Bias in Danish Medical Notes: Infection Classification of Long Texts Using Transformer and LSTM Architectures Coupled with BERT

Mehdi Parviz[1], Rudi Agius[2], Carsten Utoft Niemann[2], Rob van der Goot[3],
[1]Department of Biology, University of Copenhagen, Denmark
[2]Department of Hematology, Copenhagen University Hospital, Rigshospitalet, Denmark
[3] Department of Computer Science, IT University of Copenhagen, Denmark,
mehdi.parviz@bio.ku.dk,
{rudi.agius.01, carsten.utoft.niemann}@regionh.dk
robv@itu.dk

Abstract

Medical notes contain a wealth of information related to diagnosis, prognosis, and overall patient care that can be used to help physicians make informed decisions. However, like any other data sets consisting of data from diverse demographics, they may be biased toward certain subgroups or subpopulations. Consequently, any bias in the data will be reflected in the output of the machine learning models trained on them. In this paper, we investigate the existence of such biases in Danish medical notes related to three types of blood cancer, with the goal of classifying whether the medical notes indicate severe infection. By employing a hierarchical architecture that combines a sequence model (Transformer and LSTM) with a BERT model to classify long notes, we uncover biases related to demographics and cancer types. Furthermore, we observe performance differences between hospitals. These findings underscore the importance of investigating bias in critical settings such as healthcare and the urgency of monitoring and mitigating it when developing AI-based systems.

1 Introduction

Electronic Health Records (EHRs) provide diverse data on diagnoses, medications, and clinical tests, enabling AI-based applications for various purposes (Wang and Zhang, 2024). While medical notes contain similar information in an unstructured format, they offer deeper insights that complement other EHR data. They help cross-check information, retrieve missing details, and capture clinically relevant events like infections, which are often difficult to extract from structured EHR sources. Assessing EHRs and medical notes aids physicians in making informed decisions on treatments, medications, and patient care. Notably, prior infections are key predictors of clinical outcomes in blood cancers (Parviz et al., 2022; Packness et al., 2024). However, biases in EHR-derived medical data have been documented and can lead to performance deterioration in subpopulations with smaller sample sizes (Cobert et al., 2024). In this paper, we classify medical notes on three common blood cancers based on infection status and quantify bias related to sex and cancer type. The cancers studied are lymphoma (LYFO), multiple myeloma (MM), and chronic lymphocytic leukemia (CLL). Since medical notes often exceed model context lengths, we employ a hierarchical architecture combining a sequence model (Transformer and LSTM) with a BERT model (Pappagari et al., 2019).

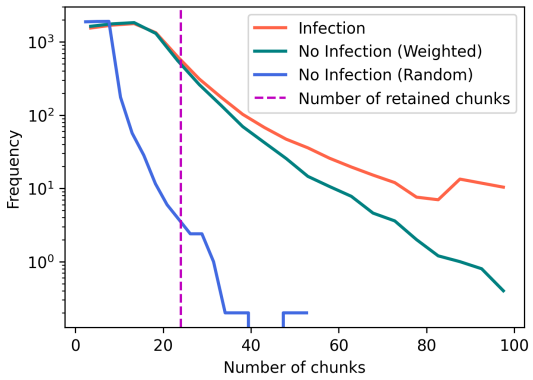

Figure 1: Document length distribution for each class before resampling, and after weighted, and random resampling. The dashed purple line indicates the number of chunks retained for modeling.

2 Method

2.1 Data

We curated a dataset of medical notes from patients diagnosed with lymphoma, CLL, or MM in Eastern Denmark, recorded between August 2016 and November 2023. For each patient, notes recorded less than two days apart were merged, as they often related to the same health-related issue. This information was extracted from data sources available through the DALY-CARE database (Brieghel et al., 2025).

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 316–320

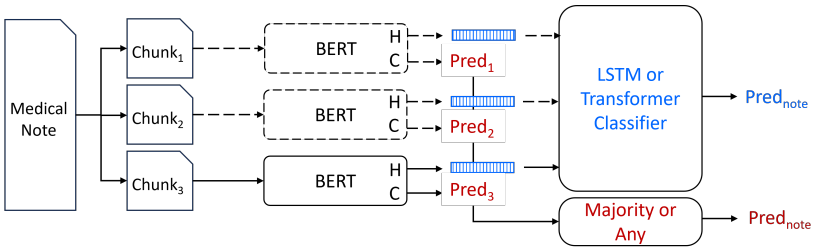

Figure 2: Schematic of modeling and data splitting strategy. A BERT model (red) is coupled with a transformer or an LSTM (blue) architecture to capture information in long medical notes.

2.2 Infection definition

While EHRs provide valuable medical information, identifying severe infections is not always straightforward. Therefore, severe infection was defined as a blood culture draw and intravenous (IV) antimicrobial administration occurring within two days. Clinically, blood cultures are taken when an infection is suspected, and IV antimicrobials are given in severe cases. Defining the outcome by both events enhances labeling precision and the likelihood that physicians mention severe infections in medical notes.

2.3 Modeling long medical notes

Due to the limited context of the BERT models, we divide medical notes into smaller chunks that fit within the maximum token limit of BERT (512). Each chunk is then assigned the same label as the full medical note. We adopt a similar approach to that of (Pappagari et al., 2019), which is presented in Figure 2. First, we fine-tune a BERT model trained on Danish medical data (Pedersen et al., 2023) to predict chunk labels. Next, we extract embeddings for each chunk from the last hidden state of BERT and model the chunk embedding sequences using either a Transformer (Vaswani et al., 2017) or an LSTM architecture (Hochreiter and Schmidhuber, 1997). We also compare the performance of these stacked methods with simpler approaches that return the chunk-level majority prediction and any (positive) prediction from BERT, which we refer to as MAJORITY and ANY.

2.4 Sensitivity to note length

We found a significant discrepancy in medical note length between classes (Figure 1); notes labeled as infection were longer than those without infection. To prevent the model from using length as a proxy for the outcome, we resample negative-class notes using a weighted approach to match the length distribution of the positive class (Figure 1). We evaluate models using both weighted resampling (Weighted) and random sampling (Random), which occur in real scenarios where one class has significantly shorter notes.

2.5 Measuring bias in subgroups

Following (Czarnowska et al., 2021), we assess potential biases in model predictions related to sex and cancer type using the false positive rate (FPR) and false negative rate (FNR). If the models are biased toward a subgroup, we expect a lower FNR and/or higher FPR compared to the other group(s). We perform binomial tests to determine whether the differences between subgroups and the majority class (male for the sex factor and lymphoma for cancer types) are statistically significant. The null hypothesis assumes that predictions for minority subgroups follow the same distribution as those for the majority subgroup.

2.6 Data splitting

To minimize data leakage or biases related to the memorization of physician-specific information (e.g., writing style or specialties) and patient history during data splitting, we ensure that training, validation, and test sets come from different hospitals. Specifically, two hospitals are used for training, the third for validation and testing, and the process is repeated for all three combinations. Figures 3 and 4 illustrate the distribution of notes across subgroups, as well as the proportion of notes labeled as infection in the training, validation, and test splits. To mitigate dataset imbalance, we resample the training set to ensure it contains an equal number of notes across female and male subgroups, cancer types, and positive and negative classes (Balanced). Since medical notes are recorded at different time points, they must be treated as a time series. Therefore, we use a time-based splitting approach when dividing the data into training, validation, and test sets. The models were trained with a learning rate

of 2×10^{-5}, a batch size of 32, and for one epoch. These parameters were selected based on preliminary experiments on the validation set.

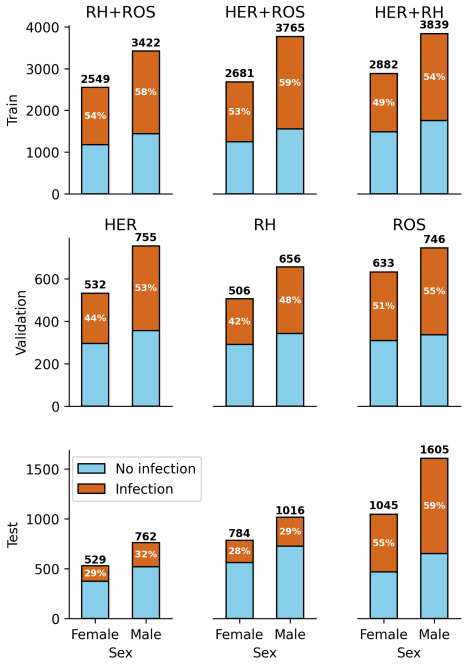

Figure 3: Notes by sex across hospitals (columns) and sets (rows). Total notes are shown atop each bar, with infection-positive percentages inside.

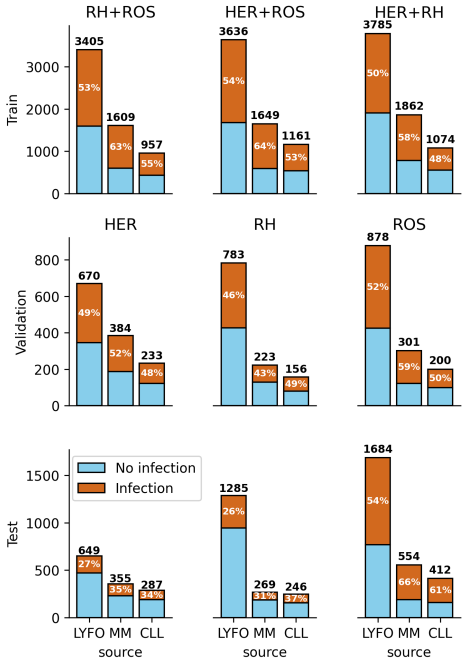

Figure 4: Notes by cancer type across hospitals (columns) and sets (rows). Total notes are shown atop each bar, with infection-positive percentages inside.

3 Results

3.1 Model performances

The results show that, overall, using sequence modeling (Transformer or LSTM) outperforms the simpler MAJORITY and ANY models at the classifier layer. Additionally, coupling a Transformer with the base BERT model performs better than coupling BERT with an LSTM (Table 1). All models tend to overclassify samples as infections, as evidenced by higher FPRs than FNRs. The FPRs of the two sampling strategies indicate that, despite being trained on artificially longer negative samples, the models achieve the same performance level on shorter texts observed in the dataset.

3.2 Variation in error rates by sex

The results in Table 2 show that, on both the Validation and test sets, FNR values remain at similar levels between males and females across the three hospitals. Without resampling (Observed), both FPR_W and FPR_R are significantly lower for females in two out of three hospitals in the test set. Although resampling (Balanced) eliminates sex differences in FPR_R, FPR_W values remain significantly lower for females. This disparity suggests that the lower FPRs observed for females may be influenced by their under-representation in the dataset, leading to biased model predictions (Figure 3). Additionally, other sources of bias, such as differences in clinical documentation patterns may have further contributed to these discrepancies.

3.3 Variation in error rates by cancer type

The results on cancer types show higher FPR_W and FPR_R on MM compared with LYFO consistently across the three hospitals (Table 3). In the test set, MM and CLL both have worse FPR_R compared to LYFO. Resampling based on cancer type has little effect on reducing the significant differences. These results highlight that the models are biased against underrepresented subgroups (Figure 4).

4 Conclusion

Medical notes supplement EHRs with information not available in structured formats. Unlike other EHR data, which are automatically compiled, medical notes are written by physicians and nurses, making them more prone to bias. In this paper, we explore potential sources of bias within the demographic population and across three types of blood cancer. The results indicate biases related to sex and among different cancer types. We also

		Validation			Test		
Sampling	Model	H_{HER}	H_{RH}	H_{ROS}	H_{HER}	H_{RH}	H_{ROS}
Weighted	ANY	66.1	72.3	69.3			
	MAJORITY	81.9	76.5	83.4			
	LSTM	85.5	81.9	82.7			
	Transformer	**86.4**	**83.9**	**85.5**	78.3	81.2	83.4
Random	ANY	83.3	**88.6**	**89.1**			
	MAJORITY	80.9	78.4	86.7			
	LSTM	81.8	77.4	84.7			
	Transformer	**84.4**	84.8	86.9	84.8	84.1	84.6

Table 1: Infection classification performance of the models, measured using balanced accuracy, on the validation and test sets constructed with weighted and random sampling across hospitals.

			Validation			Test		
Balance Method	Metric	Sex	H_{HER}	H_{RH}	H_{ROS}	H_{HER}	H_{RH}	H_{ROS}
Observed	FNR	F	6.8	11.6$^{\bullet}$	5.9$^{\bullet}$	5.8	9.0	6.9
		M	6.3	8.6	8.1	4.5	8.3	8.1
	FPR_R	F	25.3	18.7	19.4	25.9	20.2*	20.9*
		M	24.2	22.2	18.8	25.1	25.9	25.0
	FPR_W	F	20.6	21.0	22.3	35.0*	25.3*	24.0
		M	21.0	23.6	21.7	40.8	31.8	26.6
Balanced	FNR	F	10.2	21.4	13.3*	5.2	13.5	14.5
		M	9.5	19.5	17.6	4.1	15.2	15.3
	FPR_R	F	25.3	13.6	11.1$^{\bullet}$	24.6	18.0	15.2
		M	23.1	16.7	8.5	22.2	16.5	16.2
	FPR_W	F	22.6	15.8	17.5	32.6$^{\bullet}$	20.3*	16.7$^{\bullet}$
		M	22.1	17.5	14.8	36.8	26.4	19.8

Table 2: Infection classification performance of the models in male and female subpopulations, measured using FPR and FNR, on the validation and test sets constructed via weighted and random sampling across hospitals. P-values are calculated using binomial test ($\cdot$ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001).

			Validation			Test		
Balance Method	Metric	Cancer	H_{HER}	H_{RH}	H_{ROS}	H_{HER}	H_{RH}	H_{ROS}
Observed	FNR	CLL	5.4	10.5	5.9	5.2	8.9	8.0
		LYFO	6.8	10.1	8.4	4.0	9.1	8.2
		MM	6.6	8.4	4.5*	6.5	6.0	6.0$^{\bullet}$
	FPR_R	CLL	21.2	20.0	16.2	30.3**	30.6**	27.9*
		LYFO	20.8	17.7	15.3	19.1	19.9	20.6
		MM	34.0**	29.0**	29.9***	32.4***	29.4**	25.8*
	FPR_W	CLL	19.7	25.0$^{\bullet}$	20.2	38.9	29.5	24.2
		LYFO	18.3	18.5	20.0	36.7	27.9	24.4
		MM	26.3*	33.6***	30.3**	41.6$^{\bullet}$	33.9$^{\bullet}$	31.2*
Balanced	FNR	CLL	15.3	23.7	23.8	9.3	16.7	19.9
		LYFO	12.3	24.1	21.6	6.8	17.7	21.1
		MM	14.1	26.3	19.6	12.2*	13.3	18.9
	FPR_R	CLL	21.2	16.4	13.1	25.5**	21.1**	27.4*
		LYFO	18.2	13.3	18.0	14.6	11.4	21.0
		MM	25.5*	14.5	35.7***	26.7***	14.4	35.9***
	FPR_W	CLL	16.4	20.0$^{\bullet}$	14.1	31.6	20.5	18.6
		LYFO	16.2	13.4	13.2	30.9	20.1	15.8
		MM	23.1*	23.4**	18.9$^{\bullet}$	34.2	26.9*	22.8**

Table 3: Infection classification performance of the models across different cancer subpopulations, measured using FPR and FNR, on the validation and test sets constructed via weighted and random sampling across hospitals.

observe variations in classification performance across hospitals, highlighting the need for further investigation into potential differences. These discrepancies may stem from variations in data quality or differences in how information related to severe infections is recorded.

5 Limitations

One limitation of this study is the model's shorter context than the input documents. Future work could explore longer-context models like Longformer for improvement (Beltagy et al., 2020).

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Christian Brieghel, Mikkel Werling, Casper Møller Frederiksen, Mehdi Parviz, Thomas Lacoppidan, Tereza Faitova, Rebecca Svanberg Teglgaard, Noomi Vainer, Caspar da Cunha-Bang, Emelie Curovic Rotbain, Rudi Agius, and Carsten Utoft Niemann. 2025. The danish lymphoid cancer research (daly-care) data resource: The basis for developing data-driven hematology. *Clinical Epidemiology*, 17:131–145.

Julien Cobert, Hunter Mills, Albert Lee, Oksana Gologorskaya, Edie Espejo, Sun Young Jeon, W John Boscardin, Timothy A Heintz, Christopher J Kennedy, Deepshikha C Ashana, et al. 2024. Measuring implicit bias in icu notes using word-embedding neural network models. *Chest*, 165(6):1481–1490.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Esben Packness, Olafur Birgir Davidsson, Klaus Rostgaard, Michael Asger Andersen, Emelie Curovic Rotbain, Carsten Utoft Niemann, Christian Brieghel, and Henrik Hjalgrim. 2024. Infections and their prognostic significance before diagnosis of chronic lymphocytic leukemia, non-hodgkin lymphoma, or multiple myeloma. *British Journal of Cancer*.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844.

Mehdi Parviz, Christian Brieghel, Rudi Agius, and Carsten Utoft Niemann. 2022. Prediction of clinical outcome in cll based on recurrent gene mutations, cll-ipi variables, and (para)clinical data. *Blood Advances*, 6:3716–3728.

Jannik Pedersen, Martin Laursen, Pernille Vinholt, and Thiusius Rajeeth Savarimuthu. 2023. MeDa-BERT: A medical Danish pretrained transformer model. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 301–307, Tórshavn, Faroe Islands. University of Tartu Library.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Dandan Wang and Shiqing Zhang. 2024. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial Intelligence Review*, 57(11):299.