A Dataset of Brazilian Portuguese Clinical Notes for Anaphylaxis Detection

Matheus Machado and Vinícius Vanzin and Dilvan Moreira
Institute of Mathematics and Computer Science
University of São Paulo
matheusmatos@usp.br, vinicius.vanzin@usp.br, dilvan@icmc.usp.br

Luis Felipe Ensina and Fábio Lario
Dep. of Allergy and Clinical Immunology
Hospital Sírio-Libanês
100alergia@gmail.com, fabio.clario@hsl.org.br

Abstract

Anaphylaxis is an acute, potentially life-threatening allergic reaction that requires rapid recognition in clinical settings. Natural language processing (NLP) approaches for automatic detection of anaphylaxis in clinical narratives can support large-scale analysis of health records and retrospective clinical research. However, such approaches depend on high-quality labeled corpora, and resources for Portuguese remain scarce. This paper introduces a corpus of Brazilian Portuguese clinical notes annotated by domain specialists for the presence or absence of anaphylaxis. The dataset comprises 969 clinical narratives drawn from three sources: clinician-authored synthetic clinical notes designed to represent realistic scenarios, case reports from the medical literature rewritten into note-like format by specialists, and a subset of de-identified notes from the publicly available SemClinBr corpus. All texts were reviewed and labeled by allergists using established clinical diagnostic criteria, and the corpus reflects realistic prevalence conditions, with approximately 5% positive cases. We describe the corpus design, data sources, annotation methodology, and composition, discuss potential research applications, and address ethical considerations. The corpus is intended as a reusable resource for Portuguese clinical NLP, supporting future work on document classification, information extraction, and language modeling in the medical domain.

1 Introduction

Anaphylaxis is an acute, potentially life-threatening systemic hypersensitivity reaction that may progress rapidly without prompt treatment. In clinical practice, its recognition relies on patient history, physical examination, and judgment guided by standardized diagnostic criteria. The National Institute of Allergy and Infectious Disease/Food Allergy and Anaphylaxis Network (NIAID/FAAN) criteria define anaphylaxis through symptom patterns involving acute mucocutaneous manifestations with respiratory or cardiovascular compromise, rapid multi-system involvement after exposure to a likely allergen, or hypotension after exposure to a known allergen (Dribin et al., 2023). The World Allergy Organization (WAO) 2020 update simplifies these criteria and highlights additional presentations, including severe gastrointestinal symptoms and cases dominated by respiratory or cardiovascular signs (Cardona et al., 2020). Despite these guidelines, diagnosis remains challenging because clinical presentations vary across patients and episodes, and documentation may be incomplete or heterogeneous in real-world records (Simons et al., 2011).

Automatic detection of anaphylaxis in clinical narratives could support research on case identification at scale and facilitate retrospective analyses of electronic health records. However, supervised approaches depend on annotated corpora, and clinical text resources remain limited in languages other than English. In Brazilian Portuguese, SemClinBr is one of the few publicly available clinical corpora; it comprises 1,000 de-identified notes annotated with 65,117 entities and 11,263 relations (Oliveira et al., 2022). While SemClinBr is valuable for tasks such as named-entity recognition and relation extraction, it does not provide document-level labels for specific conditions such as anaphylaxis. More broadly, access to clinical narratives is constrained by privacy and governance requirements, motivating increased attention to dataset transparency

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 78–87

through documentation practices such as dataset statements and datasheets (Bender and Friedman, 2018; Gebru et al., 2021). Synthetic clinical text has also been explored as a complementary strategy to mitigate data scarcity and reduce privacy risks by generating plausible narratives that do not correspond to identifiable individuals (Mendes et al., 2025).

In this work, we introduce a dataset of Brazilian Portuguese clinical notes annotated for the presence or absence of anaphylaxis. The contribution is dataset-centric: we focus on documenting the corpus design, data sources, annotation criteria, and corpus characteristics to support reproducible research on Portuguese clinical NLP. The dataset combines expert-authored synthetic clinical notes, adapted case reports rewritten into a note-like format, and de-identified notes sampled from a public clinical corpus. We further discuss limitations and ethical considerations relevant to releasing and using a clinical text resource. The corpus is available publicly at Hugging Face: matos1012/brazilian-portuguese-anaphylaxis.

2 Related Work

Research on automatic processing of clinical text encompasses the development of linguistic resources, the application of NLP methods to clinically relevant phenomena, and approaches to mitigate data scarcity and privacy constraints. While these areas are well established for English-language clinical NLP, corresponding resources and studies remain limited for Portuguese. In this context, the availability of clearly documented datasets is essential for enabling reproducible and comparable research. This section reviews work most relevant to the present study, focusing on clinical corpora in Portuguese, NLP-based approaches to anaphylaxis detection, and the use of synthetic data in clinical NLP.

2.1 Clinical corpora in Portuguese

Publicly available resources for Portuguese clinical NLP are historically scarce, a disparity that becomes evident when compared to mature English-language benchmarks. Examples of large-scale datasets include the MIMIC-IV-Note collection, which aggregates over 331,000 de-identified discharge summaries from the Beth Israel Deaconess Medical Center (Johnson et al., 2023), and the gold-standard annotated corpora from the n2c2 (formerly i2b2) shared tasks, typically comprising 500 to 1,300 specialized notes per task (Wang et al., 2020).

In the Portuguese clinical NLP landscape, early efforts focused on specialized domains, such as the publicly-available CLINpt corpus (Lopes et al., 2019), which provided 281 neurology case reports from medical journals. Schneider et al. (2020) later utilized a private collection of 3.8 million sentences derived from Brazilian hospital EHRs and scientific abstracts to develop the BioBERTpt model. Significant progress was made with SemClinBr (Oliveira et al., 2022), the first large-scale multi-institutional corpus for Brazilian Portuguese, offering 1,000 notes semantically annotated with over 65,000 entities. Concurrently, the credentialed-access BRATECA dataset (Dias and Ulbrich, 2022) provided a leap in volume, aggregating over 2.5 million free-text notes from ten Brazilian hospitals, while da Rocha et al. (2023) introduced the private HCFMB corpus, consisting of 1,200 clinical texts derived from 30,000 records at a university hospital.

More recent contributions include AnonyMED-BR (Schiezaro et al., 2025), a publicly-available dataset of 2,962 records specifically designed for anonymization tasks, and the anonymized sepsis summaries from da Silva and Pazin-Filho (2025), featuring 200 long-form discharge summaries from a tertiary hospital. MedPT (Färber et al., 2025) represents a consumer health-oriented dataset, containing over 384,000 authentic question-answer pairs from patient-doctor interactions. Clinical NLP research in Portuguese frequently relies on private corpora, limited-access resources, or translated datasets, underscoring the need for additional publicly documented datasets targeting clinically meaningful phenomena.

2.2 Anaphylaxis detection using NLP

Anaphylaxis detection in clinical text is a complex phenotyping task that has evolved from rule-based systems to the use of large language models (LLMs). Early work by Botsis et al. (2012) utilized the VAERS dataset to extract features from thousands of vaccine safety reports using semantic text mining. Botsis and Ball (2013) furthered this by automating case definitions through literature-based reasoning, relying on mapping synonyms to Brighton criteria. Walsh et al. (2013) highlighted the limitations of structured data using a critical benchmark of 122 potential anaphylaxis events

across eight healthcare settings. Segura-Bedmar et al. (2018) shifted toward large-scale detection, applying convolutional neural networks and classical classifiers to a collection of over 219,000 clinical records.

Recent studies have emphasized the integration of clinical notes with structured records or the application of instruction-tuned models. Yu et al. (2020) developed algorithms for the Vaccine Safety Datalink, utilizing 311 potential cases to identify vaccine-related anaphylaxis with high specificity. Lo et al. (2022) addressed allergy reconciliation by detecting discrepancies in encounter notes from the Mass General Brigham healthcare system. Carrell et al. (2023) engineered a set of NLP-derived symptom and criteria covariates using data from 516 patients across the Kaiser Permanente network. In the Portuguese context, Machado et al. (2024) assessed several large language models on a smaller set of annotated clinical reports, focusing on model behavior under limited data conditions. Finally, Ensina et al. (2025) evaluated LLMs on a corpus of Portuguese medical texts annotated by physicians for anaphylaxis, exploring different prompting strategies and reporting high classification performance.

The present work differs in scope by focusing on the dataset itself. Rather than reporting additional model results, we document the construction, annotation methodology, and characteristics of the corpus to support reuse and transparent evaluation.

2.3 Synthetic clinical data

Synthetic clinical data generation has emerged as a key strategy to circumvent privacy constraints and data scarcity. Early methods, such as SynthNotes (Begoli et al., 2018), utilized semantic templates to generate note-like text from original narratives. Li et al. (2021) evaluated the utility of 500 notes generated by text generation models for downstream entity recognition. The GReaT framework (Borisov et al., 2022) introduced a method for generating realistic tabular data by treating records as sequences, and Moser et al. (2024) proposed a multi-stage pipeline using large language models to generate structured patient-physician interactions in emergency medicine.

Current research focuses on high-fidelity generation and rigorous evaluation using large language models. Synthetic4Health (Ren et al., 2025) introduced a mask-and-generate framework for creating diverse, de-identified clinical letters. SynthMedic (Grazhdanski et al., 2025) proposed generating discharge summaries grounded in standard medical references (Merck Manuals) to ensure factual consistency without requiring access to real patient records. Meoni et al. (2025) framed synthetic text generation as an intermediate step for local model training, evaluating their approach on the MIMIC-III dataset. Sarkar et al. (2025) proposed a hybrid methodology, which combines de-identification with LLM filling to safely share clinical notes. Evaluation frameworks have also been standardized through toolkits like SynthTextEval (Ramesh et al., 2025), which assesses utility and privacy in health-related text. Finally, the MedGen model (Wang et al., 2025) scaled synthetic generation to multimodal domains, achieving strong performance in medical video generation.

These lines of work motivate documenting synthetic data provenance, grounding sources, and validation procedures, as such design choices influence the linguistic and clinical properties of the generated narratives. While the aforementioned literature relies heavily on automated generation, the synthetic narratives in our corpus were manually authored by domain specialists. This manual approach was chosen to strictly guarantee clinical fidelity, ensure adherence to specific diagnostic criteria without the risk of model hallucination, and provide a high-quality gold standard. In the corpus presented in this paper, these manually authored synthetic notes are integrated with non-synthetic material (adapted case reports and de-identified notes sampled from a public clinical corpus) to support analyses that distinguish between generated and naturally occurring documentation styles.

3 Corpus Design and Construction

This section describes the design principles and construction process of the proposed corpus. The dataset was created to support research on automatic detection of anaphylaxis in Portuguese clinical narratives, while addressing recurring challenges in clinical NLP, including data scarcity, class imbalance, and privacy constraints. To this end, we combined clinical texts from complementary sources (synthetic narratives, adapted case reports from the literature, and de-identified notes from a public clinical corpus), seeking to balance linguistic realism, ethical considerations, and reproducibility. In the context of this corpus, "synthetic" refers to fictional, representative clinical narratives

manually authored or edited by domain specialists (allergists) to simulate realistic patient encounters, rather than text generated automatically by software or language models. All texts, regardless of source, were subsequently reviewed and annotated by the specialists according to standardized clinical criteria. The following subsections detail the data sources, preprocessing and de-identification steps, and the annotation guidelines adopted.

3.1 Data Sources

The corpus comprises 969 short clinical narratives in Brazilian Portuguese, organized into three clinically motivated categories to support structured analysis and annotation consistency:

Synthetic cases (75 narratives). This subset consists of anonymized clinical narratives authored by domain specialists to mimic real-world medical records. From these, 35 notes have a confirmed diagnosis of anaphylaxis, and 40 were labeled as negative. These texts constitute the largest group of positive examples in the corpus.

Case reports from the literature (24 narratives). These narratives were adapted from published medical case reports. Thirteen texts describe confirmed cases of anaphylaxis, while eleven report other clinical conditions with overlapping or potentially confounding symptoms. All case reports were rewritten into a concise, note-like format to resemble electronic medical records while preserving clinically relevant information and removing any identifying details.

SemClinBr cases (870 narratives). The remaining texts were sampled from SemClinBr, a publicly available corpus of de-identified Brazilian Portuguese clinical narratives covering a wide range of medical conditions (Oliveira et al., 2022). To ensure sufficient informational content, only texts longer than 200 characters were selected. All SemClinBr narratives were independently reviewed by allergists; although allergic manifestations were observed in some cases, none satisfied the diagnostic criteria for anaphylaxis and were therefore labeled as negative.

Among the negative cases from the synthetic and case reports sources, 35 of them were considered as differential diagnosis cases. These are clinically challenging cases that present symptoms similar to anaphylaxis but correspond to alternative diagnoses. Ten narratives were adapted from published case reports, and twenty-five were derived from synthetic medical records. None of these cases met the diagnostic criteria for anaphylaxis, and all were labeled as negative after specialist review.

Overall, the corpus contains 48 positive cases of anaphylaxis and 921 negative cases. Positive cases originate exclusively from the synthetic and literature-derived subsets. The proportion of positive cases was deliberately fixed at approximately 5% to reflect prevalence estimates reported for emergency care settings (Cardona et al., 2020). All synthetic, anonymized, and adapted narratives were reviewed to ensure that they did not correspond to identifiable patient records. Table 1 presents representative excerpts from the corpus, sampled from each data source (English translations provided for convenience, dataset contains only Brazilian Portuguese text).

3.2 Preprocessing and De-identification

All texts were normalized using a preprocessing pipeline that removed formatting artifacts and corrected obvious spelling errors. Domain-specific terminology and common clinical abbreviations were preserved to maintain linguistic authenticity. No personally identifiable information (PII) was included in the synthetic or adapted notes. The SemClinBr texts were already de-identified in accordance with Brazilian data-protection regulations. We did not observe names, dates, or institutional identifiers in the final corpus. Given the intended public release of the dataset, these measures were deemed sufficient to mitigate privacy risks.

3.3 Annotation Guidelines

This subsection outlines the clinical criteria and procedures used to annotate the corpus for the presence or absence of anaphylaxis. The annotation guidelines were designed to operationalize established diagnostic definitions in a manner compatible with short, heterogeneous clinical narratives, enabling consistent labeling across all data sources while preserving clinical interpretability.

3.3.1 Clinical criteria for anaphylaxis

Annotation was guided by the NIAID/FAAN clinical criteria for anaphylaxis. A narrative was labeled as positive if at least one of the following conditions was satisfied (Cardona et al., 2020; Dribin et al., 2023):

- **Skin or mucosal involvement with respiratory or cardiovascular compromise:** Acute onset of urticaria, angioedema, or mucosal

Source	Label	Narrative text (Brazilian Portuguese)	English translation
Synthetic	Positive	Paciente refere prurido intenso pelo corpo, coçar, inchaços (pelotas) se formando no corpo (...) após sentir picadura/ferroada de inseto na perna (...)	Patient reports intense itching over the body, scratching, swellings (welts) forming on the body (...) after feeling an insect bite/sting on the leg (...)
Case report	Positive	Mal-estar geral, tontura, sudorese após ingesta de comprimido de omeprazol para dor epigástrica (...) observou-se edema da face e disartria (...)	General malaise, dizziness, sweating after ingestion of an omeprazole table for epigastric pain (...) facial edema and dysarthria were observed (...)
SemClinBr	Negative	Às 02:45h: encontra-se consciente (...) mantém acesso venoso periférico em MSE permeável salinizado no momento, apresenta rash cutâneo corporal + prurido (...)	At 02:45h: is conscious (...) maintains peripheral venous access in LUE [Left Upper Extremity], patent and saline-locked at the moment, presents with generalized skin rash + pruritus (...)
Differential diagnosis (synthetic)	Negative	Paciente (...) com quadro de edema de língua e lábios (deformante) com duração de 72h (...) PA 130/90mmHg, Sat O2 94%aa (...) Pele sem lesões (...)	Patient (...) with presentation of tongue and lip edema (deforming) lasting 72h (...) BP 130/90mmHg, O2 Sat 94% on room air (...) Skin without lesions (...)

Table 1: Excerpts from the corpus, sampled from each data source.

swelling accompanied by respiratory compromise (e.g., dyspnea, wheezing, stridor, hypoxemia) or by reduced blood pressure or syncope.

- **Multi-system involvement after exposure to a likely allergen:** Rapid involvement of at least two organ systems (mucocutaneous, respiratory, cardiovascular, or gastrointestinal) following exposure to a likely allergen, such as urticaria with vomiting or bronchospasm with hypotension.
- **Hypotension after exposure to a known allergen:** Sudden hypotension or syncope occurring after exposure to a known or highly probable allergen (Dribin et al., 2023).

The 2020 update by the World Allergy Organization recognizes that severe gastrointestinal symptoms or isolated respiratory or cardiovascular manifestations may also indicate anaphylaxis (Cardona et al., 2020). Annotators used these updates as supporting guidance but required that the core NIAID/FAAN criteria be met to assign a positive label. Negative labels were assigned when none of these criteria were satisfied or when the narrative clearly supported an alternative diagnosis.

3.3.2 Annotation procedure

Three board-certified allergists participated in annotating the 969 narratives, assigning a binary label indicating the presence or absence of anaphylaxis. Because the SemClinBr subset originates from a general clinical corpus with a low prior probability of anaphylaxis, each of these notes was reviewed by a single expert. Narratives from other sources were independently reviewed by at least two annotators. Annotators were provided only with the narrative text and did not have access to metadata regarding data source or patient context. Following independent annotation, cases of disagreement were discussed in consensus meetings. When disagreement persisted, a senior allergist acted as adjudicator.

3.3.3 Annotation format

Each entry in the released dataset contains the following fields:

- `note_id`: unique identifier (sequential integer);
- `text`: clinical narrative in Brazilian Portuguese;
- `source`: data source category (`synthetic`, `case_report`, or `semclinbr`);
- `label`: binary indicator of anaphylaxis presence (`1`) or absence (`0`);

The corpus is distributed as a single CSV file and may be converted to other structured formats, such as JSON, for convenience. For entries sourced from SemClinBr (source `semclinbr`), the `text` field contains the corresponding note ID rather than the full clinical narrative, as the original dataset requires credentialed access.

4 Corpus Analysis

This section characterizes the corpus in terms of basic composition and linguistic variability. We report high-level statistics on document length and lexical diversity, and describe how class labels and

writing styles differ across sources. The goal is to provide sufficient quantitative context for reproducibility while keeping the analysis focused on properties that affect downstream NLP use.

4.1 Document Length and Vocabulary

The corpus contains 969 clinical narratives with substantial variability in length, reflecting differences in writing conventions across sources. SemClinBr notes include both short chart-like entries and longer narratives, whereas synthetic notes tend to be more elaborated and explanatory. Using a simple regex-based tokenizer, the corpus contains approximately 129k tokens and 11.4k unique word forms, indicating a non-trivial lexical variety for a dataset of this size. Frequent terms and expressions are consistent with the clinical domain and include references to symptoms and interventions, e.g., “urticária” (hives), “edema” (edema), “dispneia” (dyspnea), “pressão arterial” (blood pressure), “epinefrina” (epinephrine), as well as common abbreviations (e.g., “PA” for blood pressure).

Table 2 summarizes the number of documents and class labels by source.

Source	Notes	Positive	Negative
Synthetic	75	35	40
Case reports	24	13	11
SemClinBr	870	0	870
Total	969	48	921

Table 2: Corpus composition by source and class label.

The “Synthetic” category aggregates expert-authored anaphylaxis cases and differential diagnosis cases derived from anonymized medical records.

4.2 Class Balance and Category Distribution

The dataset is deliberately imbalanced, with 48 positive cases of anaphylaxis (approximately 5%) and 921 negative cases, reflecting the rarity of the condition in routine clinical documentation (Cardona et al., 2020). Positive examples originate from the synthetic and literature-derived subsets, while SemClinBr contributes only negative examples. Importantly, all SemClinBr notes were manually reviewed by allergists, and although allergy-related content was observed in that subset, no cases meeting anaphylaxis criteria were identified. This design supports evaluation in settings where the target condition is rare and must be distinguished from both unrelated medical content and clinically similar presentations.

4.3 Linguistic Characteristics

The corpus exhibits systematic stylistic variation across sources, reflecting the different conditions of clinical documentation. De-identified notes from SemClinBr present a telegraphic structure, with an average sentence length of 9.8 words and a low lexical density of 0.26. This aligns with the prevalent use of institutional shorthand, abbreviations, and fragment syntax common in real-world charting, evidenced by an out-of-vocabulary (OOV) rate of 11.5%.

In contrast, adapted case reports and synthetic narratives present more complete, formal syntactic structures. Literature-derived reports average 23 words per sentence, while synthetic notes average 18.2. Both of these authored sources exhibit higher lexical density (approximately 0.46) and substantially lower OOV rates (1.7% and 2.5%, respectively). While the authored narratives explicitly describe termporal progression, triggers, and treatment responses, the SemClinBr notes compress this information into standard charting formats.

Explicit negation markers appear consistently across all subsets, ranging from 1.37 to 1.44 instances per 100 words. Positive cases are characterized by terms denoting temporal progressions and manifestations of the WAO criteria. Negative cases are weighted toward negation markers and vocabulary associated with alternative diagnoses. Standard clinical vocabulary such as "paciente" (patient) and "uso" (use) ranks highest across both classes. This heterogeneity makes the corpus suitable for studying domain adaptation and robustness to variation in clinical writing style.

5 Potential Applications

The annotated corpus enables a range of research applications in clinical natural language processing, particularly in scenarios involving rare-event detection and heterogeneous clinical narratives.

Document classification. The primary intended use of the corpus is the development and evaluation of models for automatic detection of anaphylaxis in free-text clinical notes.

The pronounced class imbalance reflects real-world conditions and makes the corpus suitable for studying methods for rare-event detection, cost-sensitive learning, and robustness under skewed

label distributions. Because the dataset mimics the low prevalence of rare clinical presentations, it is particularly well-suited for developing outlier detection algorithms, and evaluating the capabilities of pretrained models without the need for large-scale supervised training.

Information extraction. Beyond document-level classification, the corpus can be used to investigate the linguistic cues associated with anaphylaxis, including references to cutaneous manifestations, respiratory compromise, cardiovascular instability, and gastrointestinal symptoms. This supports the development of rule-based systems, feature-driven models, or explainable approaches that aim to align model decisions with established clinical criteria.

Language modeling and domain adaptation. The combination of synthetic narratives, adapted case reports, and de-identified real clinical notes provides a controlled setting for studying domain adaptation. Researchers may explore how language models trained on synthetic or literature-derived text generalize to authentic clinical documentation, and how mixed-source training affects performance in low-resource clinical domains.

Evaluation of large language models. Existing literature (Machado et al., 2024; Ensina et al., 2025) supports the application of this corpus as a benchmark for the evaluation of large language models, including GPT-3.5 and GPT-4. Possible extensions could systematically compare prompting strategies, zero-shot and few-shot settings, and domain-adapted models, in addition to analyzing error patterns in clinically challenging or ambiguous cases.

6 Conclusion

This paper introduced a dataset of 969 Brazilian Portuguese clinical narratives annotated for the presence or absence of anaphylaxis. The corpus integrates synthetic clinical notes, adapted case reports from the literature, and de-identified records from SemClinBr, with annotations grounded in established NIAID/FAAN and WAO diagnostic criteria. By combining heterogeneous text sources and explicitly documenting the annotation process, the dataset addresses a critical gap in publicly available clinical resources for Portuguese and supports research on rare-event detection in free-text medical narratives.

The detailed description of data provenance, labeling guidelines, corpus composition, and limitations aligns with current best practices for dataset transparency and responsible reuse. Rather than proposing new modeling approaches, this work aims to provide a well-characterized and reusable resource for the community. We expect the corpus to support reproducible studies in clinical NLP, including document classification, information extraction, and evaluation of language models, and to serve as a foundation for future extensions with richer annotations or additional clinical phenomena.

7 Limitations

Despite its contributions, the proposed corpus has limitations that should be considered when interpreting results obtained with it.

First, the dataset size is relatively modest, comprising 969 narratives. This limited number of instances presents a challenge for training traditional supervised machine learning models from scratch, as these models typically require larger corpora to achieve robustness without extensive data augmentation. However, this constraint reflects the fundamental challenge of sample size in rare disease and rare clinical event research (Schaefer et al., 2020). The dataset holds utility for alternative paradigms, such as outlier detection methods and pretrained models.

Second, positive cases account for approximately 5% of the dataset. This proportion reflects real-world prevalence estimates (Cardona et al., 2020) but results in a strongly imbalanced classification problem, which may adversely affect some learning algorithms if not explicitly addressed through appropriate evaluation protocols or training strategies.

Third, although synthetic notes were authored by experienced clinicians, they may not capture the full variability of language observed in real clinical documentation. Synthetic narratives may emphasize prototypical presentations of anaphylaxis and underrepresent atypical, incomplete, or ambiguously documented cases. Similarly, adapted case reports originate from published literature and may differ stylistically from routine electronic health records, which often contain fragmented or telegraphic language.

Finally, the corpus provides only a binary label indicating the presence or absence of anaphylaxis. It does not encode information about severity, trig-

gering agents, temporal progression, or treatment outcomes. Researchers interested in more fine-grained clinical modeling will need to extend the dataset with additional annotations.

8 Ethical Considerations

This section addresses ethical aspects related to the construction, annotation, and release of the corpus. Given the sensitive nature of clinical narratives, particular attention was paid to data privacy, anonymization, and responsible reuse. We also discuss potential biases introduced during data creation and annotation, as well as the intended research-only scope of the dataset.

8.1 Data Privacy and Anonymization

Clinical text frequently contains personally identifiable information, making data sharing ethically and legally challenging. To mitigate these risks, the corpus was constructed using a combination of synthetic data, adapted case reports, and de-identified clinical notes. Synthetic narratives are entirely fictional and do not correspond to real individuals. Adapted case reports were rewritten from published sources and stripped of identifying details. The SemClinBr subset is fully anonymised and complies with the Brazilian General Data Protection Law.

The use of synthetic data aligns with current recommendations advocating privacy-preserving alternatives for clinical NLP research (Mendes et al., 2025). Nevertheless, users of the corpus are encouraged to perform their own due diligence and ensure compliance with local regulations and institutional review requirements when integrating the dataset into downstream applications.

8.2 Annotation Guidelines and Clinical Responsibility

Annotations were performed using well-established diagnostic criteria (Cardona et al., 2020; Dribin et al., 2023), originally designed for clinical practice rather than text annotation. As a result, some degree of interpretative ambiguity is unavoidable, particularly when narratives provide incomplete symptom descriptions. Annotators relied on clinical judgment to resolve such cases, and disagreements were addressed through collective discussion.

The dataset is intended exclusively for research purposes. Models trained on this corpus should not be used for clinical decision-making without rigorous validation, regulatory approval, and integration into appropriate clinical workflows.

8.3 Bias and Representativeness

The corpus may reflect biases related to linguistic style, case selection, and data provenance. Synthetic notes were authored by a limited group of clinicians and may encode their preferred terminology or narrative structure. Adapted case reports originate from a subset of medical journals and may not represent the full diversity of healthcare settings or patient populations. The SemClinBr notes were collected from specific Brazilian institutions and may not generalize to other regions, languages, or health systems.

Researchers should therefore exercise caution when extrapolating findings beyond the scope of the dataset and consider complementing it with additional resources when addressing broader clinical or linguistic questions.

Acknowledgments

This work was supported by the University of São Paulo and the Sírio-Libanês College. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001 and by grant CEPID 2013/07375-0 and grant C4AI 2019/07665-4 (C4AI), São Paulo Research Foundation (FAPESP).

References

Edmon Begoli, Kris Brown, Sudarshan Srinivas, and Suzanne Tamang. 2018. Synthnotes: A generator framework for high-volume, high-fidelity synthetic mental health notes. In *2018 IEEE international conference on big data (big data)*, pages 951–958. IEEE.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2022. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*.

Taxiarchis Botsis and R Ball. 2013. Automating case definitions using literature-based reasoning. *Applied clinical informatics*, 4(04):515–527.

Taxiarchis Botsis, Thomas Buttolph, Michael D Nguyen, Scott Winiecki, Emily Jane Woo, and Robert Ball. 2012. Vaccine adverse event text mining system for extracting features from vaccine safety reports. *Journal of the American Medical Informatics Association*, 19(6):1011–1018.

Victoria Cardona, Ignacio J Ansotegui, Motohiro Ebisawa, Yehia El-Gamal, Montserrat Fernandez Rivas, Stanley Fineman, Mario Geller, Alexei Gonzalez-Estrada, Paul A Greenberger, Mario Sanchez Borges, and 1 others. 2020. World allergy organization anaphylaxis guidance 2020. *World allergy organization journal*, 13(10):100472.

David S Carrell, Susan Gruber, James S Floyd, Maralyssa A Bann, Kara L Cushing-Haugen, Ron L Johnson, Vina Graham, David J Cronkite, Brian L Hazlehurst, Andrew H Felcher, and 1 others. 2023. Improving methods of identifying anaphylaxis for medical product safety surveillance using natural language processing and machine learning. *American Journal of Epidemiology*, 192(2):283–295.

Naila Camila da Rocha, Abner Macola Pacheco Barbosa, Yaron Oliveira Schnr, Juliana Machado-Rugolo, Luis Gustavo Modelli de Andrade, José Eduardo Corrente, and Liciana Vaz de Arruda Silveira. 2023. Natural language processing to extract information from portuguese-language medical records. *Data*, 8(1).

Rildo Pinto da Silva and Antonio Pazin-Filho. 2025. Dataset of anonymized discharge summaries of sepsis patients from a brazilian tertiary hospital for nlp applications. *Data in Brief*, page 111804.

Henrique Dias and Ana Helena Dias Pereira dos Ulbrich. 2022. BRATECA (Brazilian Tertiary Care Dataset): a Clinical Information Dataset for the Portuguese Language. PhysioNet. RRID:SCR_007345.

Timothy E Dribin, Megan S Motosue, and Ronna L Campbell. 2023. Overview of allergy and anaphylaxis. *Immunology and Allergy Clinics*, 43(3):435–451.

Luis Felipe Ensina, Matheus Matos Machado, Joice B. Machado Marques, Monica Pugliese H. dos Santos, Fábio Cerqueira Lario, Chayanne Andrade Araújo, Fabiana Andrade Nunes Oliveira, and Dilvan de Abreu Moreira. 2025. Artificial intelligence for detecting anaphylaxis in electronic medical records. *Asia Pacific Allergy*, 15(3):153–158.

Fernanda Bufon Färber, Iago Alves Brito, Julia Soares Dollis, Pedro Schindler Freire Brasil Ribeiro, Rafael Teixeira Sousa, and 1 others. 2025. Medpt: A massive medical question answering dataset for brazilian-portuguese speakers. *arXiv preprint arXiv:2511.11878*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Georgi Grazhdanski, Vasil Vasilev, Sylvia Vassileva, Dimitar Taskov, Izabel Antova, Ivan Koychev, and Svetla Boytcheva. 2025. Synthmedic: Utilizing large language models for synthetic discharge summary generation, correction and validation. *Journal of Biomedical Informatics*, 170:104906.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*. Version 2.2.

Jianfu Li, Yujia Zhou, Xiaoqian Jiang, Karthik Natarajan, Serguei Vs Pakhomov, Hongfang Liu, and Hua Xu. 2021. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *Journal of the American Medical Informatics Association*, 28(10):2193–2201.

Ying-Chih Lo, Sheril Varghese, Suzanne Blackley, Diane L Seger, Kimberly G Blumenthal, Foster R Goss, and Li Zhou. 2022. Reconciling allergy information in the electronic health record after a drug challenge using natural language processing. *Frontiers in allergy*, 3:904923.

Fábio Lopes, César Teixeira, and Hugo Gonçalo Oliveira. 2019. Contributions to clinical named entity recognition in portuguese. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 223–233.

Matheus Matos Machado, Joice Basílio Machado Marques, Fabrício A. Gualdani, Monica Pugliese Heleodoro dos Santos, Fabio Cerqueira Lario, Chayanne Andrade de Araujo, Fabiana Andrade Nunes Oliveira, Luis Felipe Chiaverini Ensina, Ricardo Marcondes Marcacini, and Dilvan Moreira. 2024. Evaluating large language models for anaphylaxis detection in clinical notes. *Journal of Health Informatics*, 16(Especial).

Jorge M Mendes, Aziz Barbar, and Marwa Refaie. 2025. Synthetic data generation: a privacy-preserving approach to accelerate rare disease research. *Frontiers in Digital Health*, 7:1563991.

Simon Meoni, Éric De La Clergerie, and Théo Ryffel. 2025. Synthetic documents for medical tasks: Bridging privacy with knowledge injection and reward mechanism. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 12–25, Albuquerque, New Mexico. Association for Computational Linguistics.

Denis Moser, Matthias Bender, and Murat Sariyar. 2024. Generating synthetic healthcare dialogues in emergency medicine using large language models. In *Collaboration across Disciplines for the Health of People, Animals and Ecosystems*, pages 235–239. IOS Press.

Lucas Emanuel Silva Oliveira, Ana Carolina Peters, Adalniza Moura Pucca da Silva, Caroline Pilatti Gebeluca, Yohan Bonescki Gumiel, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid Al Hasan, and Claudia Maria Cabral Moro. 2022.

SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, 13(1).

Krithika Ramesh, Daniel Smolyak, Zihao Zhao, Nupoor Gandhi, Ritu Agarwal, Margrét V Bjarnadóttir, and Anjalie Field. 2025. Synthtexteval: Synthetic text data generation and evaluation for high-stakes domains. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 487–499.

Libo Ren, Samuel Belkadi, Lifeng Han, Warren Del-Pinto, and Goran Nenadic. 2025. Synthetic4health: generating annotated synthetic clinical letters. *Frontiers in Digital Health*, 7:1497130.

Atiquer Rahman Sarkar, Yao-Shun Chuang, Xiaoqian Jiang, and Noman Mohammed. 2025. Not fully synthetic: Llm-based hybrid approaches towards privacy-preserving clinical note sharing. *AMIA Summits on Translational Science Proceedings*, 2025:441.

Julia Schaefer, Moritz Lehne, Josef Schepers, Fabian Prasser, and Sylvia Thun. 2020. The use of machine learning in rare diseases: a scoping review. *Orphanet journal of rare diseases*, 15(1):145.

Mauricio Schiezaro, Guilherme Rosa, Bruno Augusto Goulart Campos, and Helio Pedrini. 2025. Guardians of the data: Ner and llms for effective medical record anonymization in brazilian portuguese. *Frontiers in Public Health*, 13:1717303.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 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, Online. Association for Computational Linguistics.

Isabel Segura-Bedmar, Cristobal Colón-Ruíz, Miguél Ángel Tejedor-Alonso, and Mar Moro-Moro. 2018. Predicting of anaphylaxis in big data emr by exploring machine learning approaches. *Journal of biomedical informatics*, 87:50–59.

F Estelle R Simons, Ledit RF Ardusso, M Beatrice Bilò, Yehia M El-Gamal, Dennis K Ledford, Johannes Ring, Mario Sanchez-Borges, Gian Enrico Senna, Aziz Sheikh, Bernard Y Thong, and 1 others. 2011. World allergy organization guidelines for the assessment and management of anaphylaxis. *World Allergy Organization Journal*, 4(2):13–37.

Kathleen E Walsh, Sarah L Cutrona, Sarah Foy, Meghan A Baker, Susan Forrow, Azadeh Shoaibi, Pamala A Pawloski, Michelle Conroy, Andrew M Fine, Lise E Nigrovic, and 1 others. 2013. Validation of anaphylaxis in the food and drug administration's mini-sentinel. *Pharmacoepidemiology and drug safety*, 22(11):1205–1213.

Rongsheng Wang, Junying Chen, Ke Ji, Zhenyang Cai, Shunian Chen, Yunjin Yang, and Benyou Wang. 2025. Medgen: Unlocking medical video generation by scaling granularly-annotated medical videos. *arXiv preprint arXiv:2507.05675*.

Yanshan Wang, Sunyang Fu, Feichen Shen, Sam Henry, Ozlem Uzuner, and Hongfang Liu. 2020. The 2019 n2c2/ohnlp track on clinical semantic textual similarity: Overview. *JMIR Med Inform*, 8(11):e23375.

Wei Yu, Chengyi Zheng, Fagen Xie, Wansu Chen, Cheryl Mercado, Lina S Sy, Lei Qian, Sungching Glenn, Hung F Tseng, Gina Lee, and 1 others. 2020. The use of natural language processing to identify vaccine-related anaphylaxis at five health care systems in the vaccine safety datalink. *Pharmacoepidemiology and drug safety*, 29(2):182–188.