Leakage-Aware Evaluation of Brazilian Clinical Notes for ICU Mortality Prediction: A Study on the BRATECA Dataset

Felipe André Bach Alves[1], **Heloísa Oss Boll**[2], **Mariana Recamonde-Mendoza**[1,3]

[1]Institute of Informatics, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

[2]Dept. of Medical Informatics, Amsterdam UMC/University of Amsterdam, The Netherlands

[3]Bioinformatics Core, Hospital de Clínicas de Porto Alegre (HCPA), Brazil

{mrmendoza, felipe.alves}@inf.ufrgs.br, h.ossboll@amsterdamumc.nl

Abstract. *Early prediction of ICU mortality can support clinical decisions, but retrospective notes can contain future-event cues that cause leakage in clinical NLP. We evaluate leakage-aware modeling of Brazilian Portuguese ICU notes for mortality prediction using BRATECA. We compare 24-hour restriction, regex and LLM leakage auditing, neural and TF-IDF representations, and unrestricted-note baselines. In the 24-hour setting, models achieved AUROC 0.78-0.83; unrestricted notes reached around 0.95, indicating outcome-related information in late documentation. Regex found leakage signals in 21.13% of the notes, and the LLM audit identified additional semantic cases. Results highlight the need for temporal control and leakage auditing in clinical NLP.*

1. Introduction

Predicting clinical outcomes from *electronic health records* (EHRs) is a central task in medical *machine learning* (ML), particularly in high-risk settings such as *intensive care units* (ICUs). Early prediction of in-hospital mortality may support clinical decision-making, resource allocation, and risk stratification [Choi et al. 2022]. EHR datasets combine structured information, such as demographics, examinations, and prescriptions, with unstructured clinical narratives that capture clinical reasoning and patient trajectories. Recent studies have emphasized the importance of leveraging multimodal EHR data to improve predictive modeling and representation learning in healthcare [Boll et al. 2024].

Clinical notes are especially valuable because they include contextual information often absent from structured fields [Seinen et al. 2025]. However, they also introduce methodological risks: notes may be written retrospectively, copied forward, amended after an event, or produced near discharge or death, so they may contain information unavailable at the intended prediction time. This is a form of data leakage, that is, the model receiving information unavailable at the intended prediction time. For instance, when predicting mortality from documentation available within the first 24 hours of admission, a note written days later stating that the patient died or was discharged reveals the target and must be excluded as input. Used without temporal control, such notes let models exploit outcome-related cues instead of early clinical signals, yielding overly optimistic estimates and poor prospective validity [Chiavegatto Filho et al. 2021].

This risk is further intensified in ICU environments, where patients exhibit rapidly evolving clinical trajectories, dense monitoring, frequent interventions, and extensive documentation histories. Despite growing interest in clinical text modeling, leakage-aware

analyses of ICU narratives remain limited, particularly beyond English-language benchmark datasets such as MIMIC [van Aken et al. 2021]. In this context, clinical narratives in Brazilian Portuguese remain underexplored, and their documentation practices may differ substantially from those represented in widely used international benchmarks.

This paper investigates data leakage in ICU clinical notes for in-hospital mortality prediction using the Brazilian clinical data collection BRATECA [Dias and Ulbrich 2022, Consoli et al. 2022]. Because the dataset does not currently include metadata about the note types, robust leakage auditing is essential before such notes can be reliably incorporated into prediction pipelines. Therefore, our central *natural language processing* (NLP) questions are: 1) To what extent does the predictive performance of models using these clinical notes for ICU mortality prediction indicate real, early clinical signal rather than retrospective outcome leakage? and 2) How can temporal restriction and semantic auditing be combined to help quantify this gap in these clinical narratives?

Rather than proposing a new predictive architecture, this study makes a methodological and empirical contribution to clinical NLP. Specifically, we: (i) quantify how temporal availability of a Brazilian Portuguese clinical note dataset changes estimated mortality-prediction performance; (ii) compare lexical, dense, Portuguese-language, and translated-English representations under controlled prediction windows; and (iii) combine rule-based and LLM auditing to characterize explicit and semantic outcome cues. Our results show that conclusions about predictive quality are highly sensitive to how temporal cuts and eligible documentation are defined, offering concrete guidance for avoiding leakage-driven overestimation on BRATECA or other clinical notes datasets.

2. Related Work

Predictive modeling using ML and EHR data has been extensively studied for ICU risk assessment, particularly for mortality prediction. Most existing models rely primarily on structured variables such as vital signs, laboratory results, prescriptions, and severity scores derived from datasets such as MIMIC [Johnson et al. 2016], frequently achieving performance comparable to or exceeding traditional clinical scoring systems such as APACHE, SAPS, and SOFA [Knaus et al. 1985, Le Gall et al. 1993, Vincent et al. 1996, Olang et al. 2025]. Recent work has also shown the value of combining structured variables with clinical narratives, using multimodal EHR representations for risk prediction [Boll et al. 2024, Garriga et al. 2023].

Clinical NLP has enabled systematic use of free-text documentation through lexical and representation-learning methods. Lexical baselines such as *term frequency–inverse document frequency* (TF-IDF) remain important in clinical NLP because they provide strong, interpretable comparisons and can expose reliance on explicit outcome words [Tavabi et al. 2024]. More recent studies rely on contextual representations from transformer-based models such as BioBERT and ClinicalBERT [Alsentzer et al. 2019, Lee et al. 2020]. These approaches can capture semantic patterns in clinical notes and have been applied to mortality prediction and related ICU tasks [Vagliano et al. 2023]. More recently, general and domain-adapted decoder models, including ChatGPT, Portuguese *large language models* (LLMs) such as Sabiá, and medical LLMs such as Meditron, have expanded the range of NLP methods available for clinical and biomedical tasks [OpenAI 2023, Nogueira et al. 2023, Chen et al. 2023].

However, data leakage remains a persistent challenge in health ML. Prior work has discussed leakage as a source of biased performance estimates and poor reproducibility [Chiavegatto Filho et al. 2021]. In clinical narratives, leakage may arise when documentation includes information generated after the intended prediction time. Although methodological frameworks for the development of leakage-aware models have been proposed [Davis et al. 2023], the specific problem of temporal leakage in non-English ICU notes remains not sufficiently characterized. Here, we address this gap by exploring multiple strategies for leakage-aware modeling of Brazilian ICU narratives, including an LLM component for semantic leakage auditing in Portuguese.

3. Methodology

3.1. Dataset, ICU Cohort, and Prediction Task

We used BRATECA, a Brazilian retrospective EHR dataset containing relational tables with admissions, clinical notes, exams, prescriptions, and prescription items [Dias and Ulbrich 2022, Goldberger et al. 2000]. Table 1 summarizes the source tables used in this study. The target prediction task was in-hospital mortality. Figure 1 summarizes the cohort construction, clinical-note preprocessing, and three experimental branches used in this study.

The primary source of unstructured text was *B1_ClinicalNote*. This table contains heterogeneous Brazilian Portuguese free-text documentation produced during hospitalization, including notes consistent with admission documentation, clinical progress/evolution, nursing and medical updates, procedure-related records, discharge-oriented documentation, and administrative or care-transition records. These narratives vary in length, style, abbreviation density, and temporal proximity to clinical events, which makes them useful for clinical NLP while also increasing the risk of temporal and outcome-related leakage.

Table 1. BRATECA source tables used in this study.

Table	Cols	Rows	Description
B1_Admission	11	72,085	Encounter-level admission data, including patient and hospital identifiers and discharge information.
B1_ClinicalNote	6	2,855,818	Free-text clinical notes recorded during hospitalization.
B1_Exam	7	2,372,088	Exams per admission and their results.
B1_Prescription	30	518,420	Prescription headers, timestamps, and ICU-related flags.
B1_PrescriptionItem	30	5,737,707	Prescription items linked to prescription headers.

The ICU cohort was identified using the ICU flag in the prescription table: an admission was considered an ICU encounter when it appeared in *B1_Prescription* with at least one record where *IC* was true. This definition yielded 4,779 ICU admissions. Outcome labels were extracted from discharge disposition values and binarized as death versus discharge alive. Dispositions outside the modeling scope, such as transfers and non-standard discharges, were excluded. The modeled death-versus-discharge cohort contained 4,723 ICU admissions, including 1,098 deaths and 3,625 discharges alive.

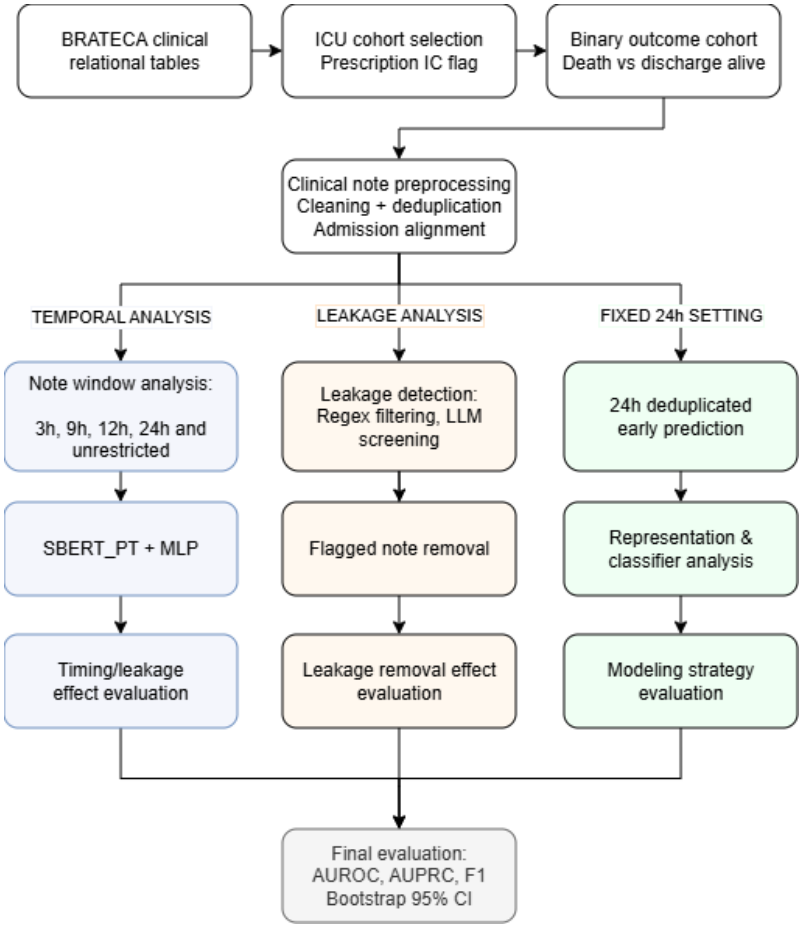

Figure 1. Overview of cohort construction and the temporal, leakage-audit, and fixed 24-hour evaluation branches.

In the modeled death-versus-discharge ICU cohort, the recorded time from admission to discharge was 18.13 $\pm$ 17.50 days (median 12.72; *interquartile range* (IQR) 6.98-23.12). Among the modeled admissions with clinical notes, the last available clinical note occurred 19.06 $\pm$ 17.99 days after admission (median 13.74; IQR 7.47-24.45), showing that unrestricted note sets frequently include documentation from late hospitalization stages. Because the main text-only experiments required at least one eligible clinical note within the 24-hour window after deduplication and embedding alignment, the aligned 24-hour matrix contained 4,668 admissions and 1,078 deaths.

Table 2 summarizes the main raw, derived, and generated fields used in the experiments. Identifiers and outcome labels were used for linkage, splitting, cohort definition, or evaluation only; they were not used as text-model predictors and were not submitted to the LLM audit.

3.2. Leakage Mitigation

We evaluated a leakage-aware preprocessing strategy with four components. First, temporal windowing restricted notes to clinically plausible prediction horizons, especially the first 24 hours after hospital admission among ICU-flagged admissions. Second, rule-based regular expressions flagged notes containing explicit or proxy outcome-related expressions, including terms related to death, discharge, transfers, failed resuscitation, ter-

Table 2. Raw, derived, and generated fields used in the experimental pipeline.

Field or group	Type	Role in the study
Admission_ID, *Patient_ID*, *Hospital_ID*	Raw identifiers	Used for relational linkage, admission-level aggregation, and split construction. They were not used as predictive features and were not submitted to the LLM.
IC	Raw structured flag	Used only to define ICU admissions from the prescription table.
Admission_Date, *Note_Date*, hours since admission	Raw/derived time fields	Used to compute temporal eligibility windows such as 3h, 9h, 12h, and 24h. These time fields were not included as text-model predictors.
Note_Text_Clean	Derived text	Cleaned Brazilian Portuguese clinical note text used to build embeddings, TF-IDF features, deduplication keys, and LLM audit inputs.
target_obito	Derived outcome	Binary in-hospital mortality label derived from discharge disposition values. It was used for training/evaluation but was not submitted to the LLM audit.
Regex leakage flag	Generated audit field	Rule-based indicator of explicit or proxy outcome-related expressions, used for leakage prevalence estimation and filtering sensitivity analyses.
Embedding vectors	Generated features	Dense note-level representations generated by pretrained encoders and aggregated to the admission level, mainly through mean pooling after deduplication.
LLM audit fields	Generated audit fields	Structured fields such as leakage presence, leakage type, evidence, explanation, and confidence. They were used only for exploratory leakage auditing and sensitivity analysis, not as predictors in the mortality model.

minal documentation, and late addenda. Third, the last note of each ICU admission was excluded in selected configurations because final notes often summarize the trajectory or outcome. Fourth, duplicated clinical notes within the same admission were removed using cleaned note text, reducing the influence of repeated or copied documentation.

The experiments used the preprocessed `Note_Text_Clean` field. The upstream cleaning procedure removed markup artifacts and textual noise and applied character normalization. The same cleaned field was used for deduplication, representation learning, TF-IDF vectorization, and LLM audit inputs. Table 3 illustrates the main lexical categories targeted by the regex audit. The examples are synthetic Brazilian Portuguese clinical-text cues derived from the rule categories, not excerpts from BRATECA.

Table 3. Synthetic examples of lexical leakage cues targeted by the regex audit.

Category	Synthetic lexical cue	Potential leakage mechanism
Explicit death	"evoluiu a obito"; "obito constatado"	Directly reveals the mortality outcome used as the prediction label.
Discharge	"alta hospitalar"; "recebeu alta"	Reveals survival/discharge information that should not be available at early prediction time.
Transfer or post-discharge trajectory	"transferido para enfermaria"; "home care"	Indicates later care trajectory or disposition after ICU documentation.
Failed resuscitation or vital-status proxy	"PCR sem retorno"; "sem sinais vitais"	Provides a strong proxy for death even when the outcome label is not explicitly named.
Terminal context	"cuidados paliativos"; "conforto exclusivo"	May reveal end-of-life trajectory or late-stage clinical decisions.
Administrative closure	"caso encerrado"; "desfecho"	Suggests documentation created after the hospitalization outcome was known.

Regex screening was designed as a baseline audit of lexical leakage indicators, targeting explicit outcome-related expressions and predefined patterns. To complement this approach, we also performed an exploratory LLM-based semantic audit aimed at identifying potentially outcome-revealing content that may not be captured by surface lexical cues alone. This evaluation uses the Azure AI Foundry OpenAI-compatible Responses API with the model `gpt-5.4-mini`, version 2026-03-17, with temperature = 0.1 and structured JSON Schema outputs. This deployment was selected for scalable, monitored batch processing with structured JSON outputs and latency/cost compatible

with full-corpus screening. Only cleaned clinical-note text was submitted, and under the enterprise/API configuration used in this study, prompts and outputs are not reused to train foundation models.

3.3. Embeddings and Aggregation

We evaluated multiple pretrained embedding configurations across language and domain settings. The multilingual sentence-level baseline used `sentence-transformers/paraphrase-multilingual-mpnet-base-v2` (SBERT), which produces 768-dimensional note representations. Additional configurations used BERTimbau for Brazilian Portuguese, Portuguese and English BioBERT variants, and ClinicalBERT for English clinical text.

For English-language configurations, Brazilian Portuguese notes were translated using `facebook/nllb-200-distilled-600M` [Costa-jussà et al. 2022]. Translation was performed in batches and persisted as Parquet shards, which were subsequently concatenated while preserving record order and alignment with the original note identifiers. This alignment allowed the same temporal filters, labels, and admission-level aggregation procedure to be applied to the original and translated notes. After within-admission deduplication and temporal filtering, note-level embeddings were aggregated into a single admission-level representation using mean pooling.

3.4. Experimental Protocol

We evaluated in-hospital mortality prediction at the admission level. The main tables use a stratified 70/20/10 admission-level train/validation/test split with a fixed random seed. The main neural classifier was a multilayer perceptron (MLP) trained on admission-level embeddings. The MLP consisted of two hidden layers with 256 and 64 units, ReLU activations, dropout of 0.30 and 0.20, and a single output logit. Models were trained with AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 128, and early stopping based on validation AUROC. Class imbalance was handled using cost-sensitive learning.

For comparison, we evaluated logistic regression, linear *support vector machine* (SVM), *Gaussian naive Bayes* (NB), and nearest centroid using the same SBERT_PT 24-hour admission-level matrix as the MLP. TF-IDF word 1-2-gram and character 3-5-gram baselines used the same eligible admissions and split, but their own sparse representations, with class-balanced logistic regression. Thresholds were selected using the validation set. We report AUROC, AUPRC, F1, precision, recall, and balanced accuracy; 95% confidence intervals were estimated from 2000 bootstrap resamples of test admissions with replacement. The five-seed LLM sensitivity analysis is reported separately.

4. Results and Discussion

4.1. Prevalence of Regex-Detected Leakage Signals

Regex-based leakage screening identified 261,395 notes with explicit or proxy outcome-related expressions among 1,236,991 evaluated notes, corresponding to 21.13% of the note corpus (Table 4). This confirms that leakage indicators are frequent in clinical narratives and motivates temporal and linguistic auditing.

Table 4. Prevalence of leakage signals identified by regex-based auditing.

Leakage detected	Notes	Share
No	975,596	78.87%
Yes	261,395	21.13%
Total	1,236,991	100.00%

4.2. Comparison of Embedding Representations for Mortality Prediction

Table 5 shows the 24-hour embedding comparison for the in-hospital mortality prediction task using an MLP classifier. Performance differs across embedding models, indicating that the choice of clinical note encoder impacts mortality prediction. SBERT-based representations obtained the highest observed AUROC values, but SBERT_EN and SBERT_PT were nearly tied and their bootstrap intervals overlap. Therefore, the small AUROC difference should not be interpreted as robust superiority. SBERT_EN achieved higher observed AUPRC in this configuration, whereas SBERT_PT remains methodologically attractive because it operates directly on the original Brazilian Portuguese notes without machine translation, avoiding translation costs and possible semantic shifts.

Table 5. Performance of different embedding methods for ICU mortality prediction under the 24-hour setting using an MLP classifier.

Model	AUROC (95% CI)	AUPRC (95% CI)	F1
SBERT_EN	0.8033 [0.760, 0.848]	0.5927 [0.497, 0.690]	0.5205
SBERT_PT	0.8025 [0.755, 0.849]	0.5676 [0.475, 0.664]	0.5190
BERTimbau_PT	0.7810 [0.733, 0.834]	0.4978 [0.402, 0.598]	0.5432
BioBERT_PT	0.7653 [0.711, 0.817]	0.4487 [0.363, 0.547]	0.5179
BioBERT_EN	0.7506 [0.707, 0.802]	0.4316 [0.350, 0.527]	0.4910
ClinicalBERT_EN	0.7482 [0.702, 0.803]	0.4719 [0.381, 0.569]	0.4873

4.3. Performance of Lexical TF-IDF Baselines

Because lexical cues are central to leakage analysis in clinical notes, we evaluated TF-IDF logistic-regression (LR) baselines on the same 24-hour admission-level split used for the embedding experiments. Unlike previous neural representations, which encode each note as a dense semantic vector before admission-level pooling, TF-IDF represents admissions through sparse word- and character-level n-gram patterns. Table 6 shows that both word- and character-level TF-IDF baselines were competitive with, and in this holdout exceeded, the neural embedding probes reported in Table 5. This result strengthens the NLP interpretation of the task: explicit lexical patterns in Brazilian Portuguese clinical notes retain considerable predictive information even under early temporal restriction.

Table 6. Performance of lexical TF-IDF baselines under the aligned 24-hour setup.

Model	AUROC (95% CI)	AUPRC (95% CI)	F1
TF-IDF word 1-2g + LR	0.8293 [0.788, 0.873]	0.6049 [0.514, 0.702]	0.5674
TF-IDF char 3-5g + LR	0.8310 [0.790, 0.873]	0.6221 [0.533, 0.715]	0.5783

4.4. Analysis of Temporal-Cutoff Sensitivity

Figure 2 reports the sensitivity of text-only SBERT_PT performance to note availability over time. Performance varied non-monotonically across the early prediction windows, with overlapping confidence intervals and differences in the amount of eligible documentation. In contrast, the unrestricted setting produced a substantially larger increase in both AUROC and AUPRC. This does not imply that early notes lack useful signal; rather, it shows that unrestricted retrospective text can inflate performance estimates when temporal availability is not controlled.

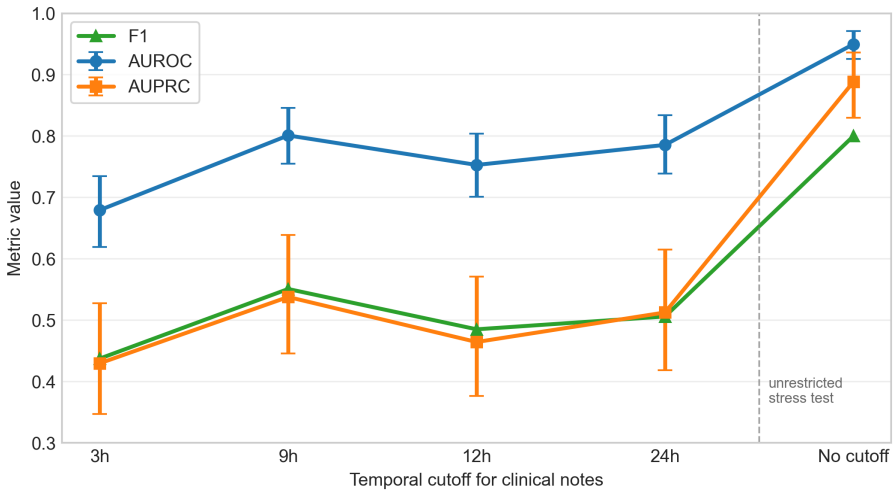

Figure 2. Text-only temporal cutoff sensitivity for SBERT_PT using mean-pooled, deduplicated note embeddings, with 95% CI (AUROC and AUPRC).

The "No cutoff" setting in Figure 2 and the "Unrestricted notes - No regex cleaning" setting in Table 8 are both unrestricted-text upper bounds, but they are not identical. Figure 2 uses the temporal-sensitivity feature lineage derived from the modeled ICU cohort of 4,723 admissions and 1,098 deaths, where as the regex-ablation pipeline reconstructs the unrestricted corpus from the regex lineage with 4,722 eligible admissions and 1,098 deaths. Therefore, the small numerical difference between AUROC 0.9496 and 0.9458 reflects slightly different eligibility/source lineage and split composition; however, the intended interpretation is the same: unrestricted retrospective notes yield substantially inflated performance.

4.5. Comparison of Traditional Classifiers Under the 24-Hour Setting

To compare classifier families under a fixed representation, we used a separately rebuilt SBERT PT 24-hour feature matrix aligned across all classifiers. To address whether neural models outperform simpler classifiers under the same representation, we compared the MLP with traditional baselines using the same SBERT_PT 24-hour mean-pooled, deduplicated admission-level matrix in the in-hospital mortality prediction task. Table 7 shows that the MLP achieved the highest observed AUROC and AUPRC, indicating better ranking performance over the shared embedding representation. Logistic regression was the closest traditional baseline, with lower AUROC/AUPRC but a similar F1, driven by higher recall and lower precision. The remaining traditional methods performed substantially worse, especially in AUPRC, suggesting a weaker ability to rank positive mortality cases

under class imbalance. Under the shared SBERT_PT representation, the MLP had the highest observed AUROC and AUPRC, while logistic regression achieved a similar F1 with higher recall and lower precision. Because the closest bootstrap intervals overlap, these results do not establish statistically robust superiority of the MLP.

Table 7. Aligned comparison between MLP and traditional classifiers using the same rebuilt SBERT_PT 24-hour text-only feature matrix.

Model	AUROC (95% CI)	AUPRC (95% CI)	F1
MLP probe	0.7855 [0.739, 0.834]	0.5125 [0.418, 0.615]	0.5059
Logistic regression	0.7527 [0.699, 0.808]	0.4806 [0.393, 0.581]	0.5112
Linear SVM	0.7113 [0.656, 0.769]	0.4128 [0.335, 0.510]	0.4944
Gaussian NB	0.6939 [0.639, 0.749]	0.3459 [0.279, 0.419]	0.4215
Nearest centroid	0.6827 [0.628, 0.738]	0.3407 [0.277, 0.414]	0.4000

4.6. Impact of Regex-Based Leakage Filtering

Table 8 compares prediction performance with and without regex-based filtering. In the 24-hour setting, filtering produced a small observed improvement, but the bootstrap intervals overlap and the result should not be interpreted as a robust gain. A plausible explanation is removal of noisy, copied, or administratively late documentation, but this remains a sensitivity finding. In the unrestricted setting, filtering reduced performance, consistent with the removal of outcome-related content from late-stage notes. These findings show that leakage is not simply the presence of outcome-correlated terms, but the use of information unavailable at prediction time. Regex filtering can mitigate explicit lexical cues, but temporal restriction addresses the root cause more directly.

Table 8. Impact of regex-based leakage filtering on SBERT_PT performance. Unrestricted notes include all eligible documentation recorded from hospital admission through discharge or death, including notes written after the 24-hour prediction cutoff. These configurations represent retrospective upper bounds rather than deployable early-prediction estimates.

Configuration	AUROC (95% CI)	AUPRC (95% CI)	F1
24h window - No regex filtering	0.7518 [0.702, 0.800]	0.4371 [0.354, 0.534]	0.5139
24h window - With regex filtering	0.7603 [0.706, 0.811]	0.4732 [0.385, 0.581]	0.5177
Unrestricted notes - No regex filtering	0.9458 [0.918, 0.966]	0.8844 [0.826, 0.929]	0.8019
Unrestricted notes - With regex filtering	0.9332 [0.905, 0.956]	0.8314 [0.762, 0.887]	0.7449

4.7. Exploratory Semantic Leakage Auditing with LLMs

In the preliminary audit, 61,852 ICU note records were evaluated and 2,430 were marked as containing potential leakage. In five-seed sensitivity analysis, LLM-based exclusion did not robustly change aggregate performance; however, among admissions affected by flagged notes, MLP AUPRC decreased from 0.8735 to 0.8321 after exclusion.

Among the flagged notes, 197 did not match any regex rule and most often contained discharge, indirect-outcome, or terminal-context cues. These cases illustrate how semantic auditing may complement predefined lexical patterns, although the annotations were not validated against a human gold standard.

5. Limitations

This study has several limitations. First, experiments used a single Brazilian ICU dataset (BRATECA), which may limit generalizability to other healthcare systems, languages, and documentation practices. Second, regex-based leakage detection captures explicit and proxy lexical cues but cannot exhaustively identify semantic or context-dependent outcome information; our preliminary LLM audit surfaced additional cases but was not validated against a human-annotated gold standard. Third, differences between Brazilian Portuguese and English representations may reflect both encoder behavior and translation artifacts from the NLLB-200 preprocessing step. Fourth, central comparisons rely on a single fixed admission-level train/validation/test split with bootstrap uncertainty rather than repeated cross-validation. Because splitting occurred at the admission level, admissions from the same patient may fall in different partitions; quantifying this overlap and testing patient-grouped splits is important future work. Sample sizes also vary slightly across experiments, since temporal cutoffs, leakage filtering, and deduplication remove different subsets of notes. Broader cross-dataset validation is needed to determine whether the leakage patterns observed here generalize beyond BRATECA.

6. Conclusion

This study evaluated distinct strategies for leakage-aware modeling of Brazilian ICU clinical notes for in-hospital mortality prediction using BRATECA. The results show that early clinical narratives contain meaningful predictive signal, with 24-hour text-based models achieving AUROC around 0.78-0.83 depending on whether neural embeddings or lexical TF-IDF representations are used. However, including unrestricted retrospective notes substantially increases text-only performance to AUROC around 0.95, which indicates the presence of outcome-related information in the documentation.

The findings support three methodological conclusions. First, strict temporal control is essential when using clinical notes for prediction. Second, regex-based filtering can identify frequent leakage indicators but should be treated as a partial mitigation strategy. Third, LLM-based semantic auditing is promising for identifying subtler forms of leakage. A qualitative inspection of LLM-only cases suggests a path for extending this research toward LLM-based mortality prediction. For example, some notes flagged only by the LLM described positive clinical and apnea tests compatible with a brain-death protocol. Although such notes may not contain the exact regex triggers for death or discharge, they semantically reveal a terminal trajectory and information unavailable at early prediction time. This qualitative inspection also motivates future work on temporally constrained LLM-based representations for mortality prediction. Overall, this work emphasizes that reliable clinical-text modeling requires not only predictive performance reporting, but also explicit auditing of what information was available at the intended prediction time.

Ethics Statement

The BRATECA dataset was accessed through PhysioNet under credentialed access and is distributed as deidentified research data. This study did not attempt to reidentify patients, clinicians, hospitals, or institutions. For the exploratory LLM audit, only cleaned clinical-note text was submitted to Azure AI Foundry; under the enterprise/API configuration used, submitted prompts and outputs are not reused for foundation-model training.

Disclosure of Generative AI Usage

Generative artificial intelligence tools were used to support language revision, manuscript organization, and the summarization of experimental findings. All scientific decisions, analyses, interpretations, and conclusions were critically reviewed by the authors and remain their sole responsibility.

References

Alsentzer, E., Murphy, J., Boag, W., et al. (2019). Publicly available clinical bert embeddings. In *Proceedings of the Clinical NLP Workshop*.

Boll, H. O., Amirahmadi, A., Ghazani, M. M., de Morais, W. O., de Freitas, E. P., Soliman, A., Etminani, F., Byttner, S., and Recamonde-Mendoza, M. (2024). Graph neural networks for clinical risk prediction based on electronic health records: A survey. *Journal of Biomedical Informatics*.

Chen, Z., Hernández Cano, A., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., et al. (2023). Meditron-70b: Scaling medical pretraining for large language models.

Chiavegatto Filho, A. D. P., Batista, A. F., and Santos, H. G. (2021). Data leakage in health outcomes prediction with machine learning. comment on "prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning". *Journal of Medical Internet Research*, 23(2).

Choi, M. H., Kim, D., Choi, E. J., Jung, Y. J., Choi, Y. J., Cho, J. H., and Jeong, S. H. (2022). Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records. *Scientific Reports*, 12(1):7180.

Consoli, B. S., dos Santos, H. D. P., Ulbrich, A. H. D. P. S., Vieira, R., and Bordini, R. H. (2022). Brateca (brazilian tertiary care dataset): a clinical information dataset for the portuguese language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, Marseille, France. European Language Resources Association (ELRA).

Costa-jussà, M. R., Cross, J., Červeňansky, F., et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Davis, S. E., Matheny, M. E., Balu, S., and Sendak, M. P. (2023). A framework for understanding label leakage in machine learning for health care. *Journal of the American Medical Informatics Association: JAMIA*, 31(1):274.

Dias, H. and Ulbrich, A. H. D. P. d. (2022). BRATECA (Brazilian Tertiary Care Dataset): a Clinical Information Dataset for the Portuguese Language. *PhysioNet*. Version 1.1.

Garriga, R., Buda, T. S., Guerreiro, J., Omaña Iglesias, J., Estella Aguerri, I., and Matić, A. (2023). Combining clinical notes with structured electronic health records enhances the prediction of mental health crises. *Cell Reports Medicine*, 4(11):101260.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Johnson, A. E., Pollard, T. J., Shen, L., et al. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Knaus, W. A., Draper, E. A., Wagner, D. P., and Zimmerman, J. E. (1985). Apache ii: a severity of disease classification system. *Critical Care Medicine*, 13(10):818–829.

Le Gall, J.-R., Lemeshow, S., and Saulnier, F. (1993). A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *JAMA*, 270(24):2957–2963.

Lee, J., Yoon, W., Kim, S., et al. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Nogueira, R., Pires, R., Abonizio, H., et al. (2023). Sabiá: Portuguese large language models.

Olang, O., Mohseni, S., Shahabinezhad, A., Hamidianshirazi, Y., Goli, A., Abolghasemian, M., Shafiee, M. A., Aarabi, M., Alavinia, M., and Shaker, P. (2025). Artificial intelligence-based models for prediction of mortality in ICU patients: a scoping review. *Journal of Intensive Care Medicine*, 40(12):1240–1246.

OpenAI (2023). Gpt-4 technical report.

Seinen, T. M., Kors, J. A., van Mulligen, E. M., and Rijnbeek, P. R. (2025). Using structured codes and free-text notes to measure information complementarity in electronic health records: Feasibility and validation study. *Journal of Medical Internet Research*, 27:e66910.

Tavabi, N., Singh, M., Pruneski, J., and Kiapour, A. M. (2024). Systematic evaluation of common natural language processing techniques to codify clinical notes. *Plos one*, 19(3):e0298892.

Vagliano, I., Dormosh, N., Rios, M., Luik, T. T., Buonocore, T., Elbers, P. W., Dongelmans, D. A., Schut, M. C., and Abu-Hanna, A. (2023). Prognostic models of in-hospital mortality of intensive care patients using neural representation of unstructured text: A systematic review and critical appraisal. *Journal of Biomedical Informatics*, 146:104504.

van Aken, B., Papaioannou, J.-M., Mayrdorfer, M., Budde, K., Gers, F., and Loeser, A. (2021). Clinical outcome prediction from admission notes using self-supervised knowledge integration. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online. Association for Computational Linguistics.

Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, K., Suter, P. M., and Thijs, L. G. (1996). The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22:707–710.