Evaluating large language models for predicting psychiatric acute readmissions from clinical notes of population-based EHR

Terne Sasha Thorn Jakobsen[1,2], Enric Cristòbal Cóppulo[1,2,3], Simon Rasmussen[2,#], Michael Eriksen Benros[1,3,#]

Affiliations

[1] Copenhagen Research Centre for Mental Health, Mental Health Centre Copenhagen, Copenhagen University Hospital, Denmark

[2] Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen N, Denmark

[3] Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

#Corresponding authors

Michael E. Benros, Copenhagen Research Centre for Biological and Precision Psychiatry, Mental Health Centre Copenhagen, Copenhagen University Hospital, Gentofte Hospitalsvej 19, ground floor, 2900 Hellerup, Denmark, email: michael.eriksen.benros@regionh.dk

Simon Rasmussen, Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Blegdamsvej 3A, 2200 Copenhagen N, Denmark, email: srasmuss@sund.ku.dk

ABSTRACT

Psychiatric patients often have complex symptoms and anamneses recorded as unstructured clinical notes. Large language models (LLM) now enable large-scale utilization of text data; however, there is a current lack of LLMs specialized for psychiatric clinical data, as well as non-English data, haltering the application of LLMs across diverse clinical domains and countries. We present PsyRoBERTa: the first LLM specialized for clinical psychiatry, using population-based data with the currently largest collection of clinical notes of psychiatric relevancy (~44 million notes) covering the eastern half of Denmark. The model was evaluated against three publicly available models, pretrained on either public general- or medical-domain text, and a baseline logistic regression classifier. Through extensive evaluations, we investigated the effect of domain-specific pretraining on predicting acute readmissions in psychiatric hospitals, explored important features, and reflected on (dis)advantages of LLMs. PsyRoBERTa succeeded in outperforming prior models (AUC=0.74), capturing information aligning with clinical practice, and additionally recognizing psychiatric diagnoses (AUC=0.85). This demonstrates the importance of domain-pretraining and the potential of LLMs to leverage psychiatric clinical notes for enhancing prediction of psychiatric outcomes.

INTRODUCTION

Large language models (LLM) now enable large-scale utilization of the complex symptom descriptions and anamneses present within clinical notes from Electronic Health Records (EHR). However, at a large scale, clinical notes remain largely unutilized for psychiatric research and outcome prediction, with most efforts utilizing structured data from clinical studies, registries and EHR, as well as genetic data.[1–3] This is partly explained by the fact that applying LLMs to clinical psychiatry data can be particularly challenging: Clinical text differs considerably, in vocabulary and structure, from the internet-originating text most general-purpose LLMs are based on, particularly regarding the many clinical abbreviations common in clinical EHR text. Moreover, LLMs are essentially large neural networks that have been trained to predict upcoming or missing words in a sequence of words, usually based on large corpora of internet-originating text. To apply an LLM for other tasks, such as predicting psychiatric outcomes from clinical notes, the LLM should also be *finetuned* specifically for this, which entails adding new learnable weights for the task, and backpropagating through the entire model to update the learned word representations (embeddings) according to the new purpose. Although a main advantage of LLMs is that they can be applied in such a *transfer learning* fashion, meaning they only need relatively brief finetuning for new tasks, because they have already learned good, general representations of words, there is substantial evidence suggesting LLMs should be further pretrained for the target domain (i.e. clinical psychiatry), before finetuning, to achieve best performance and parameter-efficiency.[4–9]

Several LLMs have been pretrained for the clinical domain with English EHR notes.[10–20] A recent review[21] identified 23 LLMs pretrained specifically on clinical text and found 17 of them to be trained on the same source data (MIMIC-III/IV),[22,23] highlighting the challenge of attaining high-quality clinical text datasets and a current lack of diversity among available models. Furthermore, data in English is by far dominating the rapid development of clinical LLMs while development for other languages is slowly progressing.[4,5,24,25] **Figure 1** gives an extended overview of LLMs pretrained on clinical notes[5,6,10–20,26–36] and commonly evaluated against their base model, e.g. BERT,[37] and models pretrained on biomedical research literature, e.g. BioBERT.[38] The availability of LLMs for investigating mental health issues is currently much more limited,[39] with most efforts focused on exploring mental health conditions in online fora using models such as MentalBERT[40] trained on text from social media. So far, utilizing clinical notes for psychiatric outcome prediction has mostly built on other natural language processing (NLP) methods for information extraction, topic modelling, and sentiment

classification for tasks such as predicting the use of mechanical restraint,[41] psychosis symptom onset,[42,43] and readmission.[44–47] Recently, Perfalk et al (2024)[48] evaluated machine learning (ML) models for predicting involuntary psychiatric admissions in the Central Denmark Region, such as a model trained with Bag-of-Words features (text represented by frequencies, disregarding order and context) from clinical notes. They highlighted the potential of a transformer-based model finetuned specifically for psychiatric clinical notes as a future direction more suitable for the sequential text data.

Figure 1. Overview of prior language models pretrained on clinical notes

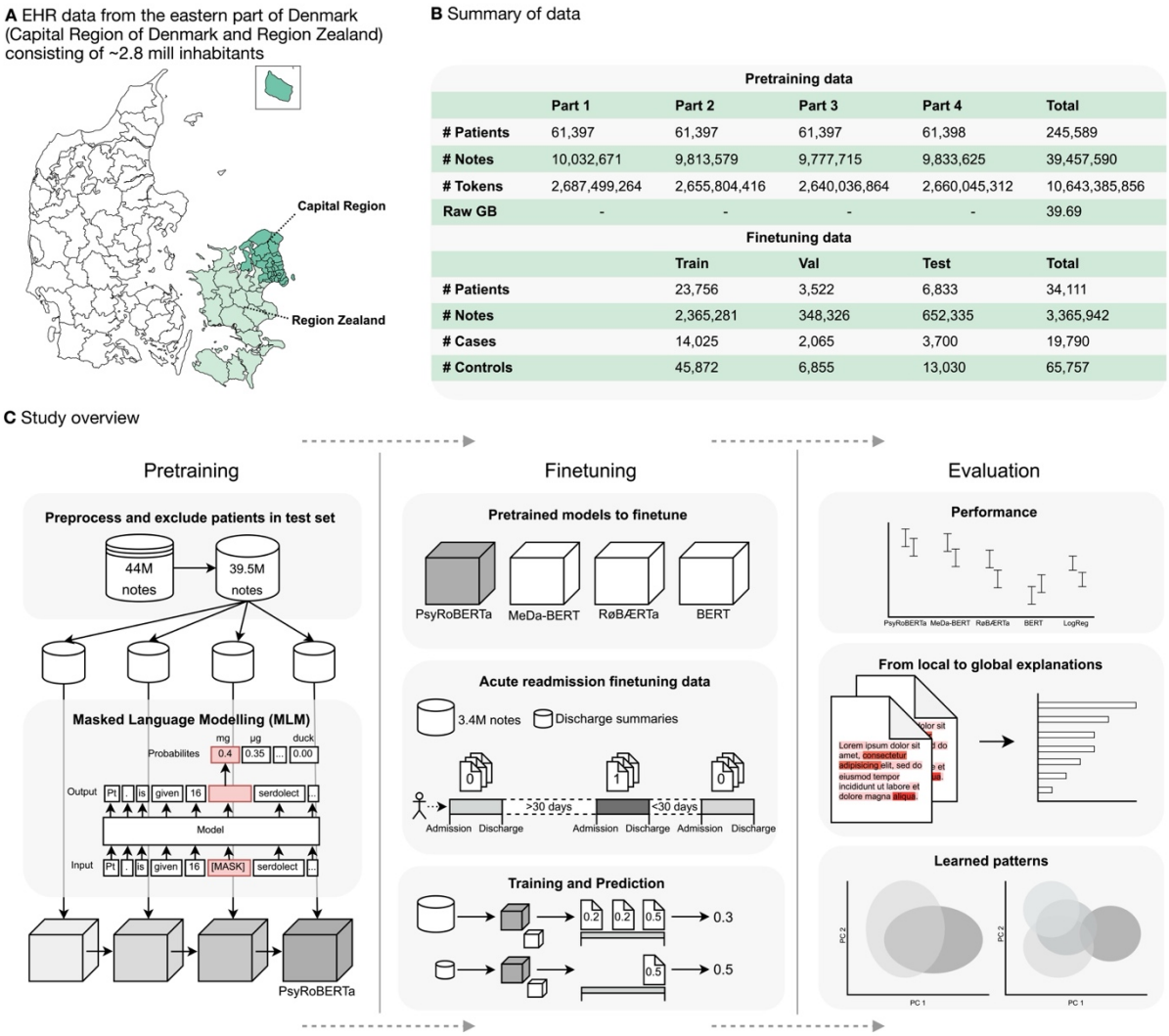
	Bio+ClinicalBERT[15]	BlueBERT[17]	Clinical Llama-LoRA[20]	Clinical-kb-bert[26]	Clinical-Longformer[14]	Clinical-T5 base (Lehman)[6]	Clinical-T5 large (Lehman)[6]	Clinical-T5 base (Lu)[27]	Clinical-T5 large (Lu)[27]	ClinicalBERT[10]	Clinical BERT embeddings[15]	ClinicalDistilBERT[18]	Clinical XLNet[28]	CORe[13]	CXR-BERT[29]	EntityBERT[30]	GatorTron small[12]	GatorTron medium[12]	GatorTron large[12]	Medical Domain BERT[31]	Medical Domain CharacterBERT[31]	medRoberta[19]	NYUTron[11]	RadBERT[32]	UCSFBert[33]	UmlsBERT[16]	Bert-for-Radiology[34]	Biomedical-clinical language model for Spanish[35]	Clinical ELECTRA[5]	MedRobBERTa.nl[36]	PsyRoBERTa
Language	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	EN	DE	ES	DA	NL	DA
Parameters (M)	110	110	7000	110	149	220	770	220	770	110	110	105	110	110	110	110	340	3900	8900	110	105	125	110	110	110	110	110	125	13	125	125
Tokens (B)	0.8	4.5	-	-	-	2	2	-	-	-	-	-	-	-	-	0.7	90	90	90	1.2	1.2	-	4.1	0.4	39	-	0.4	0.09	1.4	-	10.6
Notes (M)	2	2	2.6	2	2	4.6	4.6	2	2	2	2	2	0.07	0.04	0.15	-	292	292	292	2	2	2	7.2	4.4	75	2	3.8	0.5	0.3	10	39.5
Patients (k)	46	46	380	46	46	380	380	46	46	46	46	46	7	-	65	46	2500	2500	2500	46	46	46	380	2100	2200	46	-	-	-	-	246
Corpus source																															
MIMIC-III/IV/CXR	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x				x					
UMLS				x																						x					
MTSamples and i2b2														x																	
Local EHR																	x	x	x				x	x	x		x	x	x	x	x
Mixed w. non-EHR*	x	x												x	x		x	x	x	x	x										

Gradients of the colors signify volume, with darker colors for larger values within the property.

Here, we present the first LLM specialized for clinical psychiatry using population-based data and the currently largest collection of clinical EHR notes with psychiatric relevancy. Our model, PsyRoBERTa, is pretrained on ~40 million Danish clinical notes from 245,589 patients who have had contact with the mental health services. Clinical ML studies often only report a few performance metrics and no explainability or bias analyses. In this study, we conduct extensive evaluations when comparing our model to a logistic regression (LR) classifier (using a Bag-of-Words (BoW) representation of the frequency of unigram features) and three transformer-based LLMs pretrained on other sources and finetuned on our clinical notes. Pretraining LLMs is expensive and time-consuming, but it is a one-time expense that has the potential to enhance performance or reduce the amount of data needed for finetuning. With the task of predicting psychiatric acute readmissions as a case, we aim to answer the questions: 1) Does domain-specific pretraining enhance performance for predicting acute readmissions in psychiatry, as expected given previous research? 2) What features within clinical notes does an

LLM find important for psychiatric acute readmissions and do these align with clinical practice? 3) What are the (dis)advantages of an LLM versus a simpler ML model?

Figure 2. Graphical abstract with data description and study overview



Pretraining data					
	Part 1	Part 2	Part 3	Part 4	Total
# Patients	61,397	61,397	61,397	61,398	245,589
# Notes	10,032,671	9,813,579	9,777,715	9,833,625	39,457,590
# Tokens	2,687,499,264	2,655,804,416	2,640,036,864	2,660,045,312	10,643,385,856
Raw GB	-	-	-	-	39.69

Finetuning data				
	Train	Val	Test	Total
# Patients	23,756	3,522	6,833	34,111
# Notes	2,365,281	348,326	652,335	3,365,942
# Cases	14,025	2,065	3,700	19,790
# Controls	45,872	6,855	13,030	65,757

*A. Origin of EHR data. **B.** Size of pretraining and finetuning datasets. Raw GB refers to the size of a text file containing only the notes. Tokens were counted after tokenization with the RoBERTa tokenizer (byte-level BPE). Cases and controls refer to the admissions that lead or did not lead to acute readmissions. **C.** Study overview. This study consists of three steps: pretraining, finetuning, and evaluation. **Pretraining**: the clinical notes were preprocessed and patients within the test set of the finetuning data were excluded from pretraining. The remaining 39.5 million notes were split in four equally sized parts based on the number of patients with random selection. Starting with the RøBÆRTa model for MLM, we continued MLM training with each of the four data splits for 3 epochs, resulting in four models trained on incrementally more clinical notes. **Finetuning:** we finetuned four language models for predicting psychiatric acute readmissions; our model, PsyRoBERTa, and three publicly available models. Finetuning was carried out with two datasets, either containing all notes related to an admission or containing only the discharge summaries. An acute readmission was defined as an unplanned admission within 30 days after discharge. The models were trained to make a prediction for each textual input, with binary labels indicating whether the next admission was acute (1) or not acute (0), and we took the mean over notes within each admission as the final predictions.*

Evaluation: *We compared the predictive performance of each finetuned language model and a logistic regression classifier. We then retrieved explanations for PsyRoBERTa's predictions and visualized embeddings to inspect learned patterns.*

RESULTS

An extensive and transparent evaluation framework

From Danish population-based electronic health records, we processed 43,996,617 clinical notes from 255,944 patients with at least one hospital contact with the mental health services between 2000 and 2022 (**Figure 2A**). Data preprocessing included: de-identification of patients and health professionals; splitting data based on randomized patient IDs to ensure independence of patients across train-validation-test-splits; de-duplication of content in clinical notes. A pretraining dataset and finetuning dataset was created, with patients within the finetuning test set excluded from the pretraining data (**Figure 2B**). The study then consisted of 1) pretraining PsyRoBERTa, 2) finetuning models, and 3) evaluation (**Figure 2C**). We pretrained PsyRoBERTa in four stages to observe the effect of continuous pretraining with more or lesser data. We then finetuned PsyRoBERTa and three publicly available Danish LLMs (MeDa-BERT,[49] RøBÆRTa,[50] and Danish BERT[51], the latter hereon referred to simply as BERT) for predicting psychiatric acute readmissions and experimented with two datasets for finetuning, using either all notes within an admission (~3.4M) or using only the discharge summaries (~85k) to evaluate the difference between using the compressed information, i.e. discharge summaries, versus all notes. We reported performance of the LLMs and an LR baseline on eight metrics, with non-parametric bootstrapping confidence intervals, for a comprehensive representation of each model's strengths and weaknesses. Due to class imbalance (23% cases), we emphasized Matthew's Correlation Coefficient (MCC) as it is considered one of the most reliable metrics for imbalanced binary data.[52] We further evaluated fairness across sex, age, region, and psychiatric diagnoses. Lastly, we explained learned patterns of PsyRoBERTa, by inspecting attention scores using the Attention Rollout method,[53] and we explored the usefulness of its embeddings for psychiatric diagnosis recognition.

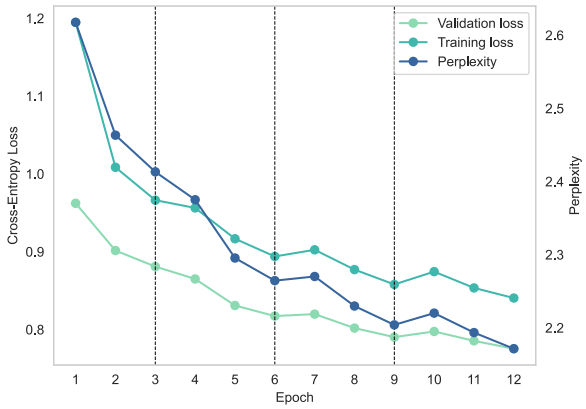
Domain-specific pretraining improved prediction of psychiatric acute readmissions

Our first question was whether domain-specific pretraining improves down-stream performance in our case as expected given the background literature. Overall, our model PsyRoBERTa consistently outperformed the non-clinical models on four metrics: Area Under the ROC curve (AUC), MCC, Area Under the Precision-Recall Curve (AUPRC), and weighted average F_1 (**Figure 3C**). The second-best model was MeDa-BERT, which was pretrained on

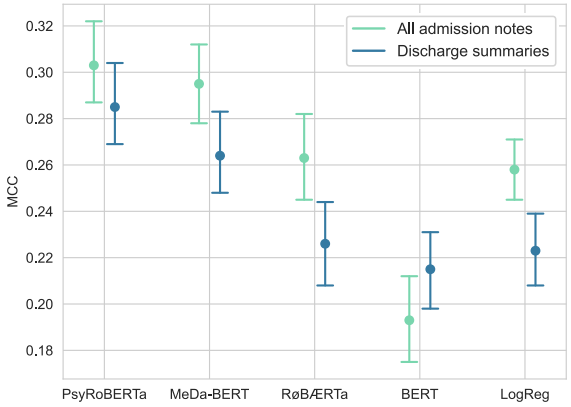
public clinical guidelines and medical websites,[47] indicating that these data sources were also highly relevant. Comparing the performance of finetuning the models with discharge summaries versus all notes revealed that the latter gave the best performance and PsyRoBERTa was the best model in both scenarios. When finetuning with discharge summaries, PsyRoBERTa (MCC=0.285) had significantly better[48] performance than all other models on the MCC metric: MeDa-BERT (MCC=0.264, $p<0.01$), RøBÆRTa (MCC=0.226, $p<0.0001$), BERT (MCC=0.215, $p<0.0001$) and LR (MCC=0.223, $p<0.0001$), and showed a 28% improvement from the LR baseline. When finetuning with all admission notes PsyRoBERTa (MCC=0.303) was still significantly better than RøBÆRTa (MCC=0.263, $p<0.0001$), BERT (MCC=0.193, $p<0.0001$) and LR (MCC=0.258, $p<0.0001$), and still better than MeDa-BERT (MCC=0.295), but by a smaller margin than previously, and showed a 17% improvement from the LR baseline. PsyRoBERTa, RøBÆRTa and MeDa-BERT were all excellent at correctly recognizing instances that did not lead to acute re-admissions (specificity: 91.9%, 94.8%, 93.1%, respectively) while recognizing positive cases was more challenging. However, PsyRoBERTa had a higher recall than the other language models, retrieving 33.2% of positive cases versus 30%, 23.7%, and 12.9% by MeDa-BERT, RøBÆRTa and BERT, respectively. We inspected the models' predicted probabilities relative to the observed probabilities, i.e. true risk (also known as *calibration curves*, see **Figure S1**) and found that all LLMs tended to overestimate the risk of readmission but exhibited better calibration when finetuned on all notes. Comparing the performance of PsyRoBERTa and MeDa-BERT with the models they were continued pretrained from (RøBÆRTa and BERT) (**Figure 3B-C**) showed that the continued pretraining resulted in a 26% improvement in MCC of RøBÆRTa and a 23% improvement in MCC of BERT when finetuning with discharge summaries. When finetuning with all admission notes, we found a 15% improvement of RøBÆRTa and a 53% improvement of BERT (the improvement should be seen in the perspective that BERT performed worst overall). Furthermore, the difference in performance of finetuning with discharge summaries versus all notes indicated that there is additional relevant information beyond the summaries, but the improvement in performance was modest and potentially hampered by increased amount of irrelevant (noisy) information. One way to reduce noise in text data is to remove repetitions (duplication), and we found that our deduplication method, applied during preprocessing of the data, improved the performance of PsyRoBERTa by 2.4% while reducing training time by ~3 hours.

Figure 3. Model performance

A Masked Language Modelling performance of PsyRoBERTa



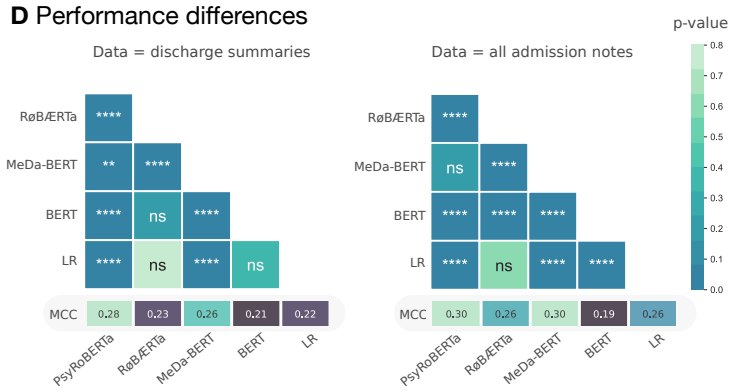
B Predictive performance of all models by MCC



C All performance results of models finetuned for predicting psychiatric acute readmission

Model	AUROC	AUPRC	MCC	F1 AVG	F1	Precision	Recall	Specificity
	Data = discharge summaries							
LogReg	0.680 [0.670,0.691]	0.390 [0.374,0.407]	0.223 [0.208,0.239]	0.691 [0.684,0.697]	0.425 [0.414,0.438]	0.341 [0.328,0.353]	0.564 [0.548,0.580]	0.695 [0.686,0.702]
BERT	0.679 [0.669,0.689]	0.368 [0.353,0.383]	0.215 [0.198,0.231]	0.742 [0.735,0.750]	0.354 [0.339,0.368]	0.425 [0.406,0.444]	0.303 [0.288,0.318]	0.885 [0.879,0.891]
MeDa-BERT	0.700 [0.691,0.711]	0.399 [0.385,0.416]	0.264 [0.248,0.283]	0.745 [0.739,0.753]	0.430 [0.417,0.446]	0.417 [0.401,0.433]	0.445 [0.429,0.463]	0.826 [0.818,0.832]
RøBÆRTa	0.689 [0.679,0.698]	0.378 [0.362,0.394]	0.226 [0.208,0.244]	0.747 [0.740,0.755]	0.351 [0.336,0.368]	0.450 [0.428,0.470]	0.288 [0.273,0.303]	0.901 [0.896,0.906]
PsyRoBERTa	0.711 [0.702,0.721]	0.420 [0.403,0.437]	0.285 [0.269,0.304]	0.759 [0.752,0.766]	0.435 [0.422,0.451]	0.452 [0.435,0.469]	0.419 [0.403,0.437]	0.858 [0.852,0.864]
	Data = all admission notes (deduplicated)							
LogReg	0.715 [0.706,0.724]	0.418 [0.400,0.433]	0.249 [0.236,0.262]	0.604 [0.598,0.612]	0.448 [0.438,0.459]	0.313 [0.304,0.323]	0.789 [0.777,0.802]	0.509 [0.500,0.518]
LogReg Not dedup.	0.718 [0.709,0.727]	0.422 [0.405,0.438]	0.258 [0.245,0.271]	0.622 [0.616,0.629]	0.454 [0.443,0.465]	0.322 [0.312,0.331]	0.774 [0.761,0.787]	0.536 [0.528,0.545]
BERT	0.718 [0.708,0.727]	0.420 [0.402,0.436]	0.193 [0.175,0.212]	0.729 [0.721,0.738]	0.211 [0.195,0.226]	0.569 [0.534,0.604]	0.129 [0.119,0.140]	0.972 [0.969,0.975]
MeDa-BERT	0.734 [0.725,0.743]	0.458 [0.440,0.474]	0.295 [0.278,0.312]	0.767 [0.760,0.774]	0.389 [0.374,0.405]	0.553 [0.529,0.573]	0.300 [0.286,0.315]	0.931 [0.926,0.935]
RøBÆRTa	0.728 [0.719,0.737]	0.447 [0.430,0.464]	0.263 [0.245,0.282]	0.756 [0.748,0.763]	0.333 [0.317,0.349]	0.562 [0.536,0.586]	0.237 [0.223,0.250]	0.948 [0.944,0.951]
PsyRoBERTa	0.736 [0.727,0.745]	0.464 [0.446,0.482]	0.303 [0.287,0.322]	0.770 [0.763,0.777]	0.410 [0.396,0.427]	0.538 [0.518,0.559]	0.332 [0.318,0.348]	0.919 [0.914,0.924]
PsyRoBERTa Not dedup.	0.732 [0.723,0.741]	0.463 [0.445,0.480]	0.296 [0.279,0.314]	0.767 [0.760,0.774]	0.388 [0.373,0.404]	0.557 [0.536,0.579]	0.298 [0.284,0.313]	0.933 [0.929,0.937]

D Performance differences



A. *Performance of the masked language modelling objective in pretraining. The striped line indicates the first epoch of the next model in our four-stage continuous pretraining.* ***B.*** *MCC of PsyRoBERTa, MeDa-BERT, RøBÆRTa, BERT and logistic regression. For each model on the x-axis, the left-most bar is the result from fine-tuning on all admission notes and the right-most is on discharge summaries.* ***C****. Performance of models fine-tuned for predicting psychiatric acute re-admissions. Top: results of finetuning with discharge summaries. Bottom: results of fine-tuning with all admission notes that are deduplicated, except the first and last row since it is not beneficial to deduplicate data for the logistic regression classifier (first row), and we show results of PsyRoBERTa when not deduplicating for comparison (last row). Green and blue highlights indicate best and worst score, respectively. 95% Confidence intervals are from non-parametric bootstrapping with 1000 resamples.* ***D.*** *Statistical significance for the observed difference in model performance by MCC. **p<0.01, ***p<0.001, ****p<0.0001.*

Fairness analysis indicates performance consistency across diagnoses and patient demographics

PsyRoBERTa consistently outperformed all models on all demographic subgroups (**Figure S2**) as well as most diagnosis subgroups apart from 'bipolar/manic' and 'other psychosis' (**Figure S3**). Cross-group performance disparity was generally small for all models (**Figure S4**), but largest across the diagnosis subgroups since these had substantial differences in frequency and within-group class proportion (**Figure S5**). The difference in performance across diagnosis subgroups, measured as the difference between the largest and smallest AUC across groups, was smallest for PsyRoBERTa (0.111) and largest for MeDa-BERT (0.139), indicating greater fairness by performance parity of PsyRoBERTa. PsyRoBERTa performed best for patients with schizophrenia (AUC=0.750) or personality disorders (AUC=0.771) and worst for patients with anxiety (AUC=0.662) or depression (AUC=0.660) – the diagnoses with lowest proportions of acute readmissions. PsyRoBERTa (all admission notes) had good performance parity for all sex-, region- and age-groups (**Figure S6**), except for patients above 54 years of age (AUC=0.669). Similar results were observed for PsyRoBERTa finetuned on discharge summaries (**Figure S7-8**).

Data size and training efficiency

We investigated performance improvements given increasing amount of pretraining data. We pretrained PsyRoBERTa on 39.5 million Danish clinical notes split in four parts to consecutively train the model for masked language modelling (MLM) for three epochs on each of the four parts of the pretraining data (i.e. three epochs with data part 1, then three epochs with data part 2 and so on, as visualized in **Figure 2C**). The MLM performance improved most in the first epochs, with an 8.4% improvement in validation loss from the first to the third epoch, and gradually diminishing gains throughout the pre-training process to 2.8% with the last part of the data (**Figure 3A**). During the validation of the four models on predicting acute

readmissions from discharge summaries, the last model improved slightly faster (**Figure S9**). Consequently, using all available data for pretraining resulted in lowest validation losses but the gain of increasing training data was diminishing for each new part. After finetuning, we observed that the difference in predictive performance given discharge summaries versus all notes was smaller for PsyRoBERTa than the other models, with an absolute difference in MCC of 0.018 versus 0.031, 0.037, 0.022, and 0.035 for MeDa-BERT, RøBÆRTa, BERT and LR, respectively (**Figure 3B**). Furthermore, the difference in performance of PsyRoBERTa and MeDa-BERT was significant when training with discharge summaries but not when training with all admission notes. This could indicate that PsyRoBERTa needed less finetuning data for performance increases. To explore this further, we re-finetuned the three LLMs on 10-50% of the discharge summary dataset (**Figure S10**) and found that both PsyRoBERTa and MeDa-BERT had better onsets in predictive performance, reflecting more efficient learning of both domain-pretrained models. The difference between the onsets (i.e. whether PsyRoBERTa or MeDa-BERT had a better onset in predictive performance with less training data than the other), varied across the 10-50% of training data usage, and it was therefore not clear which, if any, was better in terms of needing less data for finetuning (see **Figure S10** for more details).

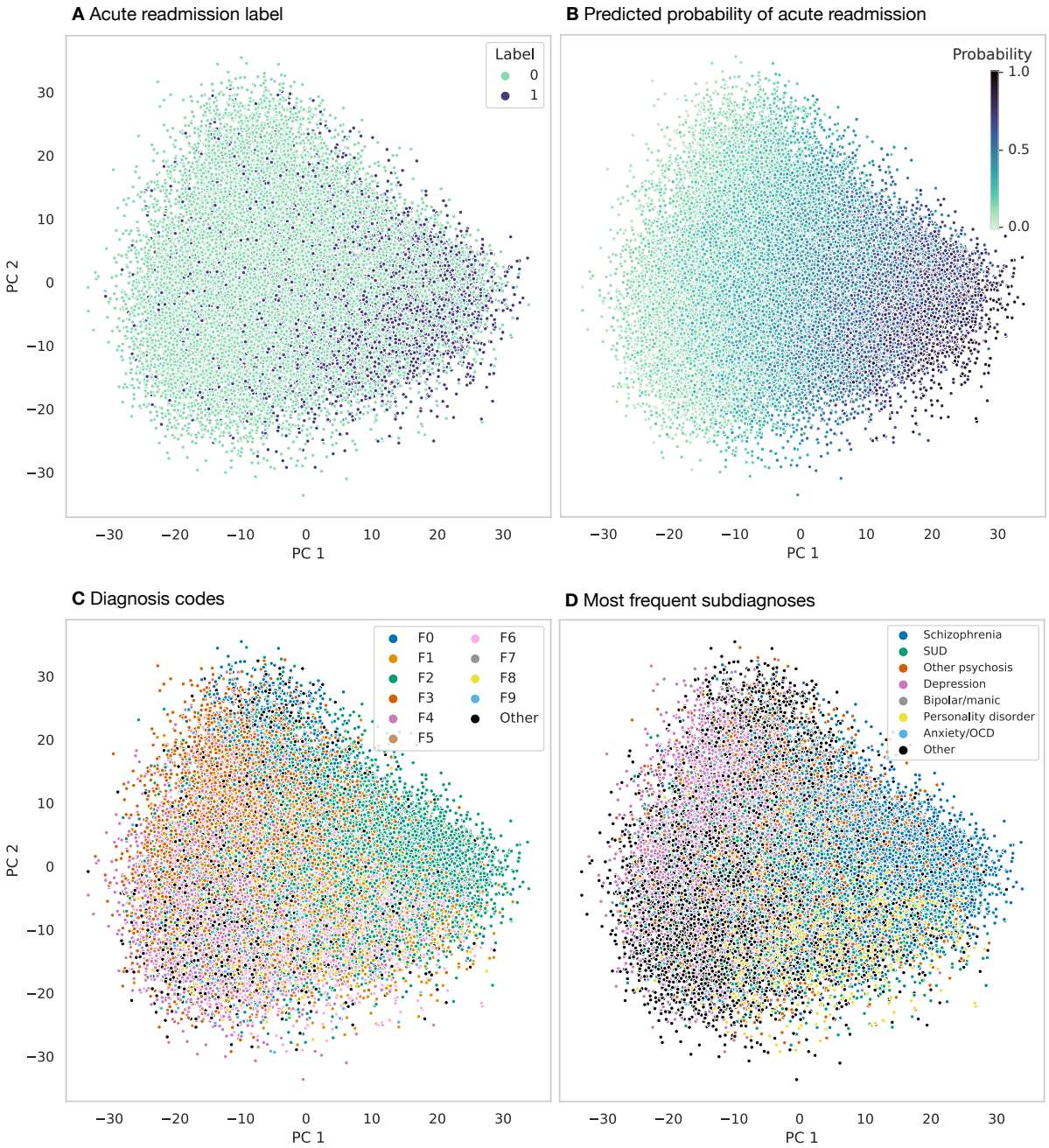
Attention-based analysis identifies key clinical factors underlying model predictions

An advantage of LLMs over Bag-of-Words models is the ability to interpret important terms in each individual note (local patterns), which can provide fine-grained explanations of the prediction for individual patients. Due to the sensitive nature of the clinical notes, we did not analyze individual notes. Instead, we aggregated and sorted terms (sequences of words) by how much attention weight PsyRoBERTa gave them locally (using the method called Attention Rollout (AR)[53]) to show the most highly weighted bigrams (**Figure 4A**), trigrams (**Figure 4B**), and trigrams of instances where the model predicted acute readmission with high certainty (Softmax[54,55] probability > 0.8). Through this, we identified five important categories of factors underlying model predictions: *psychosis*, *medication*, *level of function, alcohol and substances,* and *indications of a patient's lack of insight into their own illness* (**Figure 4E,** see also supplementary Results, **Figure S11-13**). Comparing PsyRoBERTa's AR scores to the coefficients of the LR classifier showed that both models recognized substance abuse as important (**Figure 4D**). However, LR (which was solely trained on unigram frequency and with no concept of word-order or time) seemed to rely most on terms describing a patient's history of previous readmissions, as well as indications of violence and somatic medical issues.

Figure 4. Most highly weighted terms for psychiatric acute readmission prediction

*Top 25 most highly weighted bigrams (**A**) and trigrams (**B**) sorted by their max AR scores. N-grams were translated from Danish (**Figure S6-8**). Dashes indicate ambiguity and other possible translations. **C.** Top 25 trigrams of instances where the model predicted acute re-admission with high certainty, defined as a Softmax probability > 0.8. **D.** Top 25 terms by logistic regression coefficients. **E**. The highly weighted terms were categorized as related to psychosis, medicine, level of function, alcohol and substances, and indications of lack of insight into own illness. "undifferentiated schizophrenia" appeared in various forms among the most highly weighted n-grams. The word "significant" was prominent in descriptions of the patient's level of function as well as descriptions of worsening and stabilization of the patient's condition. Additionally, "significant" was highly weighted when describing "significant abuse" (likely substance abuse), and other descriptions of alcohol and substances included the number of units (of alcohol), being sobered and "kisa" (a treatment center specialized in stabilizing patients with alcohol and substance abuse). In the last category, "accepted to come" could indicate that the patient might not want treatment; the term "refuse" could be related to some patients refusing a diagnosis or treatment; the trigram "described that pt" could indicate that the patient (abbreviated as "pt") was not themselves aware or able to describe an incidence; the term "allegedly" could indicate doubt in the veracity of the patient's statement.*

Figure 5. Principal Component Analysis with mean pooled embeddings (hidden states) of the discharge summary finetuning trainset



*The mean of embedded notes within an admission illustrated with PCA. **A**. The label indicates whether the next admission was acute (1) or not acute (0). **B**. The predicted probability of the next admission being acute. **C.** F0-9 are the 10 major psychiatric diagnosis codes. While patients may have more than one diagnosis, we show the current most severe diagnosis at the time of the admission. **D**. Most frequent subdiagnoses.*

PsyRoBERTa's representations of clinical notes enables feature clustering and diagnosis recognition

A beneficial side-effect of the contextual text representations, resulting from PsyRoBERTa, is that they can be used for other purposes than making predictions. With Principal Component Analysis (PCA), we explored PsyRoBERTa's learned patterns and ability to recognize important aspects within the notes. We found many admissions resulting in acute readmissions located closely in the embedding space (**Figure 5A**) (however with substantial overlap of clusters) and the readmission risk spectrum learned by PsyRoBERTa was clearly visible (**Figure 5B, Figure S14-15**). Exploring other aspects with PCA revealed that the model captured the age spectrum (**Figure S16B**) and to some extent sex (**Figure S16A**) and note length (**Figure S17**). Most notably, PCA analyses revealed that PsyRoBERTa effectively separated admissions related to the 11 major psychiatric diagnoses (**Figure 5C**), indicating the ability to perform diagnosis recognition as a by-product of training the model to predict acute readmissions. Patients' current diagnosis could thus be accurately recognized with a kNN classifier on pooled embeddings (AUC=0.832, MCC=0.489, k=50) (**Figure S18A**). The 7 most frequent sub-diagnoses were likewise well clustered (**Figure 5D**) and easily recognized (AUC=0.847, MCC=0.492, k=50) (**Figure S18B**).

DISCUSSION

We examined the effect of domain-specific pretraining by comparing our specialized model, pretrained on 39.5 million Danish clinical notes from 245,589 patients who have had contact with the mental health services, to those pretrained on more distant public domains, and we further analyzed data size effects in pretraining and finetuning. Our model, PsyRoBERTa, outperformed three public-domain LLMs and a logistic regression classifier on predicting psychiatric acute readmissions, with (AUC=0.736, MCC=0.303). Domain-specific pretraining with clinical notes enhanced performance by 15-26% in MCC. As expected, more pretraining, with more data, improved performance, and we found the gain of introducing new data to be diminishing for each new introduction, which is in line with previous research on pretraining RoBERTa-based models.[56,57] We made versions of PsyRoBERTa pretrained on less data, to be used for tasks with rare outcomes. In line with the related work on domain-specific pretraining, our results indicate that the closer the pretraining data is to the target domain, the more the performance can be expected to improve, since the medically-informed model MeDa-BERT performed competitively to PsyRoBERTa, particularly when using a large finetuning dataset. Since RoBERTa is an upgrade from BERT, comparing the two models does not allow for

making direct comparison of the impact of their respective pretraining data, and to do so, it would be necessary to train a PsyBERT and/or a MeDa-RoBERTa, which goes beyond our aim of comparing PsyRoBERTa to the relevant publicly available alternatives. Both PsyRoBERTa and MeDa-BERT exhibited more efficient learning during finetuning than the non-medical LLMs, meaning they require less data for finetuning for equal performance to non-medical models, which supports the need for domain-pretrained models for future clinical applications where data for finetuning is limited.[4,6,8,35,49,58,59]

PsyRoBERTa captured information about diagnoses and attended mostly to words related to psychosis, medication, the patient's level of function, alcohol and substance abuse, and indications of lack of insight into own illness. Some of these factors, such as the patient's level of function and insight into own illness, are not recorded in structured EHR and can therefore similarly not be captured by models built on structured data. Yet, these factors are closely related to clinical practice and gives insight to the clinicians' evaluations of patients from direct interactions with them. With PCA and kNN we found that PsyRoBERTa embedded information about patients' diagnoses and could therefore effectively recognize a patient's current diagnosis in unseen notes, which can be useful for making clinicians aware of diagnoses missed or unrecorded in the structured EHR data.

The simple LR classifier also recognized substance abuse as an important factor but gave more weight to a history of prior (re)admissions. we speculate that while prior readmissions are an important predictive feature easily captured by the LR classifier, more complex factors are better captured by LLMs. An important distinction between LR and PsyRoBERTa, that could explain this, is that learned coefficients of LR are static while PsyRoBERTa's weighing of terms is dynamic or *contextual*.[37] This means that the occurrence of 'bisoprolol' (medicine for high blood pressure), with a positive LR coefficient, will always count towards an increased risk of readmission by LR. PsyRoBERTa instead takes the context into account and gives different weight to the word accordingly. This is a great advantage since such a feature is likely to be spurious if given equal weigh regardless of context. However, context-dependence makes interpretability of general associations between features and predictions challenging, and learning good contextual representations requires substantially more model training. If a modest performance difference does not justify the increased complexity and cost of LLMs for a specific project, there are other advantages that might: We have shown that the contextual,

high-dimensional representations also allow for other ways of investigating patterns learned by the model and can be multi-purpose without additional training.

Our results indicated that predicting hospital readmissions is challenging, and related studies with English text data have reported similarly modest performance with AUC scores of 0.714,[10] 0.726,[60] and 0.798[11] for predicting ICU readmissions, psychiatric readmissions, and all-cause readmission, respectively. On the related task of predicting psychiatric *involuntary* admissions (which is expected to be somewhat easier as it involves a specific subgroup of patients and is likely less influenced by early discharges due to resource scarcity), Perfalk et al. (2024) reported AUCs of 0.76 with an ML model (XGBoost).[48] They proposed finetuning language models as a future direction to enhance performance compared to ML models. We found that finetuning general-domain LLMs (RøBÆRTa and BERT) did not result in significant increase of performance compared to a simple LR classifier, and hence it is crucial to apply an LLM pretrained with clinical data.

Limitations of this study includes the difficulty of providing robust explanations for LLMs' predictions.[61–63] The method used in this study, Attention Rollout, peeks into the attention mechanism, which is fundamental for transformer-based LLMs, however, it only partly explains the model's decision-making.[64] Furthermore, the LLMs used in this study were architecturally constrained to an input length that necessitated splitting some notes into smaller chunks and methods for encoding longer sequences could be a promising future direction.[14] We did not experiment with large generative language models (such as GPT, Llama, Gemini) because of their limitations: Closed-source generative models are currently not permissible for use with patient data due to privacy concerns, and open-source generative models are computationally expensive, their predictions are sensitive to prompt formulations, and recent studies have found that smaller task-specific RoBERTa-based models perform better at classification than large generative models.[65–67]

We have presented the first large language model specialized for psychiatric outcome prediction using clinical notes, with psychiatric acute readmissions as proof-of-concept. The model outperformed publicly available alternatives and achieved state-of-the-art performance in predicting readmission from Danish clinical notes, with comparable performance to related work in English. We showed that our model relied on factors related to every-day clinical practice and which are not found in structured EHR data. Furthermore, the model excelled at

diagnosis recognition without further training, highlighting the multi-purpose nature of LLMs. Importantly, PsyRoBERTa can be finetuned for many clinical use cases with the expectation of better performance and accelerated learning compared to non-clinical Danish LLMs.

METHODS

Study population and electronic health record data

Our data consists of 43,996,617 clinical notes from Danish population-based electronic health records covering around half of the Danish population (the Capital Region of Denmark and Region Zealand). The clinical notes were from 255,944 patients with at least one hospital contact with the mental health services between 2000 and 2022. The data consists of all the patients' clinical notes and was thus not limited to contact with psychiatry. We de-identified the notes by masking patients' personal and identifiable information and, additionally, developed a method for masking health professionals' names.

De-identification

Patients' personal and sensitive information, including patients' names, relatives' names, addresses, social security numbers, phone numbers, and email addresses were first masked by using Python regular expressions. Further masking of (unknown) names of doctors and other health professionals was performed with regular expressions and the use of lexicons derived from online resources.

Pretraining data

To ensure that our pretrained model could be used for unbiased evaluations of future downstream tasks, we split the data, based on randomized patient IDs, into four parts with approximately equal number of (non-overlapping) patients. This enabled pretraining in four stages with an increasing amount of data. Notes were concatenated and split to the maximum sequence length of the model and patients in the validation and test sets for our task were removed from the pretraining data. This resulted in a total of 39.5 million notes for pretraining, with 5% of these patients for MLM validation.

Finetuning data

Finetuning was based only on the clinical notes from psychiatric hospital admissions. Train, validation, and test set splits (70/10/20) were based on randomized patient IDs. Binary labels indicated whether the next admission was acute (1) or not acute (0), with acute defined as

unplanned and within 30 days after discharge. There was high class imbalance, with 19,790 cases (23%) and 65,757 controls (77%). We finetuned with datasets consisting of either 1) only the *discharge summaries* (~86 thousand), or 2) *all admission notes* (~3.4 million). For the latter, we additionally applied a deduplication method (see below). Long notes were split in chunks of the LLMs maximum input length but when finetuning on all admission notes, we concatenated notes within an admission, adding a separator token to signal the start of a new note within the text, before splitting.

De-duplication

De-duplication of notes followed the DedupCont method of Liu et al. (2022)[68] except that we only deduplicated within encounters (admissions) rather than across all notes. For each encounter, all notes were split on line-breaks (\n) into chunks. A chunk was removed if it was an exact match to any prior chunks in the encounter, and the chunks remaining were concatenated.

Pretraining of Psychiatric Clinical RoBERTa model (PsyRoBERTa)

We initialized PsyRoBERTa with weights from RøBÆRTa,[50] a RoBERTa-base[56] architecture consisting of 12 transformer encoder blocks and self-attention heads, a hidden size of 768, a vocabulary size of 50265 tokens, and a total of 125 million parameters. RøBÆRTa was pretrained on the Danish part of the multilingual MC4 dataset.[69] With our clinical notes, we continued training towards the masked language modelling (MLM) objective – a standard method for pretraining – predicting tokens that were randomly masked with a 15% probability. This pretraining was carried out in four consecutive stages, training for three epochs on each of the four pretraining data parts and resuming from the checkpoint of the previous model. The entire pretraining took ~62 hours and was carried out using HuggingFace's Accelerate library and on a Microsoft Azure Virtual Machine with 4 NVIDIA A100 GPU's (Standard_NC96ads_A100_v4) with a cost of $19.10 per hour. See all hyperparameters in **Table S1**.

Finetuning language models

To enable outcome prediction, a language model pretrained with MLM must be finetuned (trained) for the specific task. We finetuned PsyRoBERTa and three publicly available Danish LLMs (Danish BERT,[51] RøBÆRTa,[50] and MeDa-BERT[49]), for predicting psychiatric acute readmission, using either discharge summaries or all admission notes. In this process, a linear

classification head was added to each model and during training both the classification head and the LLM's weights were optimized for the prediction task, resulting in models that output a probability of the next admission being acute. All models were finetuned for 12 epochs and with equal hyperparameters (**Table S1**). The number of epochs was chosen based on inspecting the recline in training and validation losses (**Figure S19**) and with consideration for that finetuning should be cost-effective. We had chosen 12 as the maximum although the losses were persistently reclining and may have improved with more epochs, and we used the same number for all models for fair comparison. Due to high class-imbalance, we rescaled the weight given to each class in the cross-entropy loss function. The finetuning was carried out using PyTorch and Huggingface on a Microsoft Azure Virtual Machine with 4 NVIDIA A100 GPU's (Standard_NC96ads_A100_v4) with a cost of $19.10 per hour and it took ~14 hours with all admission notes and ~1.35 hours with discharge summaries for each of the four language models (BERT, MeDa-BERT, RøBÆRTa, PsyRoBERTa).

Logistic regression classifiers

The text was lowercased and some punctuation removed before training LR classifiers on the 5000 most frequent unigrams, using L2 regularization and balanced class weights. We used grid search to find the best feature representation for the classifiers (optimizing average precision): we searched two vectorizers (count and TF-IDF, see below) and n-gram ranges of (1,1), (1,2), (1,3) and (1,4), where a range of (1,4), for instance, means including both unigrams, bigrams, trigrams, and 4-grams. The vectorizers represent text in a Bag-of-Words style (where the text is represented by frequencies, disregarding order and context). The count vectorizer represents each note as a vector with counts of token occurrences, and Term Frequency-Inverse Document Frequency (TF-IDF) additionally re-weighs each token by its frequency within all notes in training set. The best feature representation of discharge summaries was counts of unigrams and the best for all admission notes was TF-IDF of unigrams.

Evaluating finetuned models

While there can be numerous clinical notes during a hospital admission, the LLMs provide predictions *per text* and some notes were split into several pieces of text. The LLMs outputs are Softmax[54,55] probabilities (between 0 and 1) for each text. To provide a single predicted probability of acute readmission, we took the mean of the texts' Softmax probabilities for the positive class for each admission. We evaluated predictive performance on a wide range of metrics and computed 95% confidence intervals with non-parametric bootstrapping using 1000

resamples. P-values were also calculated with bootstrapping (1000 resamples and 95% confidence level), calculating differences in the metric (MCC and AUC) of two models and computing the p-value for the null hypothesis that there is no difference in the performance. PsyRoBERTa was evaluated against three public LLMs to show the performance when pretraining on millions of clinical notes is not possible. However, the public models were also finetuned on clinical notes since they cannot predict outcomes out-of-the-box. We emphasized Matthews Correlation Coefficient (MCC) given its greater reliability for binary classification with imbalanced data.[52] The deduplication method was only evaluated on PsyRoBERTa and the LR classifier.

Explainability

We explained the predictions of PsyRoBERTa with Attention Rollout (AR),[53] which attributes a score (attention weight) to each input token indicating the token's importance for the prediction *locally* (i.e. specific for the input-prediction pair). Since weights were assigned to tokens of the RoBERTa tokenizer, which sometimes splits words into several tokens, we merged tokens and their weights to reconstruct the complete words. Additionally, we experimented with either using the mean or max of the weights when merging. Then, *to approximate global explanations*, we summed weights given to unique, lowercased words and normalized them by their frequency before inspecting highly weighted words. We focused the analysis on the 10% most frequent words, since infrequent words could be highly weighted but less relevant overall, and we only inspected words that were used in notes of at least three patients. We inspected highly weighted unigrams (1-word sequence), bigrams (2-word sequences) and trigrams (3-word sequences), including punctuations as "words". Treating LR coefficients as global explanations for this model, we compared highly weighted terms of LR to the AR scores.

Clustering and recognition

The transformation of clinical notes to multi-dimensional embeddings with finetuned PsyRoBERTa creates representations of words that are dynamic (meaning the representation changes depending on the context) and captures semantic knowledge (such that words used in similar contexts will have more similar representations). We used principal component analyses (PCA) to visualize these high-dimensional embeddings of PsyRoBERTa finetuned on discharge summaries. (Additional visualizations with UMAP in **Figure S20-23**). We performed inference on the train and test set of discharge summaries and mean pooled the last hidden

state. We then further mean pooled the several splits of notes, that may exist within an admission, to finally have one vector representation of each admission. We reduced the vectors for the train set to two dimensions with PCA and visualized the embedding space in relation to seven variables: true class, predicted probability, predicted class, sex, age, amount of note splits, and main psychiatric diagnosis. k-nearest neighbors (kNN) classifiers were trained and tested on the vectors, with k={5,10,50}, to evaluate the model's ability to cluster and recognize these important variables.

Fairness and bias

The term *fairness* is inherently value-based, and there exists several definitions that can appear conflicting.[70,71] In this study, we evaluate *group fairness*, which we define as non-discrimination for socio-demographic groups. This is rooted in the ideal that clinical prediction models should not discriminate based on attributes such as age, sex/gender, race/ethnicity and socio-economic status.[72] We evaluated group fairness through the lens of *equalized odds* (equality in true positive and true negative rates) and *performance* parity.[73,74] Furthermore, the explainability methods described above were used to reveal and better understand potential biases. (Un)fairness and bias are connected phenomena, since unintentional biases can result in disparate performance across groups. We investigated the demographic attributes sex (binary), four age groups (below 18, 18-34, 35-54, and above 54 years of age) and region (Capital Region and Region Zealand). Ethnicity, although a relevant factor for a bias analysis, was unfortunately underreported. We further investigated fairness across the most frequent (sub)diagnoses: schizophrenia, other psychosis, substance use disorder (SUD), depression, bipolar/manic, personality disorder, and anxiety/obsessive compulsive disorder (OCD).

Data and ethics approvals

This project is part of the PRECISE project where the overall aims is to utilizing health data to increase the understanding of mental disorders and to pave the way for improvements of current treatments for mental disorders. The project is approved by the Data Protection Agency of the Capital Region of Denmark and the Danish Society for Patient Safety (approval numbers: P-2020-101 and R-22002033). According to Danish regulations, the analyses do not require informed consent. All data are available to researchers in the regions after the relevant approvals from the Danish Health authorities. All patient data were pseudonymized and, as far as possible, only analyzed as summary data consisting of at least 3 patients in each group.

DATA AVAILABILITY

The data used in this study are not publicly available to ensure the security and compliance with patient privacy regulations. Data can only be available for those researchers that meet the criteria for access to the confidential data, to be used in the work for improving the evidence base and quality in the patient treatment. The data and model cannot be shared openly according to Danish regulations due to potential personal identifiable information but is accessible for researchers with appropriate legal permissions on secured servers.

CODE AVAILABILITY

The non personally identifiable part of the code will be shared on www.github.com/terne/PsyRoBERTa.

ACKNOWLEDGEMENTS

The work was supported by unrestricted grants from the Lundbeck Foundation (grant number R278-2018-1411 and R383-2022-285) and from the Mental Health Services of the Capital Region of Denmark. S.R. was supported by the Novo Nordisk Foundation (NNF23SA0084103).

AUTHOR CONTRIBUTIONS

Dr Benros had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Drs Rasmussen and Benros are considered co–senior authors.

Concept and design: Jakobsen, Rasmussen, Benros.

Acquisition, analysis or interpretation of data: Jakobsen, Cóppulo, Rasmussen, Benros.

Drafting of the manuscript: Jakobsen, Benros.

Critical review of the manuscript for important intellectual content: All authors.

Statistical analysis: Jakobsen.

Obtained funding: Benros, Rasmussen,

Administrative, technical, or material support: Cóppulo, Benros.

Supervision: Rasmussen, Benros.

COMPETING INTERESTS

S.R. is the founder and owner of the Danish company BioAI and have performed consulting for Sidera Bio ApS. No other disclosures were reported.

REFERENCES

1. Nielsen SD, Christensen RHB, Madsen T, Karstoft KI, Clemmensen L, Benros ME. Prediction models of suicide and non-fatal suicide attempt after discharge from a psychiatric inpatient stay: A machine learning approach on nationwide Danish registers. *Acta Psychiatr Scand*. 2023;148(6):525-537. doi:10.1111/acps.13629
2. Allesøe RL, Thompson WK, Bybjerg-Grauholm J, et al. Deep Learning for Cross-Diagnostic Prediction of Mental Disorder Diagnosis and Prognosis Using Danish Nationwide Register and Genetic Data. *JAMA Psychiatry*. 2023;80(2):146-155. doi:10.1001/jamapsychiatry.2022.4076
3. Allesøe RL, Nudel R, Thompson WK, et al. Deep learning-based integration of genetics with registry data for stratification of schizophrenia and depression. *Sci Adv*. 2022;8(26):eabi7293. doi:10.1126/sciadv.abi7293
4. Carrino CP, Armengol-Estapé J, Gutiérrez-Fandiño A, et al. Biomedical and Clinical Language Models for Spanish: On the Benefits of Domain-Specific Pretraining in a Mid-Resource Scenario. *ArXiv*. Published online 2021:2109.03570v2. http://arxiv.org/abs/2109.03570v2
5. Pedersen JS, Laursen MS, Soguero-Ruiz C, Savarimuthu TR, Hansen RS, Vinholt PJ. 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)*. IEEE; 2022. doi:10.1109/bhi56158.2022.9926955
6. Lehman E, Hernandez E, Mahajan D, et al. Do We Still Need Clinical Language Models? In: *Proceedings of Machine Learning Research*. Vol 209. 2023:2023. https://github.com/elehman16/clinical
7. Gururangan S, Marasović A, Swayamdipta S, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,*. Association for Computational Linguistics; 2020.
8. Gu Y, Tinn R, Cheng H, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans Comput Healthc*. 2021;3(1):1-23. doi:10.1145/3458754
9. Wang B, Xie Q, Pei J, et al. Pre-trained Language Models in Biomedical Domain: A Systematic Survey. *ArXiv*. Published online 2021:2110.05006v3. http://arxiv.org/abs/2110.05006v3
10. Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:190405342*. Published online 2020. https://arxiv.org/pdf/1904.05342
11. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. Published online 2023. doi:10.1038/s41586-023-06160-y
12. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5(1):194. doi:10.1038/s41746-022-00742-2
13. Van Aken B, Papaioannou JM, Mayrdorfer M, Budde K, Gers FA, Löser A. Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics; 2021:881-893.
14. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. *J Am Med Inform Assoc*. 2023;30(2):340-347. doi:10.1093/jamia/ocac225

15. Alsentzer E, Murphy J, Boag W, et al. Publicly Available Clinical BERT Embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2019:72-78. doi:10.18653/v1/W19-1909
16. Michalopoulos G, Wang Y, Kaka H, Chen H, Wong A. UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics; 2021:1744-1753. doi:10.18653/v1/2021.naacl-main.139
17. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics; 2019:58-65. doi:10.18653/v1/W19-5006
18. Rohanian O, Nouriborji M, Jauncey H, et al. Lightweight transformers for clinical natural language processing. *Nat Lang Eng*. 2024;30(5):887-914. doi:10.1017/S1351324923000542
19. Rawat BPS, Kovaly S, Yu H, Pigeon W. ScAN: Suicide Attempt and Ideation Events Dataset. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics; 2022:1029-1040. doi:10.18653/v1/2022.naacl-main.75
20. Gema A, Minervini P, Daines L, Hope T, Alex B. Parameter-Efficient Fine-Tuning of LLaMA for the Clinical Domain. In: *Proceedings of the 6th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2024:91-104. doi:10.18653/v1/2024.clinicalnlp-1.9
21. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit Med*. 2023;6(1):135. doi:10.1038/s41746-023-00879-8
22. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. doi:10.1038/sdata.2016.35
23. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10(1):1. doi:10.1038/s41597-022-01899-x
24. Lentzen M, Madan S, Lage-Rupprecht V, et al. Critical assessment of transformer-based AI models for German clinical notes. *JAMIA Open*. 2022;5(4):ooac087. doi:10.1093/jamiaopen/ooac087
25. Sundahl Laursen M, Skyttegaard Pedersen J, Søgaard Hansen R, Rajeeth Savarimuthu T, Just Vinholt P. Danish Clinical Named Entity Recognition and Relation Extraction. In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Association for Computational Linguistics; 2023:655-666. Accessed January 6, 2025. https://aclanthology.org/2023.nodalida-1.65/
26. Hao B, Zhu H, Ch Paschalidis I. Enhancing Clinical BERT Embedding using a Biomedical Knowledge Base. In: *Proceedings of the 28th International Conference on Computational Linguistics,*. 2020:657-661. Accessed January 16, 2025. https://github.com/noc-lab/clinical-kb-bert
27. Lu Q, Dou D, Nguyen TH. ClinicalT5: A Generative Language Model for Clinical Text. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics; 2022:5436-5443. Accessed January 16, 2025. https://aclanthology.org/2022.findings-emnlp.398.pdf
28. Huang K, Singh A, Chen S, et al. Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation. In: *Proceedings of the 3rd Clinical*

Natural Language Processing Workshop. Association for Computational Linguistics; 2020:94-100.
29. Boecking B, Usuyama N, Bannur S, et al. Making the Most of Text Semantics to Improve Biomedical Vision--Language Processing. Published online April 20, 2022. doi:10.1007/978-3-031-20059-5_1
30. Lin C, Miller T, Dligach D, Bethard S, Savova G. EntityBERT: Entity-centric Masking Strategy for Model Pretraining for the Clinical Domain. In: *Proceedings of the 20th Workshop on Biomedical Language Processing*. Association for Computational Linguistics; 2021:191-201. doi:10.18653/v1/2021.bionlp-1.21
31. El Boukkouri H, Ferret O, Lavergne T, Noji H, Zweigenbaum P, Tsujii J. CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations From Characters. In: *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics; 2020:6903-6915. doi:10.18653/v1/2020.coling-main.609
32. Yan A, McAuley J, Lu X, et al. RadBERT: Adapting Transformer-based Language Models to Radiology. *Radiol Artif Intell*. 2022;4(4):e210258. doi:10.1148/ryai.210258
33. Sushil M, Ludwig D, Butte AJ, Rudrapatna VA. *Developing a General-Purpose Clinical Language Inference Model from a Large Corpus of Clinical Notes*.; 2022.
34. Bressem KK, Adams LC, Gaudin RA, et al. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. Wren J, ed. *Bioinformatics*. 2021;36(21):5255-5261. doi:10.1093/bioinformatics/btaa668
35. Carrino CP, Llop J, Pàmies M, et al. Pretrained Biomedical Language Models for Clinical NLP in Spanish. In: *Proceedings of the 21st Workshop on Biomedical Language Processing*. Association for Computational Linguistics; 2022:193-199. doi:10.18653/v1/2022.bionlp-1.19
36. Verkijk S, Vossen P, Nl PTJMV. MedRoBERTa.nl: A Language Model for Dutch Electronic Health Records. In: *Computational Linguistics in the Netherlands Journal*. Vol 11. 2021:141-159.
37. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*. Published online 2018:1810.04805v1. http://arxiv.org/abs/1810.04805v1
38. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240. doi:10.1093/bioinformatics/btz682
39. Greco CM, Simeri A, Tagarelli A, Zumpano E. Transformer-based language models for mental health issues: A survey. *Pattern Recognit Lett*. 2023;167:204-211. doi:10.1016/j.patrec.2023.02.016
40. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In: *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. 2022.
41. Danielsen AA, Fenger MHJ, Østergaard SD, Nielbo KL, Mors O. Predicting mechanical restraint of psychiatric inpatients by applying machine learning on electronic health data. *Acta Psychiatr Scand*. 2019;140(2):147-157. doi:10.1111/acps.13061
42. Viani N, Yin L, Kam J, et al. Time Expressions in Mental Health Records for Symptom Onset Extraction. In: *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018)*. Association for Computational Linguistics; 2018:183-192.

43. Viani N, Kam J, Yin L, et al. Temporal information extraction from mental health records to identify duration of untreated psychosis. *J Biomed Semantics*. 2020;11(1):2. doi:10.1186/s13326-020-00220-2
44. Holderness E, Miller N, Cawkwell P, et al. Analysis of risk factor domains in psychosis patient health records. *J Biomed Semantics*. 2018;10:1-10. https://link.springer.com/article/10.1186/s13326-019-0210-8
45. Alvarez-Mellado E, Holderness E, Miller N, et al. Assessing the Efficacy of Clinical Sentiment Analysis and Topic Extraction in Psychiatric Readmission Risk Prediction. In: *Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*. Association for Computational Linguistics; 2019:81-86.
46. Ding X, Hall MH, Miller T. Incorporating Risk Factor Embeddings in Pre-trained Transformers Improves Sentiment Prediction in Psychiatric Discharge Summaries. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2020:35-40. doi:10.18653/v1/2020.clinicalnlp-1.4
47. Holderness E, Cawkwell P, Bolton K, Pustejovsky J, Hall MH. Distinguishing Clinical Sentiment: The Importance of Domain Adaptation in Psychiatric Patient Health Records. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2019:117-123. doi:10.18653/v1/W19-1915
48. Perfalk E, Damgaard JG, Bernstorff M, Hansen L, Danielsen AA, Østergaard SD. Predicting involuntary admission following inpatient psychiatric treatment using machine learning trained on electronic health record data. *Psychol Med*. 2024;54(15):4348-4361. doi:10.1017/S0033291724002642
49. Jannik Skyttegaard Pedersen Pernille Just Vinholt Thiusius Rajeeth Savarimuthu MSL. MeDa-BERT: A medical Danish pretrained transformer model. In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. doi:10.18653/v1/w19-5006
50. Højmark-Bertelsen M, Saattrup Nielsen D, Junge K, Kloster Pedersen M, Kummervold P, Moëll B. RøBÆRTa. Hugging Face. 2021. Accessed January 6, 2025. https://huggingface.co/DDSC/roberta-base-danish
51. Certainly. Danish BERT. Hugging Face. 2021. Accessed January 1, 2024. https://huggingface.co/Maltehb/danish-bert-botxo
52. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1):6. doi:10.1186/s12864-019-6413-7
53. Abnar S, Zuidema W. Quantifying Attention Flow in Transformers. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2020:4190-4197. doi:10.18653/v1/2020.acl-main.385
54. Bridle JS. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In: *Proceedings of the 3rd International Conference on Neural Information Processing Systems*. MIT Press; 1989:211-217.
55. Bishop CM. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag; 2006.
56. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Published online July 26, 2019.
57. Pérez-Mayos L, Ballesteros M, Wanner L. How much pretraining data do language models need to learn syntax? In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2021:1571-1582. doi:10.18653/v1/2021.emnlp-main.118

58. Pais C, Liu J, Voigt R, Gupta V, Wade E, Bayati M. Large language models for preventing medication direction errors in online pharmacies. *Nat Med*. 2024;30(6):1574-1582. doi:10.1038/s41591-024-02933-8
59. Muizelaar H, Haas M, van Dortmont K, van der Putten P, Spruit M. Extracting patient lifestyle characteristics from Dutch clinical text with BERT models. *BMC Med Inform Decis Mak*. 2024;24(1):151. doi:10.1186/s12911-024-02557-5
60. Boag W, Kovaleva O, McCoy TH, Rumshisky A, Szolovits P, Perlis RH. Hard for humans, hard for machines: predicting readmission after psychiatric hospitalization using narrative notes. *Transl Psychiatry*. 2021;11(1):32. doi:10.1038/s41398-020-01104-w
61. Søgaard A. *Explainable Natural Language Processing*. 1st ed. Springer International Publishing; 2021. doi:10.1007/978-3-031-02180-0
62. Li D, Hu B, Chen Q, He S. Towards Faithful Explanations for Text Classification with Robustness Improvement and Explanation Guided Training. In: *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. Association for Computational Linguistics; 2023.
63. Jacovi A, Goldberg Y. Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness? In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2020:4198-4205.
64. Achtibat R, Hatefi SMV, Dreyer M, et al. AttnLRP: Attention-Aware Layer-Wise Relevance Propagation for Transformers. *ArXiv*. Published online 2024:2402.05602v2. http://arxiv.org/abs/2402.05602v2
65. Rogers A, Luccioni AS. Position: Key Claims in LLM Research Have a Long Tail of Footnotes. In: *Proceedings of the 41 St International Conference on Machine Learning*. 2024.
66. Kostina A, Dikaiakos MD, Stefanidis D, Pallis G. Large Language Models For Text Classification: Case Study And Comprehensive Review. Published online January 14, 2025.
67. Edwards A, Camacho-Collados J. Language Models for Text Classification: Is In-Context Learning Enough? In: *LREC-COLING*. ELRA Language Resource Association; 2024:10058-10072.
68. Liu J, Capurro D, Nguyen A, Verspoor K. "Note Bloat" impacts deep learning-based NLP models for clinical prediction tasks. *J Biomed Inform*. 2022;133:104149. doi:10.1016/j.jbi.2022.104149
69. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*. 2020;21(140):1-67. https://www.jmlr.org/papers/v21/20-074.html
70. Friedler SA, Scheidegger C, Venkatasubramanian S. The (Im)possibility of fairness. *Commun ACM*. 2021;64(4):136-143. doi:10.1145/3433949
71. Binns R. On the apparent conflict between individual and group fairness. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM; 2020. doi:10.1145/3351095.3372864
72. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378
73. Williamson RC, Menon AK. Fairness risk measures. In: *Proceedings of the 36th International Conference on Machine Learning*. 2019:6786-6797.

74. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput Surv*. 2021;54(6):1-35. doi:10.1145/3457607