Investigating anatomical bias in clinical machine learning algorithms

Jannik Skyttegaard Pedersen[*]
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
jasp@mmmi.sdu.dk

Martin Sundahl Laursen[*]
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
msla@mmmi.sdu.dk

Pernille Just Vinholt
Department of Clinical Biochemistry
Odense University Hospital

Anne Bryde Alnor
Department of Clinical Biochemistry
Odense University Hospital

Thiusius Rajeeth Savarimuthu
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark

Abstract

Clinical machine learning algorithms have shown promising results and could potentially be implemented in clinical practice to provide diagnosis support and improve patient treatment. Barriers for realisation of the algorithms' full potential include bias which is systematic and unfair discrimination against certain individuals in favor of others.

The objective of this work is to measure *anatomical bias* in clinical text algorithms. We define anatomical bias as unfair algorithmic outcomes against patients with medical conditions in specific anatomical locations. We measure the degree of anatomical bias across two machine learning models and two Danish clinical text classification tasks, and find that clinical text algorithms are highly prone to anatomical bias. We argue that datasets for creating clinical text algorithms should be curated carefully to isolate the effect of anatomical location in order to avoid bias against patient subgroups.

1 Introduction

Research in clinical machine learning algorithms have shown promising results for automating clinical tasks. The algorithms could potentially be implemented in clinical practice to provide diagnosis support, improve patient treatment and provide time-savings for medical doctors (Topol, 2019; Matheny et al., 2020).

However, despite appealing research results, there are currently limited examples of algorithms being successfully deployed into clinical practice (Kelly et al., 2019). Barriers for realisation of the algorithms' full potential include bias and generalisation issues (Char et al., 2018; Hovy and Prabhumoye, 2021; Carrell et al., 2017).

Algorithmic bias can be defined as systematic and unfair discrimination against certain individuals or groups of individuals in favor of others (Friedman and Nissenbaum, 1996). Previous studies have raised serious concerns of algorithms that contain age, gender and racial bias (Sun et al., 2019; Davidson et al., 2019) — even for algorithms that have been taken into use (Obermeyer et al., 2019). Although machine learning algorithms are trained to be able to generalise to previously unseen data, they tend to overfit to the data they have been trained on. As a consequence of this, bias can unintendedly arise if some subgroups of the target population are not represented in the data used to train the algorithm. Moreover, if the training data itself include biases against some populations, e.g. data reflecting a negative attitude against people with disabilities (Hutchinson et al., 2020), these biases might be encoded and reinforced.

If biased algorithms are adopted, healthcare systems risk doing injustice to certain patient groups and harming patient safety (Obermeyer et al., 2019). Therefore, identifying and mitigating bias is important for successful implementation of novel clinical machine learning algorithms.

This paper investigates *anatomical bias* in clinical machine learning algorithms developed to classify and extract specific medical conditions from the narrative text of electronic health records (EHR). We define anatomical bias as unfair algorithmic outcomes against a subgroup of patients with the same medical condition, where the algorithm performs differently depending on the anatomical location of the condition. If the performance of clinical algorithms varies depending

[*] Equal contribution

Findings of the Association for Computational Linguistics: EACL 2023, pages 1398–1410
May 2-6, 2023 ©2023 Association for Computational Linguistics

on the anatomical location, it is reflected in some patient subgroups receiving worse treatment than others.

We hypothesised that careful dataset curation is needed to measure and mitigate anatomical bias because the text description of medical conditions in EHRs varies depending on the location, e.g. 'epistaxis' is a location-specific word describing nose bleedings.

Specifically, this paper investigates anatomical bias for classification of bleeding and venous thromboembolism (VTE) mentions in the narrative text of Danish EHRs. Automatic extraction of these conditions could be valuable for medical doctors in clinical practice, e.g. to guide diagnostic decision making and treatment options (Decousus et al., 2011). Previous papers (Hinz et al., 2013; Lee et al., 2017; Taggart et al., 2018; Li et al., 2019; Mitra et al., 2020, 2021; Elkin et al., 2021; Pedersen et al., 2021; Shi et al., 2021; Verma et al., 2022) have shown promising results for automatic extraction of these medical conditions but they did not investigate the performance of the algorithms across anatomical subgroups.

Our main contributions are:

- We find that clinical text algorithms are highly prone to anatomical bias.
- The performance of state-of-the-art algorithms developed to extract specific medical conditions varies significantly across anatomical locations with performance drops up to 89.1 percentage points (PP).
- We argue that datasets for creating clinical text algorithms should be curated carefully to isolate the effect of anatomical location in order to avoid bias against patient subgroups.

2 Methods

To investigate if machine learning algorithms are prone to anatomical bias, we performed two experiments. We (1) investigated the performance of a binary classifier on different anatomical subgroups of a medical condition when that subgroup was left out of the training set, and (2) measured how the performance on an anatomical subgroup varied depending on the amount of samples from that subgroup included in the training set.

Table 1: Distribution of the training, validation, and test samples for the balanced bleeding detection dataset.

Label	Location	Train	Validation	Test
Positive for bleeding	Gastrointestinal	750	250	250
	Urogenital	750	250	250
	Internal	750	250	250
	Otorhinolaryngeal	750	250	250
	Dermatological	750	250	250
	Gynecological	750	250	250
	Cerebral	750	250	250
	Ophthalmological	750	250	250
Negative for bleeding		6,000	2,000	2,000
Sum		12,000	4,000	4,000

Table 2: Distribution of the training, validation, and test samples for the balanced VTE detection dataset.

Label	Location	Train	Validation	Test
Positive for VTE	Lower extremity	1,600	200	200
	Lung	1,600	200	200
	Liver	0	0	239
	Cerebral	0	0	218
	Upper extremity	0	0	176
Negative for VTE		3,200	400	1,033
Sum		6,400	800	2,066

2.1 Datasets

We used the binary bleeding classification dataset from Pedersen et al. (2022) and present a new binary VTE classification dataset. The bleeding dataset consists of 20,000 sentences from Danish EHRs labeled as either positive or negative for bleeding mentions. The VTE classification dataset consists of 9,266 sentences from Danish EHRs labeled as either positive or negative for VTE mentions. Both datasets were constructed from Danish EHRs from Odense University Hospital and were labeled with a consensus label from three medical doctors.

In addition to the main labels of each dataset (positive and negative for bleeding or VTE), we created a subgroup label for the positive samples describing the anatomical location of either the bleeding or VTE mention. Samples that did not describe the anatomical location or described multiple locations were omitted.

For the bleeding dataset, we used the following eight anatomical locations: gastrointestinal, urogenital, internal, otorhinolaryngeal, dermatological, gynecological, cerebral, and ophthalmological.

For the VTE dataset, we used the following five anatomical locations: lower extremity, lung, liver, cerebral, and upper extremity.

The locations included for each medical condition were selected by two medical doctors.

For each dataset, we created a balanced training, validation, and test set containing an equal amount of positive and negative samples. Moreover, for the bleeding dataset, the positive samples of the training, validation, and test sets were distributed equally between anatomical locations. For the VTE dataset, only samples from the lower extremity and lung locations were distributed equally between the train, validation, and test sets. The liver, cerebral, and upper extremity locations were only used for the test set because of a limited number of samples.

All samples were preprocessed by removing special characters, superfluous spaces, and duplicate samples. After preprocessing the samples, the bleeding and VTE datasets had an average token length of 13.3 and 13.6, respectively. The dataset distributions can be seen in Table 1 and Table 2.

2.2 Training set distributions

To measure performance differences for specific anatomical locations, we systematically removed all samples from a specific location, x, from the training set, creating the training set $\mathcal{T}_{\not\subset x}$, trained a deep learning model on $\mathcal{T}_{\not\subset x}$, and evaluated it on the test set. For example, for the bleeding dataset, we created 8 different training sets, one for each anatomical location being removed, containing 10,500 samples.

For comparison, we created a balanced training set, $\mathcal{T}$, which included the same amount of samples as $\mathcal{T}_{\not\subset x}$, distributed equally between the positive and negative classes, and between anatomical locations.

2.3 Deep learning models

The deep learning models were a transformer-based ELECTRA model (Clark et al., 2020) and a Long Short-Term Memory (LSTM) model (Hochreiter and Schmidhuber, 1997).

The ELECTRA model was a Danish clinical ELECTRA (Clin-ELECTRA) (Pedersen et al., 2022) pretrained on the narrative text from 299,718 EHRs from Odense University Hospital. The model had ~13M parameters and consisted of 12 transformer layers with 4 attention heads. We initialised Clin-ELECTRA from its pretrained checkpoint and followed the HuggingFace (Wolf et al., 2019) implementations for binary text classification.

The LSTM model had ~4M parameters and consisted of a bidirectional LSTM layer with a hidden layer size of 512. The last hidden state of the LSTM was followed by a dropout layer with probability 0.2, a dense layer of size 256, a ReLU activation function, a dropout layer of probability 0.2, and a dense classification layer. For word representation, the LSTM model used 300-dimensional FastText (Bojanowski et al., 2017) word embeddings pretrained on Danish EHRs consisting of 1.4B tokens.

2.4 Model evaluation

For each of the ELECTRA and LSTM deep learning models and training sets $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$, we:

1. Trained the deep learning model with five different learning rates and random initialisations.

2. Computed the test set accuracy of the best performing model based on the loss on the validation set.

3. Repeated step 1 and 2 five times.

We used the five accuracies to perform bootstrapping with 9,999 replicates and calculated mean accuracy, standard error (SE), and 95% confidence interval (CI) for $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$. Moreover, we computed the bootstrapped difference of means between $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$ to evaluate statistically significant differences in performance.

For both deep learning models, we used the Adam optimizer (Kingma and Ba, 2015) and searched for the best model using learning rates 7e-5, 8e-5, 9e-5, and 1e-4. Clin-ELECTRA was finetuned for a maximum of 10 epochs and the LSTM for a maximum of 30 epochs. One epoch was trained in <1 and ~5 seconds for the LSTM and ELECTRA model, respectively, using an NVIDIA v100 GPU.

We measured anatomical bias as the difference in sensitivity on a specific location, x, between two deep learning models trained on $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$.

3 Results

3.1 Bleeding classification

Table 3 shows the binary accuracy of the bleeding classifiers on the test set for each of the training sets $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$. Appendix A shows additional metrics. With the exception of $\mathcal{T}_{\not\subset Otorhinolaryngeal}$ for the ELECTRA model, all training sets with an anatomical location removed resulted in models with significantly worse performance than when trained on $\mathcal{T}$.

The decreases in accuracy were caused by a significant drop in sensitivity for the anatomical locations which had been removed from the training

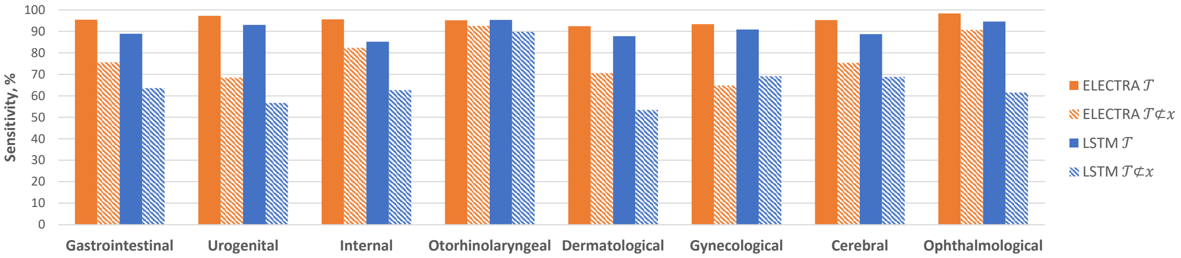


Figure 1: Sensitivity of models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$ for each anatomical location, x.

Table 3: Accuracy (%), standard error (SE), and 95% confidence interval (CI) for the bleeding classification dataset. $\mathcal{T}_{\not\subset x}$ denotes the training set from which an anatomical location, x, has been removed. * denotes a significant difference at the 0.05 level between models trained on $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$.

	ELECTRA		LSTM	
	Accuracy±SE	CI	Accuracy±SE	CI
$\mathcal{T}$	95.6 ± 0.1	95.4 - 95.8	90.6 ± 0.1	90.4 - 90.7
$\mathcal{T}_{\not\subset Gastrointestinal}$	94.4 ± 0.2*	94.0 - 94.7	88.8 ± 0.1*	88.6 - 88.9
$\mathcal{T}_{\not\subset Urogenital}$	93.9 ± 0.1*	93.7 - 94.2	88.5 ± 0.1*	88.3 - 88.6
$\mathcal{T}_{\not\subset Internal}$	95.0 ± 0.1*	94.8 - 95.2	88.8 ± 0.1*	88.5 - 89.0
$\mathcal{T}_{\not\subset Otorhinolaryngeal}$	95.6 ± 0.1	95.4 - 95.8	90.0 ± 0.2*	89.8 - 90.4
$\mathcal{T}_{\not\subset Dermatological}$	94.3 ± 0.1*	94.0 - 94.5	88.2 ± 0.1*	88.1 - 88.3
$\mathcal{T}_{\not\subset Gynecological}$	93.8 ± 0.1*	93.6 - 94.0	89.1 ± 0.1*	88.9 - 89.4
$\mathcal{T}_{\not\subset Cerebral}$	94.6 ± 0.2*	94.3 - 94.9	89.0 ± 0.1 *	88.8 - 89.2
$\mathcal{T}_{\not\subset Ophthalmological}$	95.3 ± 0.1*	95.1 - 95.4	88.2 ± 0.2 *	87.8 - 88.5

Table 4: Accuracy (%), standard error (SE), and 95% confidence interval (CI) for the VTE classification dataset. $\mathcal{T}_{\not\subset x}$ denotes the training set from which an anatomical location, x, has been removed. * denotes a significant difference at the 0.05 level between models trained on $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$.

	ELECTRA		LSTM	
	Accuracy±SE	CI	Accuracy±SE	CI
$\mathcal{T}$	84.8 ± 0.3	84.2 - 85.4	75.9 ± 0.3	75.4 - 76.4
$\mathcal{T}_{\not\subset Lower\ extremity}$	67.6 ± 0.6*	66.5 - 68.7	71.6 ± 0.4*	70.8 - 72.6
$\mathcal{T}_{\not\subset Lung}$	74.0 ± 1.1*	71.9 - 76.1	69.9 ± 0.1*	69.7 - 70.2

data. Figure 1 shows the test set sensitivity for each anatomical location, x, for each training set $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$. The sensitivity for all anatomical locations was significantly worse when not present in the training set with performance drops up to 28.8 PP for ELECTRA and 36.3 PP for the LSTM model. On average, the sensitivity on the anatomies decreased with 17.8 PP (standard deviation ± 8.8 PP) for ELECTRA and 24.5 PP (standard deviation ± 9.4 PP) for the LSTM model.

Moreover, it is seen that even though models trained on $\mathcal{T}_{\not\subset x}$ achieved high accuracies on the test set overall, the sensitivity on the anatomical location not present in the training set was low. E.g., for ELECTRA, $\mathcal{T}_{\not\subset Gynecological}$ had a 93.8% accuracy on the test set, but the sensitivity for gynecological bleedings was only 64.8%. Appendix A shows the sensitivity, SE, and the differences of means for all anatomical locations and training sets.

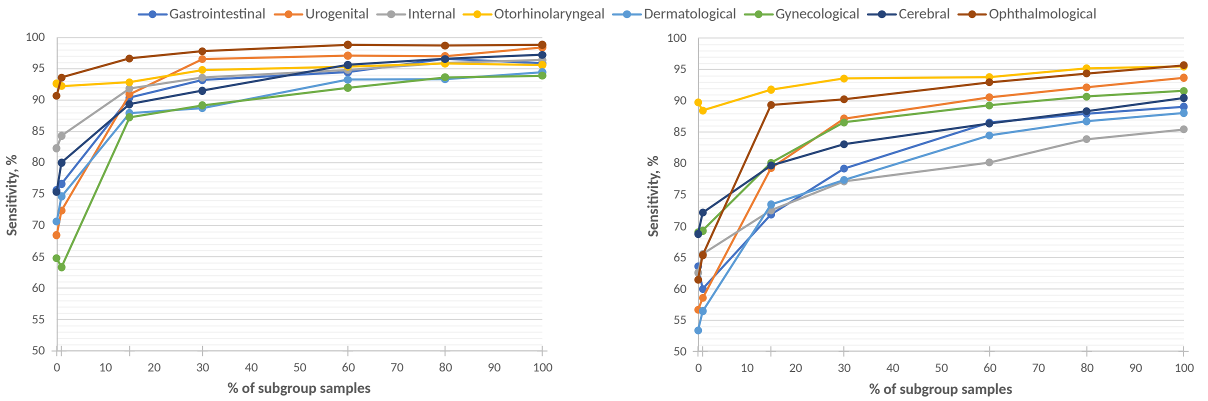
Figure 2 shows the sensitivity on each anatomical location by the percentage of total subgroup samples in the training set. It is seen that the accuracy increased as more samples were present in the training set. For the LSTM model, the sensitivity on gastrointestinal, urogenital, cerebral, and ophthalmological bleedings was significantly worse - even when 80% of samples were present in the training set. For ELECTRA, the sensitivity on urogenital and internal bleedings was significantly worse when 80% of samples were present in the training set. Appendix A shows the accuracies and differences of means.

3.2 Venous thromboembolism classification

Table 4 shows the binary accuracy of the VTE classifiers on the test set for each of the training sets $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$. Appendix B shows additional metrics. Models trained on $\mathcal{T}_{\not\subset Lower\ extremity}$ and $\mathcal{T}_{\not\subset Lung}$ performed significantly worse than those trained on $\mathcal{T}$.

Similar to the bleeding classifier results, the decrease in the overall accuracy was caused by a significant drop in sensitivity on the anatomical locations which had been removed from the training data. Figure 3 shows the sensitivity for each anatomical location, x, for each training set $\mathcal{T}$ and $\mathcal{T}_{\not\subset x}$. The sensitivity on liver, cerebral, and lower extremity VTEs is only reported when not being present in the training set because of limited samples.

The sensitivity on lower extremity and lung VTEs was significantly worse when not present in the training set, e.g. the performance for the ELECTRA classifier decreased by 89.1 PP for lower extremity VTEs and 81.0 PP for lung VTEs. Appendix B shows the sensitivity, SE, and the differences of means for all anatomical locations and



(a) ELECTRA

(b) LSTM

Figure 2: Test set sensitivity on the anatomical locations when removing a fraction of samples from that anatomy from the training set.

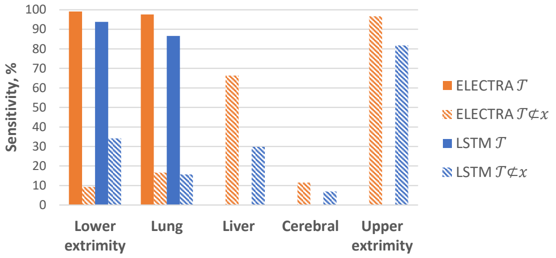


Figure 3: Sensitivity of models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$ for each anatomical location, x. The sensitivity on liver, cerebral, and lower extremity VTEs is only reported when not being present in the training set because of limited samples.

training sets.

Figure 4 shows the sensitivity for lower extremity and lung VTEs by the percentage of total subgroup samples in the training set. Both locations performed significantly worse when 15% and 30% of samples from that location were present in the training set for the ELECTRA and LSTM classifier, respectively. Appendix B shows the accuracies and differences of means.

Table 5: Most frequent words used to describe VTE events for each anatomical location of the VTE classification dataset. Words are translated from Danish to English and, therefore, some cells include two words. PE = pulmonary embolism.

Word	Frequency	Location uniqueness
Location: Lower extremity		
dvt	1384	0.92
thrombus	135	0.70
blood clot	108	0.47
Location: Lung		
pulmonary embolism	1058	0.98
le (PE)	483	0.99
pulmonary embolisms	242	0.99
Location: Liver		
porta thrombosis	71	1.00
thrombosis	70	0.36
thrombus	22	0.11
Location: Cerebral		
infarct	93	0.97
sinus thrombosis	32	0.97
blood clot	26	0.11
Location: Upper extremity		
dvt	99	0.07
thrombus	28	0.14
thrombosis	14	0.07

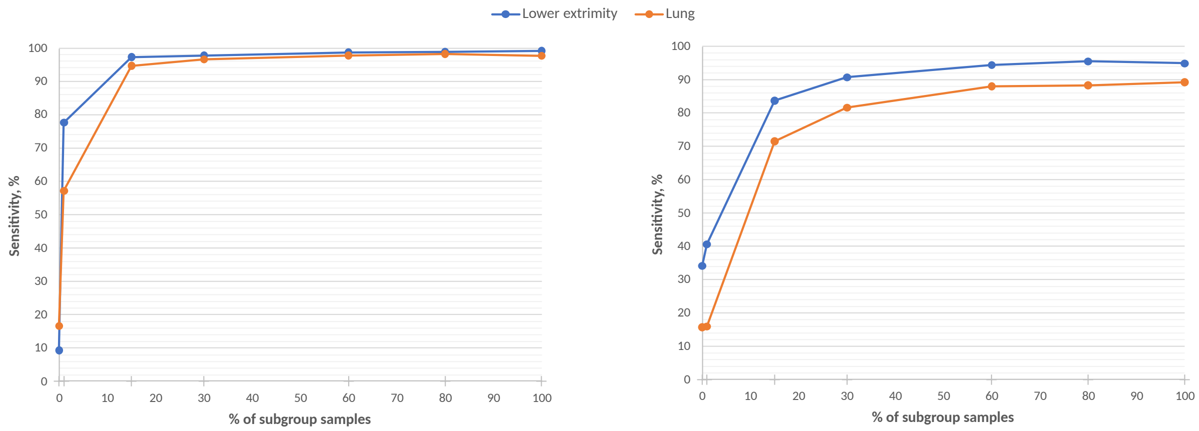
4 Analysis of word distributions

Medical conditions are often described using different words depending on the anatomical location for which the condition occurs. Table 5 shows the top-3 most frequent words used to describe VTE events for each anatomical location. The column *Location uniqueness* shows the fraction of times a word appears in samples from a specific anatomical location compared to the complete dataset:

$$Location\ uniqueness = \frac{f_x}{f_D} \tag{1}$$

where f_x is the frequency of the word in samples from anatomical location x and f_D is the total frequency of the word in the dataset, i.e. a value of 1 means that the word is unique for an anatomical location.

The top-3 words for upper extremity had a low uniqueness score (<0.15). This indicates that the vocabulary used to describe VTEs in the upper extremity was also used for other locations — e.g.



(a) ELECTRA

(b) LSTM

Figure 4: Test set sensitivity on the anatomical locations when removing a fraction of samples from that anatomy from the training set.

'dvt' (deep vein thrombosis) was the most frequent word but the uniqueness score was only 0.07. This might explain why the sensitivity for upper extremity was relatively high, as seen in Figure 3, even when samples from that location were not present in the training set. On the contrary, some of the frequent words from the lower extremity, lung, and cerebral locations were close to unique which could explain why the sensitivity of those locations were low. Appendix C shows word frequency and location uniqueness for the bleeding classification dataset which shows similar results.

5 Discussion

This paper has presented evidence that clinical natural language processing (NLP) classification algorithms are prone to anatomical bias, which is unfair algorithmic outcomes against patients with medical conditions in specific anatomical locations. We found that the performance of algorithms for both bleeding and VTE classification can vary significantly depending on the anatomical location with differences up to 36.3 PP and 89.1 PP, respectively.

Moreover, we found that small fluctuations in the training set distribution of anatomical locations can lead to significant performance drops for the underrepresented anatomical locations. For the datasets presented in this study, we showed that the words used to describe medical conditions vary depending on the anatomical location. If classifiers do not learn to properly represent the full vocabulary for describing a medical condition, its performance will decrease for some anatomical locations.

We argue that datasets for clinical NLP algorithms should be created to be able to carefully measure anatomical bias, e.g. by subdividing each sample into an anatomical location. This is essential to avoid implementing clinical algorithms that might discriminate against specific subgroups of patients. For example, one of the developed VTE classifiers in this study performed with sensitivities of >96% for VTEs in the lungs and lower extremity while it performed with a sensitivity of only 11.5% for cerebral VTEs. Applying such a model in clinical practice or research would provide unfair algorithmic outcomes against patients with cerebral VTEs. We also showed that an algorithm not exposed to gynecological bleedings would perform worse on this anatomical location. This would lead to unfair algorithmic outcomes against woman with gynecological bleedings. Similarly, because alcoholics have an increased prevalence of gastrointestinal bleedings (Singal et al., 2018), this group of people would have a higher risk of unfair algorithmic outcomes if the algorithm has not been trained on such bleeding locations.

To the best of our knowledge, anatomical bias has not been investigated in previous research. However, some studies tried to automatically create datasets distributed between different patient groups by extracting data based on International Classification of Diseases 10 (ICD) codes — e.g. Pedersen et al. (2021) extract data based on different bleeding disorders. While this approach could, to some degree, mitigate the problem, studies (Valkhoff et al., 2014; Øie et al., 2018) found that ICD codes have low accuracy and, therefore, this does not ensure an evenly distributed dataset.

Moreover, in order to isolate and measure the performance on different anatomical locations, the dataset should be constructed with a known distribution of these anatomies.

Our work is closely related to the field of domain adaption. For example, MacAvaney et al. (2017) find that an algorithm trained to extract temporal information from a specific patient population performs worse on another related patient population. Their results highlight that it is a challenging task to develop algorithms that can generalise well across domains. The main difference between our study and theirs is that the algorithms described in this paper are not developed to work on different domains. Rather, the algorithms are specifically developed to work on a specialised domain in the clinical field, e.g. bleeding detection. As our results have shown, the algorithms perform worse on some subpopulations of the population it is supposed to work on, and therefore, we describe this as a bias issue.

6 Conclusion

This paper presented evidence that clinical NLP algorithms are prone to anatomical bias. We found that the performance of clinical classification algorithms for both bleeding and VTE classification can vary significantly depending on the anatomical location of the medical condition. We argue that anatomical bias should be carefully examined when developing clinical text algorithms in order to avoid unfair algorithm performance against patient subgroups.

7 Limitations

Future work should investigate the degree of anatomical bias in other clinical areas and tasks, e.g. named entity recognition, to be able to compare the severity of the bias problem between algorithms and other clinical areas. Moreover, as the datasets used in this study are only from a single institution, the findings of the paper might not be widely representative.

The objective of this work was to stress the need for measuring anatomical bias. We leave it to future work to investigate algorithmic solutions other than dataset balancing for mitigating the problem, e.g. using techniques such as oversampling and data augmentation. Such techniques could also help mitigating anatomical bias in algorithms for which training set balancing is not sufficient.

The classification datasets and machine learning models presented in this paper cannot be shared publicly due to privacy concerns but we advise interested researchers to contact us for sharing possibilities.

Ethics Statement

Machine learning researchers must be proactive in recognising and counteracting biases such as the one described in this paper. We hope that the findings and focus of this paper will lead other researchers to test and mitigate other kinds of algorithmic biases.

All datasets used in this research were obtained according to each dataset's respective data usage policy. The datasets were stored and processed on a secure platform[1] in compliance with GDPR regulations. According to section 14(2) of the Danish Act on Research Ethics Review of Health Research Projects[2], studies using retrospective data that do not involve human biological material do not require ethical approval.

References

Piotr Bojanowski, Édouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

David S Carrell, Robert E Schoen, Daniel A Leffler, Michele Morris, Sherri Rose, Andrew Baer, Seth D Crockett, Rebecca A Gourevitch, Katie M Dean, and Ateev Mehrotra. 2017. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *Journal of the American Medical Informatics Association*, 24(5):986–991.

Danton S Char, Nigam H Shah, and David Magnus. 2018. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, 378(11):981.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35.

[1] https://docs.cloud.sdu.dk/intro/security.html

[2] https://www.retsinformation.dk/eli/lta/2011/593 (English version is unfortunately not available)

Hervé Decousus, Victor F Tapson, Jean-François Bergmann, Beng H Chong, James B Froehlich, Ajay K Kakkar, Geno J Merli, Manuel Monreal, Mashio Nakamura, Ricardo Pavanello, et al. 2011. Factors at admission associated with bleeding risk in medical patients: findings from the improve investigators. *Chest*, 139(1):69–79.

Peter L Elkin, Sarah Mullin, Jack Mardekian, Christopher Crowner, Sylvester Sakilay, Shyamashree Sinha, Gary Brady, Marcia Wright, Kimberly Nolen, JoAnn Trainer, et al. 2021. Using artificial intelligence with natural language processing to combine electronic health record's structured and free text data to identify nonvalvular atrial fibrillation to decrease strokes and death: Evaluation and case-control study. *Journal of medical Internet research*, 23(11):e28946.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3):330–347.

Eugenia R McPeek Hinz, Lisa Bastarache, and Joshua C Denny. 2013. A natural language processing algorithm to define a venous thromboembolism phenotype. In *AMIA Annual Symposium Proceedings*, volume 2013, page 975. American Medical Informatics Association.

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

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):1–9.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Hee-Jin Lee, Min Jiang, Yonghui Wu, Christian M Shaffer, John H Cleator, Eitan A Friedman, Joshua P Lewis, Dan M Roden, Josh Denny, and Hua Xu. 2017. A comparative study of different methods for automatic identification of clopidogrel-induced bleedings in electronic health records. *AMIA Summits on Translational Science Proceedings*, 2017:185.

Rumeng Li, Baotian Hu, Feifan Liu, Weisong Liu, Francesca Cunningham, David D McManus, Hong Yu, et al. 2019. Detection of bleeding events in electronic health record notes using convolutional neural network models enhanced with recurrent neural network autoencoders: deep learning approach. *JMIR medical informatics*, 7(1):e10788.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2017. Guir at semeval-2017 task 12: a framework for cross-domain clinical temporal information extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1024–1029.

Michael E Matheny, Danielle Whicher, and Sonoo Thadaney Israni. 2020. Artificial intelligence in health care: a report from the national academy of medicine. *Jama*, 323(6):509–510.

Avijit Mitra, Bhanu Pratap Singh Rawat, David McManus, Alok Kapoor, and Hong Yu. 2020. Bleeding entity recognition in electronic health records: A comprehensive analysis of end-to-end systems. In *AMIA Annual Symposium Proceedings*, volume 2020, page 860. American Medical Informatics Association.

Avijit Mitra, Bhanu Pratap Singh Rawat, David D McManus, Hong Yu, et al. 2021. Relation classification for bleeding events from electronic health records using deep learning systems: an empirical study. *JMIR medical informatics*, 9(7):e27527.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Lise R Øie, Mattis A Madsbu, Charalampis Giannadakis, Anders Vorhaug, Heidi Jensberg, Øyvind Salvesen, and Sasha Gulati. 2018. Validation of intracranial hemorrhage in the norwegian patient registry. *Brain and behavior*, 8(2):e00900.

Jannik S Pedersen, Martin S Laursen, Thiusius Rajeeth Savarimuthu, Rasmus Søgaard Hansen, Anne Bryde Alnor, Kristian Voss Bjerre, Ina Mathilde Kjær, Charlotte Gils, Anne-Sofie Faarvang Thorsen, Eline Sandvig Andersen, et al. 2021. Deep learning detects and visualizes bleeding events in electronic health records. *Research and practice in thrombosis and haemostasis*, 5(4):e12505.

Jannik S Pedersen, Martin S Laursen, Cristina Soguero-Ruiz, Thiusius R Savarimuthu, Rasmus Søgaard Hansen, and Pernille J Vinholt. 2022. Domain over size: Clinical electra surpasses general bert for bleeding site classification in the free text of electronic health records. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE.

Jianlin Shi, John F Hurdle, Stacy A Johnson, Jeffrey P Ferraro, David E Skarda, Samuel RG Finlayson, Matthew H Samore, and Brian T Bucher. 2021. Natural language processing for the surveillance of postoperative venous thromboembolism. *Surgery*, 170(4):1175–1182.

Ashwani K Singal, Ramon Bataller, Joseph Ahn, Patrick S Kamath, and Vijay H Shah. 2018. Acg clinical guideline: alcoholic liver disease. *The American journal of gastroenterology*, 113(2):175.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640.

Maxwell Taggart, Wendy W Chapman, Benjamin A Steinberg, Shane Ruckel, Arianna Pregenzer-Wenzler, Yishuai Du, Jeffrey Ferraro, Brian T Bucher, Donald M Lloyd-Jones, Matthew T Rondina, et al. 2018. Comparison of 2 natural language processing methods for identification of bleeding among critically ill patients. *JAMA network open*, 1(6):e183451–e183451.

Eric Topol. 2019. *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK.

Vera E Valkhoff, Preciosa M Coloma, Gwen MC Masclee, Rosa Gini, Francesco Innocenti, Francesco Lapi, Mariam Molokhia, Mees Mosseveld, Malene Schou Nielsson, Martijn Schuemie, et al. 2014. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. *Journal of clinical epidemiology*, 67(8):921–931.

Amol A Verma, Hassan Masoom, Chloe Pou-Prom, Saeha Shin, Michael Guerzhoy, Michael Fralick, Muhammad Mamdani, and Fahad Razak. 2022. Developing and validating natural language processing algorithms for radiology reports compared to icd-10 codes for identifying venous thromboembolism in hospitalized medical patients. *Thrombosis Research*, 209:51–58.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Bleeding classification results

Table 6: Precision, recall, and F1 performance for the bleeding classification dataset.

Table 7: Sensitivity and standard error for all anatomical locations of the bleeding classification dataset.

Table 8: Bootstrapped 95% confidence intervals for difference of means between models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$ of the bleeding classification dataset.

Table 9: Bleeding test set sensitivity and standard error on an anatomical location by percentage of subgroup samples in the modified training set.

Table 10: Bootstrapped 95% confidence intervals for difference of means between models trained on a modified training set, including a percentage of subgroup samples, and models trained on the full training set, $\mathcal{T}$, of the bleeding classification dataset.

B VTE classification results

Table 11: Precision, recall, and F1 performance for the VTE classification dataset.

Table 12: Sensitivity and standard error for all anatomical locations of the VTE classification dataset.

Table 13: Bootstrapped 95% confidence intervals for difference of means between models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$ of the VTE classification dataset.

Table 14: VTE test set sensitivity and standard error on an anatomical location by percentage of subgroup samples in the modified training set.

Table 15: Bootstrapped 95% confidence intervals for difference of means between models trained on a modified training set, including a percentage of subgroup samples, and models trained on the full training set, $\mathcal{T}$, of the VTE classification dataset.

C Bleeding word distribution

Table 16: Most frequent words used to describe bleeding mentions for each anatomical location of the bleeding classification dataset.

Table 6: Precision, recall, and F1 performance for the bleeding classification dataset. $\mathcal{T}_{\not\subset x}$ denotes the training set from which an anatomical location, x, has been removed. SE = Standard error. CI = 95% confidence interval.

	ELECTRA			**LSTM**		
	Precision ± SE (CI)	**Recall ± SE (CI)**	**F1 ± SE (CI)**	**Precision ± SE (CI)**	**Recall ± SE (CI)**	**F1 ± SE (CI)**
$\mathcal{T}_{\not\subset Gastrointestinal}$	94.7 ± 0.2 (94.2 - 95.0)	94.2 ± 0.4 (93.4 - 95.0)	94.4 ± 0.2 (94.0 - 94.8)	89.1 ± 0.7 (8.77 - 90.5)	88.5 ± 0.7 (87.2 - 89.8)	88.8 ± 0.1 (88.6 - 88.9)
$\mathcal{T}_{\not\subset Urogenital}$	94.3 ± 0.3 (93.6 - 94.8)	93.5 ± 0.4 (92.6 - 94.3)	93.9 ± 0.1 (93.6 - 94.2)	89.5 ± 0.1 (89.3 - 89.8)	87.3 ± 0.2 (87.0 - 87.7)	88.5 ± 0.1 (88.3 - 88.6)
$\mathcal{T}_{\not\subset Internal}$	95.0 ± 0.4 (94.2 - 95.0)	95.1 ± 0.6 (93.8 - 96.1)	95.0 ± 0.1 (94.8 - 95.2)	89.9 ± 0.4 (89.3 - 90.7)	87.6 ± 0.5 (86.6 - 88.6)	88.8 ± 0.1 (88.5 - 89.0)
$\mathcal{T}_{\not\subset Otorhinolaryngeal}$	95.2 ± 0.2 (94.8 - 95.6)	96.2 ± 0.1 (96.0 - 96.3)	95.7 ± 0.1 (95.5 - 95.8)	89.7 ± 0.3 (89.1 - 90.3)	90.7 ± 0.6 (89.6 - 91.7)	90.1 ± 0.2 (89.8 - 90.5)
$\mathcal{T}_{\not\subset Dermatological}$	94.7 ± 0.5 (93.6 - 95.5)	93.7 ± 0.4 (93.1 - 94.5)	94.2 ± 0.1 (94.0 - 94.4)	89.5 ± 0.3 (89.0 - 90.2)	87.0 ± 0.3 (86.5 - 87.6)	88.2 ± 0.1 (88.1 - 88.3)
$\mathcal{T}_{\not\subset Gynecological}$	94.2 ± 0.2 (93.7 - 94.6)	93.5 ± 0.2 (93.1 - 93.8)	93.8 ± 0.1 (93.6 - 94.0)	89.6 ± 0.1 (89.4 - 89.8)	88.7 ± 0.2 (88.3 - 89.0)	89.1 ± 0.1 (88.9 - 89.4)
$\mathcal{T}_{\not\subset Cerebral}$	95.2 ± 0.2 (94.7 - 95.6)	94.0 ± 0.2 (93.5 - 94.5)	94.6 ± 0.1 (94.3 - 94.8)	89.1 ± 0.2 (88.7 - 89.5)	88.9 ± 0.2 (88.6 - 89.4)	89.0 ± 0.1 (88.8 - 89.2)
$\mathcal{T}_{\not\subset Ophthalmological}$	95.0 ± 0.1 (94.7 - 95.3)	95.6 ± 0.2 (95.2 - 96.0)	95.3 ± 0.1 (95.1 - 95.5)	88.8 ± 0.4 (88.0 - 89.6)	87.7 ± 0.5 (87.0 - 88.8)	88.2 ± 0.2 (87.8 - 88.5)
$\mathcal{T}$	95.9 ± 0.2 (95.5 - 96.2)	95.4 ± 0.3 (94.7 - 96.0)	95.6 ± 0.1 (95.4 - 95.8)	90.5 ± 0.5 (89.6 - 91.4)	90.4 ± 0.6 (89.3 - 91.7)	90.5 ± 0.1 (90.3 - 90.6)

Table 7: Sensitivity (%) and standard error for all anatomical locations of the bleeding classification dataset. * denotes a significant difference at the 0.05 level between models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$.

	Gastrointestinal	**Urogenital**	**Internal**	**Otorhinolaryngeal**	**Dermatological**	**Gynecological**	**Cerebral**	**Ophthalmological**
	ELECTRA							
$\mathcal{T}_{\not\subset Gastrointestinal}$	75.7± 2.4*	98.6± 0.3*	96.2± 0.3	96.2± 0.4*	94.3± 0.4*	95.9± 0.4*	97.3± 0.3*	99.1± 0.1*
$\mathcal{T}_{\not\subset Urogenital}$	97.1±0.3*	68.5± 2.2*	97.4± 0.1*	96.6± 0.2*	95.8± 0.6*	95.8± 0.5*	98.1± 0.1*	98.7± 0.3
$\mathcal{T}_{\not\subset Internal}$	96.6±0.2*	98.8± 0.2*	82.3± 2.0*	96.2± 0.3*	94.6± 0.6*	95.6± 1.0*	97.3± 0.3*	99.0± 0.2*
$\mathcal{T}_{\not\subset Otorhinolaryngeal}$	96.2±0.4	98.2± 0.2*	96.2± 0.3	92.6± 0.4*	94.6± 0.2*	95.1± 0.5*	97.4± 0.1*	99.1± 0.2*
$\mathcal{T}_{\not\subset Dermatological}$	96.6±0.4	98.2± 0.3*	96.2± 0.4*	95.8± 0.5	70.6± 1.4*	96.0± 0.7*	97.6± 0.2*	98.7± 0.2
$\mathcal{T}_{\not\subset Gynecological}$	97.0±0.2*	98.6± 0.3*	97.2± 0.3*	97.2± 0.2*	95.5± 0.4*	64.8± 0.8*	97.9± 0.4*	99.4± 0.1*
$\mathcal{T}_{\not\subset Cerebral}$	96.2±0.3	98.3± 0.3*	96.6± 0.4	99.1± 0.1*	94.4± 0.5*	95.8± 0.4*	75.4± 0.7*	99.1± 0.1*
$\mathcal{T}_{\not\subset Ophthalmological}$	96.6±0.2*	98.5± 0.2*	96.5± 0.2*	95.2± 0.3	93.8± 0.6	95.4± 0.6*	97.4± 0.3*	90.7± 0.5*
$\mathcal{T}$	95.4± 0.4	97.3± 0.4	95.6± 0.5	95.2± 0.2	92.4± 0.8	93.4± 0.2	95.3± 0.4	98.4± 0.3
	LSTM							
$\mathcal{T}_{\not\subset Gastrointestinal}$	63.6 ± 3.0*	94.6 ± 0.7*	85.8 ± 1.0*	96.1 ± 0.5*	88.2 ± 0.9	92.2 ± 0.8	90.6 ± 0.7*	96.6 ± 0.6*
$\mathcal{T}_{\not\subset Urogenital}$	90.2 ± 0.4	56.7 ± 1.6*	86.0 ± 0.5	97.4 ± 0.2*	89.8 ± 0.3*	91.6 ± 0.4	90.9 ± 0.3*	96.2 ± 0.1
$\mathcal{T}_{\not\subset Internal}$	86.9 ± 0.6	93.0 ± 0.7	62.6 ± 1.2*	95.1 ± 0.5	86.9 ± 0.5	89.7 ± 0.4	90.6 ± 0.5	96.1 ± 0.4
$\mathcal{T}_{\not\subset Otorhinolaryngeal}$	88.5 ± 0.6	93.8 ± 0.6	85.2 ± 1.0	89.8 ± 0.5*	89.4 ± 0.8	92.4 ± 0.7	90.2 ± 0.5*	96.1 ± 0.4*
$\mathcal{T}_{\not\subset Dermatological}$	90.6 ± 0.3	93.4 ± 0.5	85.4 ± 0.4*	96.0 ± 0.1*	53.4 ± 0.5*	91.8 ± 0.5	90.4 ± 0.6*	95.2 ± 0.2
$\mathcal{T}_{\not\subset Gynecological}$	89.3 ± 0.5	92.4 ± 0.3	86.4 ± 0.2*	96.1 ± 0.3	89.2 ± 0.4	69.1 ± 1.2*	90.8 ± 0.4*	96.2 ± 0.2
$\mathcal{T}_{\not\subset Cerebral}$	87.8 ± 0.6	93.7 ± 0.2	85.6 ± 0.6	96.8 ± 0.3*	89.0 ± 0.4	93.0 ± 0.2*	68.8 ± 0.5*	96.7 ± 0.2*
$\mathcal{T}_{\not\subset Ophthalmological}$	89.5 ± 1.0	93.1 ± 0.7	86.6 ± 0.9*	96.9 ± 0.4*	89.4 ± 1.0	93.6 ± 0.3*	90.9 ± 0.3*	61.5 ± 2.1*
$\mathcal{T}$	88.9 ± 1.1	93.0 ± 0.8	85.2 ± 0.8	95.4 ± 0.2	87.8 ± 0.8	90.9 ± 0.6	88.7 ± 0.3	94.6 ± 0.7

Table 8: Bootstrapped 95% confidence intervals for difference of means between models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$ of the bleeding classification dataset. Means are computed as performance of models trained on $\mathcal{T}_{\not\subset x}$ minus $\mathcal{T}$. Total = difference of means on the full test set.

	Total	**Gastrointestinal**	**Urogenital**	**Internal**	**Otorhinolaryngeal**	**Dermatological**	**Gynecological**	**Cerebral**	**Ophthalmological**
	ELECTRA								
$\mathcal{T}_{\not\subset Gastrointestinal}$	-1.5 , -0.9	-23.6 , -15.0	0.2 , 2.3	-0.6 , 2.0	0.5 , 1.5	0.6 , 3.3	1.6 , 3.3	1.4 , 2.6	0.5 , 1.1
$\mathcal{T}_{\not\subset Urogenital}$	-2.0 , -1.4	0.6 , 2.8	-33.2 , -24.7	1.0 , 2.7	0.9 , 2.1	1.8 , 5.0	1.2 , 3.6	1.8 , 3.8	-0.6 , 1.3
$\mathcal{T}_{\not\subset Internal}$	-0.8 , -0.4	0.2 , 2.4	0.7 , 2.4	-18.5 , -9.1	0.1 , 2.2	0.1 , 4.6	0.4 , 3.9	1.2 , 3.0	0.1 , 1.2
$\mathcal{T}_{\not\subset Otorhinolaryngeal}$	-0.3 , 0.4	-0.2 , 1.6	0.2 , 1.7	-0.5 , 1.9	-2.9 , -2.1	0.7 , 4.0	0.7 , 2.5	1.5 , 2.6	0.2 , 1.4
$\mathcal{T}_{\not\subset Dermatological}$	-1.8 , -1.1	0.0 , 2.4	0.4 , 1.7	0.1 , 1.3	-0.2 , 1.6	-23.5 , -19.8	1.8 , 3.6	1.9 , 2.9	-0.1 , 1.2
$\mathcal{T}_{\not\subset Gynecological}$	-2.2 , -1.5	1.1 , 2.3	0.7 , 2.1	0.8 , 2.2	1.3 , 2.7	1.6 , 4.9	-29.5 , -27.3	2.1 , 3.4	0.6 , 1.4
$\mathcal{T}_{\not\subset Cerebral}$	-1.5 , -0.5	-0.1 , 1.6	0.2 , 1.9	-0.2 , 2.3	0.5 , 1.6	0.4 , 3.5	1.7 , 3.3	-21.3 , -18.7	0.3 , 1.3
$\mathcal{T}_{\not\subset Ophthalmological}$	-0.6 , -0.1	0.5 , 2.3	0.5 , 2.1	0.1 , 1.7	-0.7 , 0.7	-0.1 , 3.7	1.0 , 3.4	1.5 , 2.9	-8.6 , -6.5
	LSTM								
$\mathcal{T}_{\not\subset Gastrointestinal}$	-1.9 , -1.4	-30.3 , -18.6	1.0 , 2.1	1.0 , 2.2	0.1 , 1.4	-0.6 , 1.4	-0.2 , 2.6	0.8 , 2.9	1.2 , 3.0
$\mathcal{T}_{\not\subset Urogenital}$	-2.1 , -1.8	-1.7 , 4.0	-38.6 , -34.0	-0.5 , 3.6	1.4 , 2.8	0.1 , 4.1	-1.3 , 2.6	1.2 , 3.1	-0.1 , 3.0
$\mathcal{T}_{\not\subset Internal}$	-1.8 , -1.4	-4.6 , 0.7	-1.4 , 1.7	-23.0 , -20.3	-1.4 , 1.1	-2.1 , 1.0	-2.6 , 0.2	-2.6 , 0.2	0.0 , 3.0
$\mathcal{T}_{\not\subset Otorhinolaryngeal}$	-0.7 , -0.1	-2.7 , 1.8	-0.8 , 2.1	-1.2 , 3.4	-6.9 , -4.5	-1.3 , 4.1	-0.6 , 3.0	0.1 , 2.6	0.2 , 2.7
$\mathcal{T}_{\not\subset Dermatological}$	-2.2 , -2.0	-1.1 , 4.2	-2.2 , 2.5	-1.3 , 3.0	0.2 , 1.0	-36.3 , -32.1	-0.2 , 2.0	0.4 , 3.4	-1.0 , 2.2
$\mathcal{T}_{\not\subset Gynecological}$	-1.5 , -1.1	-1.9 , 2.7	-2.0 , 0.4	0.3 , 3.8	-0.1 , 1.6	-0.4 , 3.7	-24.3 , -19.0	1.4 , 2.8	0.0 , 2.8
$\mathcal{T}_{\not\subset Cerebral}$	-1.7 , -1.2	-3.4 , 0.4	-1.0 , 1.7	-0.3 , 3.0	1.0 , 1.8	-0.6 , 3.4	1.4 , 2.7	-21.0 , -19.0	0.6 , 3.4
$\mathcal{T}_{\not\subset Ophthalmological}$	-2.6 , -1.8	-2.9 , 3.8	-1.1 , 1.4	0.1 , 4.8	0.7 , 2.4	-0.2 , 3.4	1.4 , 3.9	1.3 , 3.0	-38.5 , -28.5

Table 9: Bleeding test set sensitivity (%) and standard error on an anatomical location by percentage of subgroup samples in the modified training set. * denotes a significant difference at the 0.05 level between models trained on the modified training set and the full training set, $\mathcal{T}$.

	Anatomical subgroup fraction						
	0.0	**0.01**	**0.15**	**0.30**	**0.60**	**0.80**	**1.0**
	ELECTRA						
Gastrointestinal	75.7±2.4*	76.6± 0.6*	90.4± 1.1*	93.2± 0.3*	94.5± 0.5*	96.6± 0.5	95.8± 0.4
Urogenital	68.5±2.2*	72.4± 1.1*	91.0± 1.3*	96.6± 0.3*	97.1± 0.1*	97.0± 0.1*	98.4± 0.4
Internal	82.3±2.0*	84.3± 1.8*	91.8± 0.6*	93.6± 0.2*	94.8± 0.5*	95.9± 0.4*	96.4± 0.4
Otorhinolaryngeal	92.6±0.4*	92.2± 0.7*	92.9± 0.3*	94.8± 0.2*	95.4± 0.6	95.8± 0.2	95.6± 0.3
Dermatological	70.6±1.4*	74.6± 1.3*	87.9± 0.7*	88.7± 0.9*	93.3± 0.7	93.4± 0.3	94.5± 0.6
Gynecological	64.8±0.8*	63.4± 1.3*	87.3± 0.6*	89.2±0.3*	92.0± 0.7*	93.7± 0.2	93.9± 0.4
Cerebral	75.4±0.7*	80.0± 1.5*	89.4± 0.4*	91.5± 0.7*	95.6± 0.5*	96.6± 0.6	97.2± 0.3
Ophthalmological	90.70±0.5*	93.6± 0.3*	96.7± 0.4*	97.8± 0.3*	98.8± 0.1	98.7± 0.1	98.9± 0.2
	LSTM						
Gastrointestinal	63.6 ± 3.0*	60.0 ± 1.0*	71.9 ± 1.7*	79.2 ± 0.7*	86.6 ± 0.5*	88.0 ± 0.9*	89.1 ± 0.6
Urogenital	56.7 ± 1.6*	58.6 ± 0.9*	79.3 ± 0.7*	87.2 ± 0.6*	90.6 ± 0.2*	92.2 ± 0.3*	93.7 ± 0.1
Internal	62.6 ± 1.2*	65.6 ± 0.8*	72.6 ± 0.8*	77.2 ± 0.5*	80.2 ± 0.8*	83.9 ± 0.5	85.5 ± 0.5
Otorhinolaryngeal	89.8 ± 0.5*	88.5 ± 0.3*	91.8 ± 0.3*	93.6 ± 0.3*	93.8 ± 0.1*	95.2 ± 0.5	95.5 ± 0.3
Dermatological	53.4 ± 0.5*	56.5 ± 1.5*	73.5 ± 2.4*	77.4 ± 0.6*	84.5 ± 0.3*	86.8 ± 0.4	88.1 ± 0.5
Gynecological	69.1 ± 1.2*	69.3 ± 0.5*	80.1 ± 0.7*	86.6 ± 0.7*	89.3 ± 0.5*	90.7 ± 0.4*	91.6 ± 0.3
Cerebral	68.8 ± 0.5 *	72.2 ± 1.6*	79.7 ± 0.4*	83.1 ± 0.6*	86.4 ± 0.4*	88.4 ± 0.5*	90.5 ± 0.2
Ophthalmological	61.5 ± 2.1*	65.4 ± 1.4*	89.4 ± 0.4*	90.3 ± 0.5*	93.0 ± 0.7*	94.4 ± 0.4*	95.7 ± 0.4

Table 10: Bootstrapped 95% confidence intervals for difference of means between models trained on a modified training set, including a percentage of subgroup samples, and models trained on the full training set, $\mathcal{T}$, of the bleeding classification dataset. Means are computed as performance of models trained on the modified training set minus $\mathcal{T}$.

	Anatomical subgroup fraction					
	0.0	**0.01**	**0.15**	**0.30**	**0.60**	**0.80**
	ELECTRA					
Gastrointestinal	-23.6 , -15.0	-20.2 , -18.2	-8.3 , -2.6	-4.2 , -1.2	-2.6 , -0.2	-0.2 , 1.8
Urogenital	-33.2 , -24.7	-27.7 , -24.0	-10.0 , -4.8	-2.6 , -1.4	-2.2 , -0.4	-2.2 , -0.6
Internal	-18.5 , -9.1	-15.2 , -9.0	-5.8 , -3.4	-3.8 , -1.8	-2.4 , -1.0	-3.0 , -0.1
Otorhinolaryngeal	-2.9 , -2.1	-4.8 , -2.1	-3.5 , -1.8	-1.5 , -0.2	-0.7 , 0.3	-0.6 , 1.0
Dermatological	-23.5 , -19.8	-22.6 , -16.5	-7.8 , -5.0	-7.8 , -3.4	-3.0 , 1.0	-2.4 , 0.1
Gynecological	-29.5 , -27.3	-33.6 , -27.5	-8.1 , -5.3	-5.5 , -3.8	-3.6 , -0.8	-0.8 , 1.5
Cerebral	-21.3 , -18.7	-20.1 , -13.8	-8.4 , -7.3	-7.0 , -4.4	-2.6 , -0.7	-1.3 , 0.0
Ophthalmological	-8.6 , -6.5	-6.0 , -4.5	-2.7 , -1.8	-1.8 , -0.5	-0.6 , 0.6	-0.2 , 0.6
	LSTM					
Gastrointestinal	-31.7 , -19.8	-31.0 , -27.2	-21.4 , -14.3	-12.3 , -7.6	-4.1 , -0.6	-1.9 , -0.1
Urogenital	-39.8 , -33.8	-36.7 , -32.9	-15.8 , -13.0	-7.8 , -4.8	-3.5 , -2.5	-2.1 , -0.9
Internal	-25.7 , -19.6	-22.3 , -17.4	-15.0 , -10.6	-9.5 , -7.1	-7.4 , -3.1	-3.4 , 0.4
Otorhinolaryngeal	-7.0 , -4.5	-8.2 , -6.2	-4.5 , -3.0	-2.5 , -1.4	-2.3 , -1.1	-1.8 , 1.0
Dermatological	-35.6 , -33.8	-35.5 , -28.6	-19.0 , -10.1	-12.8 , -9.0	-4.9 , -2.5	-3.1 , 0.0
Gynecological	-25.4 , -19.5	-23.3 , -21.2	-13.4 , -10.3	-6.5 , -3.4	-3.9 , -0.9	-1.2 , -0.6
Cerebral	-22.3 , -21.2	-21.1 , -15.1	-11.8 , -9.8	-8.7 , -5.8	-5.0 , -3.2	-3.4 , -0.8
Ophthalmological	-38.4 , -30.5	-33.4 , -27.2	-7.4 , -5.1	-6.5 , -4.1	-3.9 , -1.1	-2.1 , -0.6

Table 11: Precision, recall, and F1 performance for the VTE classification dataset. $\mathcal{T}_{\not\subset x}$ denotes the training set from which an anatomical location, x, has been removed. SE = Standard error. CI = 95% confidence interval.

	ELECTRA			**LSTM**		
	Precision ± SE (CI)	**Recall ± SE (CI)**	**F1 ± SE (CI)**	**Precision ± SE (CI)**	**Recall ± SE (CI)**	**F1 ± SE (CI)**
$\mathcal{T}_{\not\subset Lower\ extremity}$	86.3 ± 0.7 (84.9 - 87.7)	41.8 ± 1.5 (39.1 - 44.7)	56.3 ± 1.3 (53.8 - 58.8)	77.2 ± 0.2 (76.9 - 77.6)	61.4 ± 1.5 (58.8 - 64.5)	68.3 ± 0.8 (66.8 - 70.1)
$\mathcal{T}_{\not\subset Lung}$	86.1 ± 1.0 (84.1 - 87.9)	57.4 ± 3.2 (51.8 - 63.8)	68.6 ± 2.0 (64.9 - 72.7)	78.6 ± 0.5 (77.6 - 79.6)	54.8 ± 0.7 (53.5 - 56.1)	64.6 ± 0.3 (63.9 - 65.3)
$\mathcal{T}$	96.4 ± 0.3 (95.9 - 97.0)	72.3 ± 0.7 (71.0 - 73.6)	82.7 ± 0.4 (81.9 - 83.5)	91.4 ± 0.1 (91.2 - 91.6)	57.2 ± 0.6 (56.0 - 58.4)	70.4 ± 0.4 (69.4 - 71.2)

Table 12: Sensitivity (%) and standard error for all anatomical locations of the VTE classification dataset. * denotes a significant difference at the 0.05 level between models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$.

	Lower extremity	**Lung**	**Liver**	**Cerebral**	**Upper extremity**
			ELECTRA		
$\mathcal{T}_{\not\subset Lower\ extremity}$	9.3 ± 0.8*	98.6 ± 0.2*	50.1 ± 3.6*	23.8 ± 1.9*	25.3 ± 1.4*
$\mathcal{T}_{\not\subset Lung}$	99.7 ± 0.2*	16.6 ± 3.5*	65.7 ± 7.7	13.6 ± 4.1	99.0 ± 0.3*
$\mathcal{T}$	99.1 ± 0.2	97.6 ± 0.3	66.3 ± 1.6	11.5 ± 1.6	96.7 ± 0.3
			LSTM		
$\mathcal{T}_{\not\subset Lower\ extremity}$	34.2 ± 1.5*	92.6 ± 0.5*	76.2 ± 2.2*	52.6 ± 2.2*	47.6 ± 2.8*
$\mathcal{T}_{\not\subset Lung}$	95.3 ± 0.2*	15.7 ± 0.4*	54.2 ± 1.8*	24.7 ± 1.3*	91.4 ± 0.4*
$\mathcal{T}$	93.8 ± 0.6	86.6 ± 0.6	29.8 ± 1.6	7.0 ± 0.5	81.8 ± 1.3

Table 13: Bootstrapped 95% confidence intervals for difference of means between models trained on $\mathcal{T}_{\not\subset x}$ and $\mathcal{T}$ of the VTE classification dataset. Means are computed as performance of models trained on $\mathcal{T}_{\not\subset x}$ minus $\mathcal{T}$. Total = difference of means on the full test set.

	Total	**Lower extremity**	**Lung**	**Liver**	**Cerebral**	**Upper extremity**
				ELECTRA		
$\mathcal{T}_{\not\subset Lower\ extremity}$	-18.0 , -16.3	-91.3 , -88.3	0.2 , 1.6	-23.9 , -7.3	8.0 , 16.6	-74.0 , -68.7
$\mathcal{T}_{\not\subset Lung}$	-12.4, -9.2	0.3 , 0.9	-87.2 , -73.7	-12.6 , 11.6	-2.0 , 8.3	1.8 , 2.7
				LSTM		
$\mathcal{T}_{\not\subset Lower\ extremity}$	-5.1 , -3.5	-63.0 , -56.5	4.3 , 7.7	43.8 , 50.0	40.3 , 49.6	-41.9 , -28.2
$\mathcal{T}_{\not\subset Lung}$	-6.6 , -5.4	0.6 , 2.6	-72.5 , -69.5	18.9 , 28.8	14.9 , 19.7	7.5 , 11.9

Table 14: VTE test set sensitivity (%) and standard error on an anatomical location by percentage of subgroup samples in the modified training set. * denotes a significant difference at the 0.05 level between models trained on the modified training set and the full training set, $\mathcal{T}$.

				Anatomical subgroup fraction			
	0.0	**0.01**	**0.15**	**0.30**	**0.60**	**0.80**	**1.0**
				ELECTRA			
Lower extremity	9.3 ± 0.8*	77.6 ± 3.7*	97.2 ± 0.3*	97.7 ± 0.4	98.6 ± 0.3	98.8 ± 0.2	98.4 ± 0.2
Lung	16.6 ± 3.5*	57.1 ± 8.1*	94.6 ± 0.3*	96.6 ± 0.3	97.7 ± 0.2	98.2 ± 0.1	97.6 ± 0.3
				LSTM			
Lower extremity	34.2 ± 1.5	40.6 ± 0.7	83.7 ± 0.6	90.7 ± 0.5	94.4 ± 0.6	95.5 ± 0.6	94.9 ± 0.7
Lung	15.7 ± 0.4	16.0 ± 0.9	71.5 ± 1.1	81.6 ± 0.7	88.0 ± 0.9	88.3 ± 0.9	89.2 ± 0.8

Table 15: Bootstrapped 95% confidence intervals for difference of means between models trained on a modified training set, including a percentage of subgroup samples, and models trained on the full training set, $\mathcal{T}$, of the VTE classification dataset. Means are computed as performance of models trained on the modified training set minus $\mathcal{T}$.

			Anatomical subgroup fraction			
	0.0	**0.01**	**0.15**	**0.30**	**0.60**	**0.80**
			ELECTRA			
Lower extremity	-91.3 , -88.3	-28.8 , -15.0	-2.2 , -0.4	-1.9 , 0.4	-0.2 , 0.7	-0.4 , 1.1
Lung	-87.2 , -73.7	-58.8 , -27.4	-3.6 , -2.2	-2.0 , 0.0	-0.7 , 0.8	-0.1 , 1.2
			LSTM			
Lower extremity	-63.0 , -56.5	-56.2 , -52.2	-13.2 , -9.4	-5.6 , -2.9	-2.3 , 1.2	-1.3 , 2.2
Lung	-72.5 , -69.5	-76.3 , -70.1	-20.6 , -14.4	-9.2 , -6.0	-3.5 , 0.8	-3.6 , 1.9

Table 16: Most frequent words used to describe bleeding mentions for each anatomical location of the bleeding classification dataset. Words are translated from Danish to English and, therefore, some cells include two words.

Word	**Frequency**	**Location uniqueness**	**Word**	**Frequency**	**Location uniqueness**
Location: Otorhinolaryngeal			Location: Gynecological		
bleeding	324	0.13	bleeding	714	0.29
nose bleeding	273	1.0	uterus	108	0.97
epistaxis	254	0.99	allowable	78	0.94
nostril	148	1.0	vagina	69	1.0
Location: Dermatological			Location: Cerebral		
haematoma	354	0.56	sah	217	0.99
bleeding	170	0.07	bleeding	190	0.08
skin	122	0.73	ct	185	0.63
right	97	0.29	haematoma	161	0.25
Location: Urogenital			Location: Internal		
haematuria	536	0.99	bleeding	273	0.11
urine	311	0.98	haemothorax	249	1.0
blood	205	0.23	fluid	174	0.80
macroscopic	186	0.99	blood	140	0.16
Location: Gastrointestinal			Location: Ophthalmological		
bleeding	560	0.23	corpus hemorrhagicum	270	1.0
blood	247	0.28	corpus hem	205	1.0
fresh	180	0.48	bleeding	198	0.08
melaena	165	0.98	haemorrhage	180	0.70

.