotipagem molecular* are annotated, while a bare *exame* or *laboratório* is not. For MANIFESTACAO_CLINICA (κ = 0.36), the question is how much of a multiword condition to include: we annotate the full named condition (*meningite bacteriana* as a single span) and established domain labels such as *doença invasiva*, but exclude a bare *doença* or *infecção*. These decisions are now recorded as explicit boundary rules in the released guidelines, so that future annotation rounds can apply them from the start.

5. The EpiCorpus-BR

EpiCorpus-BR is built from 12 annual SIREVA-SUS reports (2013–2024) and 10 complementary epidemiological surveillance documents, totalling 22 documents. The SIREVA-SUS sub-corpus totals **3,343 sentences** and **19,762 tokens**, with a mean of 5.9 tokens

per sentence. The low mean reflects short text fragments from captions, lists, and section headings that passed the narrative prose filter, coexisting with full epidemiological sentences. Sentences are spread fairly evenly across the 2013–2024 series, and the 2023 and 2024 reports concentrate the highest volumes (385 and 370 sentences) due to the expansion of surveillance scope during that period.

5.1. Entity Distribution

Of the 3,343 sentences in the SIREVA-SUS silver set, **772 (23.1%)** contain at least one annotated entity, totalling **1,139 mentions** distributed across seven of the eight tagset categories. Table 3 presents the distribution by category in the SIREVA-SUS silver set, alongside the gold-standard counts.

Table 3. Mention distribution per category, in the SIREVA-SUS silver set (3,343 sentences) and in the gold standard (280 units: 180 text sentences + 100 table rows).

	Silver (SIREVA)		Gold (280 units)		
Category	**Ment.**	**%**	**Text**	**Tab.**	**Total**
LOCAL	592	52.0	154	0	154
PATOGENO	379	33.3	54	2	56
FAIXA_ETARIA	57	5.0	46	116	162
METODO_LAB	54	4.7	38	54	92
SOROTIPO	24	2.1	21	0	21
PERIODO_TEMPORAL	18	1.6	5	2	7
MANIFESTACAO_CLINICA	15	1.3	6	31	37
METRICA_EPI	0	0.0	22	638	660
Total	**1,139**	**100.0**	**346**	**843**	**1,189**

LOCAL and PATOGENO account for 85.3% of mentions, consistent with the structure of SIREVA-SUS reports, where surveillance is organized primarily by etiological agent and by federal unit. The absence of METRICA_EPI mentions in narrative text was confirmed as structural: the domain's quantitative metrics (such as strain counts by age group and susceptibility percentages) are concentrated in epidemiological tables, inaccessible to the narrative prose filter. The tabular extraction described in Section 4.2 covers this content; its 100-row gold standard is evaluated in Section 6.

The gold-standard columns of Table 3 confirm the same tabular concentration: METRICA_EPI and FAIXA_ETARIA are predominantly represented in tables, while LOCAL and PATOGENO concentrate in narrative text, reflecting the structure of SIREVA-SUS reports.

5.2. Complementary Documents

Beyond the SIREVA-SUS reports, the corpus incorporates 10 complementary surveillance documents: three *Electronic Epidemiological Bulletins*, five numbered *Epidemiological Bulletins*, and two *Meningitis Advisories*, all from the Ministry of Health or the Health Surveillance Secretariat. They add **3,015 sentences** and **3,020 silver mentions**, and their distinct structure and terminology broaden the linguistic diversity of the corpus; the same extraction and silver-annotation pipeline was applied to them. Their entity

density is much higher than in the SIREVA-SUS sub-corpus, **1.00 mention per sentence** against **0.34**, because the thematic bulletins describe specific cases and outbreaks rather than long-term surveillance prose. The two meningitis advisories are the densest documents, with 56.9% and 48.9% of their sentences carrying at least one entity.

6. Results

The evaluation compares silver annotations against the gold standard using **seqeval** [Nakayama 2018] in strict matching (IOB2) mode. Table 4 presents F_1 per category for the three evaluated configurations: narrative text (180 sentences, 346 mentions), table rows (100 rows, 843 mentions), and combined corpus (280 units, 1,189 mentions).

Table 4. Per-category F_1 (strict seqeval, IOB2): GPT-4o-mini silver annotation in the three configurations, a second-LLM annotator (Gemini 2.5 Flash, same prompt), and the BERTimbau baseline (trained on silver, combined-gold test, Section 6.2). N = support in the combined corpus.

	GPT-4o-mini (silver)			Gemini	BERTimbau	
Category	**Text**	**Tab.**	**Comb.**	**Comb.**	**Comb.**	**N**
PATOGENO	0.97	0.00	0.95	0.96	0.98	56
SOROTIPO	0.92	0.00	0.92	0.62	0.76	21
PERIODO_TEMPORAL	0.89	0.00	0.73	0.62	0.77	7
METODO_LAB	0.88	0.62	0.71	0.65	0.63	92
FAIXA_ETARIA	0.85	0.25	0.43	0.45	0.43	162
LOCAL	0.61	0.00	0.59	0.26	0.30	154
MANIFESTACAO_CLINICA	0.50	0.59	0.58	0.68	0.51	37
METRICA_EPI	0.00	0.91	0.89	0.94	0.89	660
Micro avg	0.73	0.79	0.77	0.77	0.73	1189
Macro avg	0.70	0.34	0.73	0.65	0.66	1189

In narrative text, PATOGENO and SOROTIPO exceed $F_1 = 0.92$, helped by their consistent Latin nomenclature. METRICA_EPI scores zero in text, because its metrics are almost entirely tabular, yet it reaches 0.91 on table rows. FAIXA_ETARIA drops from 0.85 in text to 0.25 in tables, and LOCAL shows only moderate precision (0.57). The combined silver micro-F_1 is 0.77. Section 6.1 analyses the error patterns behind these numbers.

Robustness across LLM annotators. To check whether the annotation quality is specific to GPT-4o-mini, we ran the same prompt with a second, more recent model, Gemini 2.5 Flash, over the same 280 gold units. It reaches the *same* combined micro-F_1 of 0.77 (Table 4). This suggests that the silver annotation is robust to the choice of LLM and is not an artefact of a single model. The two models redistribute the errors rather than removing them. Gemini is stronger on the tabular split (0.86 vs. 0.79), on METRICA_EPI (0.94), and on MANIFESTACAO_CLINICA (0.68), but weaker on SOROTIPO (0.62). Like GPT-4o-mini, it still collapses on the ambiguous LOCAL category (0.26), which shows that this category is a guideline problem, not a model one.

6.1. Error Analysis

Manual inspection of the disagreements between the silver annotation and the gold standard shows that the errors are not spread evenly across categories. They concentrate in

three systematic patterns, and all three are **omissions** (false negatives) rather than confusion between categories. Table 5 gives one representative example of each pattern. Below, we explain what each example is meant to show and how much F_1 it costs (Table 4).

(E1) Omitted numeric values in narrative text. This pattern alone explains the zero F_1 of METRICA_EPI in text. The `(n=166)` example isolates the failure: the model reads the surrounding context correctly, since it annotates *cultura* as METODO_LAB, yet it does not extend the annotation to the adjacent count that the guidelines mark as METRICA_EPI. The problem is therefore recall of the numeric span, not a wrong label, because the same model annotates these counts correctly when they appear alone in table cells ($F_1 = 0.91$).

(E2) Institutional acronyms used as locations. This pattern explains the moderate score of LOCAL (0.59). The *IAL* (Instituto Adolfo Lutz) example shows the boundary the model misses: it captures explicit place names such as *Brasil* and *São Paulo*, but not domain acronyms that stand for a place in context, here the location of the surveillance laboratory network. This is the same LOCAL boundary that drove the low inter-annotator agreement for the category (Section 4.4), which confirms it is a guideline issue, not a model one.

(E3) Compact age formats in tables. This pattern explains why FAIXA_ETARIA falls from $F_1 = 0.85$ in text to 0.25 in tables. The example, an abbreviated age band such as `<1 ano` in a table header, shows that the failure is one of surface form, not of concept: the few-shot examples cover spelled-out ranges (*"menor que 12 meses"*) but not the short notation of tabular headers, so the model omits them or misplaces their boundaries.

Together, the three patterns show that the errors come from domain-specific surface forms that are under-represented in the few-shot prompt, not from conceptual ambiguity between categories. Two routes address this: richer few-shot examples and guidelines, and the supervised model of Section 6.2, which learns these surface forms from the distribution of the silver corpus.

Table 5. The three systematic silver-annotation errors. Each row shows the token that should have been annotated, its gold label, and the silver output. O means the model left the token unannotated, so all three patterns are omissions (false negatives).

Type	Token (in context)	Gold label	Silver
E1: omitted value	`(n=166)` in "…por cultura `(n=166)`"	B-METRICA_EPI	O
E2: locative acronym	*IAL* in "o *IAL* monitora … os estados"	B-LOCAL	O
E3: compact age	`<1 ano` in a tabular age-group header	B-FAIXA_ETARIA	O

6.2. A Supervised Baseline under Silver Supervision

To assess whether the corpus can train a dedicated Portuguese NER model, and not only measure the LLM annotator, we fine-tune BERTimbau [Souza et al. 2020] (base, cased)

as a token classifier under *silver (distant) supervision*. The model is trained on the silver-annotated corpus (narrative text and table rows) with the 280 gold units removed to prevent test leakage (8,034 training units), and evaluated on the gold standard using the same seqeval strict-IOB2 protocol as in Section 6. Word-level labels are aligned to the first subword token; we train for 4 epochs with learning rate 3×10^{-5}, batch size 16, maximum sequence length 192, and a fixed seed for reproducibility. This setup directly tests the downstream utility of the corpus while keeping the gold set untouched during training.

The rightmost column of Table 4 reports the results. Trained only on the silver corpus, BERTimbau reaches a combined micro-F_1 of **0.73**, close to the 0.77 of the silver annotation it learned from. This shows that the corpus is usable to train a dedicated model and that its silver labels are consistent enough to be learned. The model is strongest where the annotation is terminologically regular, namely PATOGENO (0.98, Latin nomenclature) and METRICA_EPI (0.89, and 0.90 on table rows alone). It also matches the silver micro-F_1 on the tabular split (0.79), which confirms that the tabular sub-corpus transfers well to a trained model. Performance drops on the ambiguous categories, with LOCAL falling to 0.30. This mirrors error type E2 (Section 6.1) and confirms that the bottleneck is institutional and locative ambiguity, not model capacity. Because the model is supervised by the silver labels themselves, it cannot exceed the silver ceiling by much. Reaching 0.73 without ever seeing the gold is therefore a lower bound, and a larger, gold-supervised training set should raise it.

7. Limitations

The gold standard covers only 3.4% of the corpus (280 of 8,314 units); although stratified, the evaluation reflects this subset, and the silver annotation comes from a single model (GPT-4o-mini). The IAA (κ = 0.76, Section 4.4) used two annotators who both started from the silver pre-annotations, which may inflate it relative to annotation from scratch, and the low κ for LOCAL, METODO_LAB, and MANIFESTACAO_CLINICA reflects residual boundary ambiguity. More importantly, the review corrects and reclassifies the model's spans but does not systematically add entities the model never proposed, so the gold standard measures the *precision* of the silver annotation reliably while its true *recall* may be overestimated.

8. Conclusion

This paper presented EpiCorpus-BR, to the best of our knowledge the first public NER corpus for epidemiological surveillance documents in Portuguese. It brings together 8,314 textual units from 22 SIREVA-SUS and Ministry of Health documents, a tagset of eight specialized categories absent from existing Portuguese resources [Oliveira et al. 2022], and a gold standard of 280 manually reviewed units. The combined evaluation (micro-F_1 = 0.77) showed that METRICA_EPI, at zero F_1 in narrative text, reaches 0.89 in tables, so the tabular sub-corpus is essential for a representative evaluation of the domain. A BERTimbau baseline trained on the silver corpus, with the gold held out, reached micro-F_1 = 0.73, confirming that the corpus supports training a dedicated NER model.

As future work, we plan to integrate the three annotation layers already built for SIREVA-SUS documents, Universal Dependencies [Freitas et al. 2026], named entities (this work), and RAG-based semantic retrieval [Freitas et al. 2025], into a unified analysis pipeline.

Availability

EpiCorpus-BR and the pipeline code are publicly available at `https://github.com/ChristianSF/EpiCorpus-BR`, in JSONL format (fields `tokens`, `bio_tags` in IOB2, and `spans`) with documentation and the extraction, annotation, and evaluation modules. The corpus is released under CC BY 4.0 and the code under MIT.

Acknowledgements

The authors thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support.

Ethics Statement

The documents used are official publications made available by the Oswaldo Cruz Foundation, the Ministry of Health, and the Health Surveillance Secretariat without access restrictions. The corpus contains no patient personal data, consisting exclusively of epidemiological text aggregated at the population level. During the preparation of this article, generative artificial intelligence tools were used to support language and structure review; the authors reviewed and are responsible for the final content of the publication.

References

Albuquerque, H. O., Souza, E., Gomes, C., Pinto, M. H. d. C., Filho, R. P. S., Costa, R., Lopes, V. T. d. M., da Silva, N. F. F., de Carvalho, A. C., and Sampaio, A. d. O. (2023). Named entity recognition: a survey for the Portuguese language. *Procesamiento del Lenguaje Natural*, 70:171–185.

Brito, M., Pinheiro, V., Furtado, V., Monteiro Neto, J. A., Bomfim, F. d. C. J., da Costa, A. C. F., and Silveira, R. (2023). CDJUR-BR – uma coleção dourada do judiciário brasileiro com entidades nomeadas refinadas. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 177–186, Porto Alegre. Sociedade Brasileira de Computação.

Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A. T., and Joty, S. (2024). Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 8764–8786.

Doğan, R. I., Leaman, R., and Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

DS4SD Team (2024). Docling technical report. *arXiv preprint arXiv:2408.09869*.

Freitas, C., Rabonato, R. T., and Berton, L. (2025). Enhancing epidemiological insights with RAG for SIREVA-SUS reports. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 1364–1375, Fortaleza/CE, Brazil. Sociedade Brasileira de Computação.

Freitas, C., Real, L., Berton, L., and Paiva, V. d. (2026). Towards a Universal Dependencies corpus for Portuguese epidemiological reports. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 228–237, Salvador, Brazil. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Moreira, H., Silva, P. F. d., Vieira, R., and Moreira, V. (2025). PetroGeoNER: um conjunto de dados refinado e unificado para NER no domínio de petróleo e gás. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 259–271, Porto Alegre. Sociedade Brasileira de Computação.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nakayama, H. (2018). seqeval: A python framework for sequence labeling evaluation. `https://github.com/chakki-works/seqeval`.

Oliveira, L. E. S. e., Peters, A. C., da Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Al Hasan, S., and Moro, C. M. C. (2022). SemClinBr – a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, 13(1).

OpenAI (2024). GPT-4o mini: Advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`.

Piai, L., Di-Felippo, A., and Romano, N. T. (2025). Entidades nomeadas em tweets do mercado de ações: uma anotação refinada e com motivação linguística. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 654–663, Porto Alegre. Sociedade Brasileira de Computação.

Pontes, M. F., Pedrosa, R. C., Lopes, P. H., and Luz, E. J. (2024). Avaliando aprendizagem federada com criptografia homomórfica para reconhecimento de entidades médicas nomeadas utilizando modelos compactos de BERT. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 48–56, Porto Alegre. Sociedade Brasileira de Computação.

Ramos, F. d. O., Marvila, C. G., and Colombo, C. d. S. (2026). Uma abordagem híbrida para reconhecimento de entidades nomeadas em relatos de casos clínicos. In *Anais do XXII Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 216–233, Porto Alegre. Sociedade Brasileira de Computação.

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., Oliveira, L. E. S. e., Copara, J., Gumiel, Y. B., Oliveira, L. F. A. d., Paraiso, E. C., Teodoro, D., and Barra, C. M. C. M. (2020). BioBERTpt – a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72. Association for Computational Linguistics.

Secretaria de Vigilância em Saúde (2024). Vigilância epidemiológica. Ministério da Saúde do Brasil. Disponível em: `https://www.gov.br/saude/pt-br/composicao/svs`. Acesso em: mai. 2026.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Intelligent Systems (BRACIS 2020)*, pages 403–417. Springer.

Tkachenko, M., Malyuk, M., Holmberg, N., Liubimov, N., and Strulov, A. (2020). Label Studio: Data labeling software. `https://github.com/heartexlabs/label-studio`.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.