PsyRoBERTa: A Large Language Model for Predicting Psychiatric Outcomes from Clinical Notes - A Population-Based Danish EHR study

Authors

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

Key points

Question: Can large language models be used for psychiatric outcome prediction based on clinical notes?

Findings: In this prognostic study including 255,944 individuals and ~40 million clinical notes, the performance of a large language model for predicting psychiatric acute readmission improved when adapting the model for the clinical domain by pretraining on clinical notes. The model showed descent classification performance and demonstrated multi-purpose abilities with its excellence in recognizing psychiatric diagnoses.

Meaning: These findings demonstrate that adapting large language models to the clinical domain facilitates large-scale utilization of clinical notes for both psychiatric outcome prediction and recognition.

ABSTRACT

Importance: Machine Learning has been widely applied on *structured* Electronic Health Records (EHR) to predict psychiatric outcomes. However, the complex symptom descriptions and anamneses of psychiatric patients are found in the *unstructured* clinical notes. which have been rarely utilized for psychiatric research and outcome prediction. However, large language models (LLM) now enable large-scale utilization of clinical notes.

Objective: To develop a LLM for predicting psychiatric outcomes from clinical notes using population-based EHR data from the entire eastern part of Denmark.

Design, setting and participants:
This prognostic study included ~44 million Danish clinical notes, written between 2000 and 2022, from 255,944 patients who have had contact with the mental health services. A LLM was pretrained on ~40 million notes and finetuned on a dataset of 85,547 psychiatric admissions for predicting psychiatric acute readmissions. The model was evaluated against three publicly available models, pretrained on either public general- or medical-domain text, and a baseline logistic regression classifier. Data was analyzed between April 2023 and November 2024.

Exposure: At least one contact with the mental health services in the eastern part of Denmark.

Main outcomes and measures: Predictability of 1) masked tokens (word-pieces) measured by cross-entropy loss, 2) psychiatric acute readmission, defined as an unplanned readmission within 30 days after discharge, and 3) psychiatric diagnosis recognition, evaluated by AUC, MCC and additional metrics, as well as explainability and fairness analyses.

Results: Our specialized LLM, PsyRoBERTa, outperformed three Danish LLMs in predicting psychiatric acute readmissions (AUC:0.736; MCC:0.303) and was significantly better ($p<0.05$) than a baseline LR classifier (AUC:0.718; MCC:0.258). A public medical-domain model, MeDa-BERT, was a close second (AUC:0.734; MCC:0.295). Five categories of important features were identified (psychosis, medication, level of function, alcohol and substances, and lack of insight into own illness). PsyRoBERTa was furthermore able to recognize patients' main current diagnoses through diagnosis-based clustering of the clinical notes (AUC:0.832; MCC:0.489).

Conclusions and relevance: To the best of our knowledge, we present the first clinical LLM specialized for the psychiatric domain. This is a first step towards large-scale utilization of the unstructured EHR data with the prospect of improving patient care.

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) to predict health outcomes. 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 data is challenging. Although a main advantage of LLMs is that they can be applied in a *transfer learning* fashion, meaning they only need brief *finetuning* for new tasks, there is substantial evidence suggesting they should be pretrained for the target domain to achieve best performance and parameter-efficiency.[4–7] This is especially important for clinical text, which differs considerably, in vocabulary and structure, from the internet-originating text most general-purpose LLMs are based on.

Several LLMs have been pretrained for the clinical domain with English EHR notes.[8–18] A recent review[19] identified 23 LLMs pretrained specifically on clinical text and found 17 of them to be trained on the same source data (MIMIC-III/IV),[20,21] highlighting the challenge of attaining large, high-quality clinical text datasets and a current lack of diversity among available models. Furthermore, English data is by far dominating the rapid development of clinical LLMs while development for other languages, such as Danish, is only slowly progressing.[4,5,22,23] **Figure 1** gives an extended overview of LLMs pretrained on clinical notes.[5,7–18,24–34] The availability of LLMs for investigating mental health issues is currently much more limited[35] and so far, utilizing clinical notes for psychiatric outcome prediction has 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,[36] psychosis symptom onset,[37,38] and readmission.[39–42]

To the best of our knowledge, 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. We evaluate our model on predicting acute readmissions in psychiatry and recognizing psychiatric diagnoses, and we use explainability methods for thorough analyses of the model's behavior.

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 (**eMethods**).

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 (**Figure 2C**). 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 (**Figure 2B**), 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 (**Figure 2B**). 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 (**Figure 2C**). 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 (**eMethods**). 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.

Pretraining of Psychiatric Clinical RoBERTa model (PsyRoBERTa)

We initialized PsyRoBERTa with weights from RøBÆRTa,[43] a RoBERTa-base[44] 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.[45] 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 (**Figure 2C**). The entire pretraining took ~62 hours using 4 A100 GPUs (**eMethods,** hyperparameters in **eTable 1**).

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,[46] RøBÆRTa,[43] and MeDa-BERT[47]), 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 (**eMethods, eTable 1**). Due to high class-imbalance, we rescaled the weight given to each class in the cross-entropy loss function.

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. Hyperparameters for feature representation were optimized using grid search (**eMethods**).

Evaluation and analysis methods

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. 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 (**Figure 2C**). We evaluated predictive performance on a wide range of metrics and computed 95% confidence intervals with non-parametric bootstrapping. 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.[48] The deduplication method was only evaluated on PsyRoBERTa and the LR classifier.

Explainability

We explained the predictions of PsyRoBERTa with Attention Rollout (AR),[49] 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). To approximate *global explanations*, we summed weights given to unique, lowercased words (after merging tokens, **eMethods**), normalized by frequency and focused on the 10% most frequent words and inspected highly weighted unigrams (1-word sequences), 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 (although LR coefficients are towards the positive class while AR scores are undirected).

Clustering and recognition

Principal Component Analysis (PCA) was used to visualize embeddings produced by PsyRoBERTa finetuned on discharge summaries. We mean-pooled embeddings of discharge summaries into one representation per admission and visualized the train set with 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 representations to evaluate the model's ability to cluster and recognize these important variables (**eMethods**).

Fairness and bias

Group fairness was evaluated through the lens of *equalized odds* (equality in true positive and true negative rates) and *performance parity.*[50,51] We investigated the demographic attributes sex (binary), age groups (below 18; 18-34; 35-54; 55 and above) and region (Capital Region and Region Zealand). We further investigated fairness across the most frequent (sub)diagnoses (**eMethods**).

RESULTS

Development of PsyRoBERTa

We developed PsyRoBERTa, a large language model pretrained on 39.5 million Danish clinical notes consecutively trained for three epochs on each of the four parts of the pretraining data. 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 of increasing training data 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 (**eFigure 2**), the last model improved slightly faster, with a 1% mean difference in validation loss over 12 epochs of finetuning. Consequently, using all available data for pretraining resulted in the lowest validation loss for both MLM and prediction but the gain of increasing training data was diminishing for each new data part.

Predicting psychiatric acute readmissions with clinical notes

Domain-specific pretraining with clinical notes improves prediction of psychiatric acute readmissions

PsyRoBERTa consistently outperformed the non-clinical models on four metrics (AUC, MCC, 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] When finetuning with *discharge summaries*, PsyRoBERTa (MCC:0.285; 95% CI: 0.269-0.304) was significantly better ($p<0.05$)[52] than RøBÆRTa (MCC:0.226; CI:0.208-0.244), BERT (MCC:0.215; CI:0.198-0.231) and LR (MCC:0.223; CI:0.208-0.239), but not MeDa-BERT (MCC:0.264; CI:0.248-0.283), on the MCC metric, and showed a 28% improvement from the LR baseline. Finetuning with *all admission notes* improved all models, PsyRoBERTa (MCC:0.303; CI:0.287-0.322) still being significantly better than RøBÆRTa (MCC:0.263; CI:0.245-0.282), BERT (MCC:0.193; CI:0.175-0.212) and LR (MCC:0.258; CI:0.245-0.271), but not MeDa-BERT (MCC:0.295; CI:0.278-0.312), with a 17% improvement from the LR baseline. Deduplication improved the performance of PsyRoBERTa by 2.4% while reducing training time by ~3 hours but did not improve the LR classifier. All LLMs tended to overestimate the risk of readmission but exhibited better calibration when finetuned on all notes (**eResults, eFigure 18**).

Domain-specific pretraining reduces data requirements

The difference in performance given discharge summaries versus all notes was smaller for PsyRoBERTa (**Figure 3B**), with an absolute difference of 0.018 on the MCC metric for PsyRoBERTa versus 0.031, 0.037, 0.022, and 0.035 for MeDa-BERT, RøBÆRTa, BERT and LR, respectively, indicating that PsyRoBERTa needed less finetuning data for performance increases.

Attention-based analysis identifies key clinical factors underlying model predictions

Terms were sorted by their Attention Rollout (AR)[49] scores (**eFigure 7-8**, **eResults)** to show the top 25 most highly weighted bigrams (**Figure 4A**), trigrams (**Figure 4B**), and trigrams of instances where the model predicted acute readmission with high certainty (SoftMax probability > 0.8). Five important categories were identified in the highly weighted terms: *psychosis*, *medication*, *level of function, alcohol and substances,* and *indications of a patient's lack of insight into their own illness* (**Figure 4E**). Comparing PsyRoBERTa's AR scores to the coefficients of the LR classifier showed that both models recognized alcohol and substance abuse as important (**Figure 4D**). However, LR seemed to rely more on terms related to a patient's history of being hospitalized and previous readmissions.

Principal Component Analysis reveals class- and diagnosis-based clustering of clinical note representations

PCA analyses revealed that PsyRoBERTa effectively separated admissions related to the 11 major psychiatric diagnoses (**Figure 5C**). Patients' current diagnosis could thus be accurately recognized with a kNN classifier on pooled embeddings (AUC:0.832; MCC:0.489, K=50) (**eFigure 13A**). The 7 most frequent sub-diagnoses were likewise well clustered (**Figure 5D**) and easily recognized (AUC:0.847; MCC: 0.492, K=50) (**eFigure 13B**). Many admissions resulting in acute readmissions were located closely in the embedding space (**Figure 5A-B**), however, visible overlap of the two classes also occurred. Using a kNN classifier to predict the classes resulted in a worse classification performance (AUC:0.698; MCC:0.219, K=50) than the prior approach with a neural network classification head (AUC:0.711; MCC:0.285) (**eFigure 10**). It was predominantly patients with schizophrenia who were located within the space of admissions leading to acute readmissions, which aligned with the AR scores indicating the model attended to terms related to psychosis.

Fairness analysis indicates performance consistency across diagnoses and patient demographics

PsyRoBERTa performed best for patients with schizophrenia (TPR:0.480, AUC:0.750) or personality disorders (TRP:0.531, AUC:0.771) and worst for patients with anxiety (TPR:0.061; AUC:0.662) or depression (TPR:0.005, AUC:0.660) – the diagnoses with lowest proportions of acute readmissions (**eFigure 14, eFigure 16**). With PsyRoBERTa finetuned on all admission notes, we observed small differences in true positive rates regarding sex and age (**eFigure 13**), with a higher rate for males (TPR:0.365) than females (TPR:0.293), and a higher rate for the age groups 18-34 (TPR:0.325) and 35-54 (TPR:0.432), than below 18 (TPR:0.193) and above 54 (TPR:0.127) years of age. However, these differences were opposite for true negative rates, a trade-off reflected in good performance parity for all sex-, region- and age-groups, except for the patients above 54 years of age (AUC:0.669), where the model performed slightly below average (AUC:0.736). Similar results were observed for PsyRoBERTa finetuned on discharge summaries (**eFigure 15**), but with additional slight performance disparity for patients below 18 years of age (AUC:0.677) compared to the average (AUC:0.711). This aligned with PCA plots (**eFigure 10B**) where embeddings clearly captured the age spectrum.

DISCUSSION

We developed the first large language model specialized for psychiatric outcome prediction, based on the currently largest sample for psychiatry, consisting of 245,589 patients and more than 40 million Danish clinical notes. Our model, PsyRoBERTa, outperformed three public-domain LLMs and a logistic regression classifier on predicting psychiatric acute readmissions with an AUC of 0.736. When finetuned using the same data the medically informed model, MeDa-BERT, performed competitively to PsyRoBERTa, supporting the benefit of domain-specific pretraining with clinical notes and of pretraining on data related to this clinical domain.

Few studies have utilized NLP for predicting *psychiatric* readmissions.[39–41,53] Boag et al. (2021) reported an AUC of 0.726 (±0.001) with a LR classifier trained on a combination of demographic features, features from discharge summaries, and the number of previous psychiatric admissions – the latter being in itself a highly predictive feature.[54] Although existing mono-lingual models trained on other languages cannot be directly compared to ours, as they cannot be applied on Danish, we note that previous studies predicting *non-psychiatric* readmission based on *English* clinical notes have reported similar performance. ClinicalBERT, pretrained on MIMIC, achieved an AUC of 0.714 (±0.018) for predicting ICU readmissions.[8]

NYUTron, pretrained on the notes of ~387k patients from hospitals in New York, achieved an AUC of 0.798 (±0.002) for all-cause readmission prediction.[9]

PsyRoBERTa captured information about diagnoses and seemed to attend 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. Whereas the simple LR classifier also recognized alcohol and substance abuse as an important factor but gave more weight to a history of prior (re)admissions. This might indicate that while prior readmissions are an important predictive feature easily captured by the LR classifier, the other more complex factors are better captured by LLMs. 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.

With PCA and kNN we found that PsyRoBERTa effectively embedded information about patients' diagnoses. Although this kNN classifier was inapplicable for predicting *future* risk, since diagnoses were written in the notes, the analysis showed PsyRoBERTa's ability to capture this information, which could be useful for making clinicians aware of diagnoses missed or unrecorded in the structured EHR data.

Strengths and limitations

Main strengths are the population-based data containing the largest collection of clinical notes with psychiatric relevancy, spanning multiple hospitals, departments, and psychiatric diagnoses. Since most research in clinical NLP is based on the US-based MIMIC datasets, which are relatively small and have limited geographical and departmental representation, this study is an important contribution towards assessing LLMs for real-world clinical outcome prediction. Another strength of this study is the novel pretraining strategy which considered the need for unbiased evaluations of future tasks by training models on incrementally more data. Limitations of this study includes the difficulty of explaining LLMs' predictions.[55] The method used in this study, Attention Rollout, peeks into the attention mechanism, which is fundamental for transformers-based LLMs, however, it only partly explains the model's decision-making.[56] Due to the vastness of our dataset, we opted for pretraining for only three epochs with each data part (12 epochs and 243,612 steps in total). This is comparable to previous studies pretraining clinical language models with approximately 150,000-200,000 steps.[8,13,57] Whereas, MeDa-BERT was pretrained on a smaller dataset but with four times the amount of

epochs (48), which might contribute to the small performance difference of PsyRoBERTa and MeDa-BERT. Furthermore, our finetuning datasets were large, giving public-domain models a better chance of "catching up" to PsyRoBERTa, and the upper bound on performance for predicting readmissions is low within the literature, as outlined above. Lastly, the LLMs used in this study were constrained to an input length that meant some notes had to be split into smaller chunks and future advances in methods for encoding longer sequences might further improve the clinical use of LLMs.[12]

CONCLUSIONS

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 clinical notes. Furthermore, the model excelled at recognizing current psychiatric diagnoses, highlighting the flexibility 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 LLMs.

ACKNOWLEDGEMENTS

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.

Funding/support

The work was supported by unrestricted grants from the Lundbeck Foundation (grant number R278-2018-1411 and R380-2021-1225) and from the Mental Health Services of the Capital

Region of Denmark. S.R. was supported by the Novo Nordisk Foundation (NNF23SA0084103).

Role of the Funder/Sponsor

The funders of the study had no role in the study design, data collection, data analysis, or in preparation, review, or approval of the manuscript, or the decision to submit the manuscript for publication.

Conflict of interest Disclosures

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.

Data and Ethics statement

The non personally identifiable part of the code will be shared. The data underlying this article are not publicly available according to Danish regulations to ensure the security and compliance with patient privacy regulations. Data are only 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. All patient data were pseudonymized for research purposes, and only accessible through secure servers provided by the region, with access control and the use is under strict confidentiality. Patient data are, as far as possible, only analyzed as summary data consisting of at least 3 in each group for non-person identifiable use of the data and only exported for research purposes in this manner. The overall aim of the larger project is to utilize health data to increase the understanding of mental disorders and to pave the way for improvements to the current treatments of 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.

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. 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; 2020:8342-8360. doi:10.18653/v1/2020.acl-main.740
7. 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
8. 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
9. 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

10. 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
11. 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.
12. 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
13. 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
14. 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
15. 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
16. 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
17. 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
18. 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

19. 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
20. 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
21. 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
22. 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
23. 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/
24. 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
25. 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
26. 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.
27. 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
28. 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

29. 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
30. 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
31. Sushil M, Ludwig D, Butte AJ, Rudrapatna VA. *Developing a General-Purpose Clinical Language Inference Model from a Large Corpus of Clinical Notes*.; 2022.
32. 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
33. 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
34. 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.
35. 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
36. 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
37. 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.

38. 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
39. 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
40. 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.
41. 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
42. 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
43. 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
44. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Published online July 26, 2019.
45. 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
46. Certainly. Danish BERT. Hugging Face. 2021. Accessed January 1, 2024. https://huggingface.co/Maltehb/danish-bert-botxo
47. 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

48. 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
49. 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
50. Williamson RC, Menon AK. Fairness risk measures. In: *Proceedings of the 36th International Conference on Machine Learning*. ; 2019:6786-6797.
51. 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
52. Cumming G. Inference by eye: reading the overlap of independent confidence intervals. *Stat Med*. 2009;28(2):205-220. doi:10.1002/sim.3471
53. Rumshisky A, Ghassemi M, Naumann T, et al. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl Psychiatry*. 2016;6(10):e921. doi:10.1038/tp.2015.182
54. 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
55. Søgaard A. *Explainable Natural Language Processing*. 1st ed. Springer International Publishing; 2021. doi:10.1007/978-3-031-02180-0
56. 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
57. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. *ArXiv*. Published online 2022:2201.11838v3. http://arxiv.org/abs/2201.11838v3

FIGURES

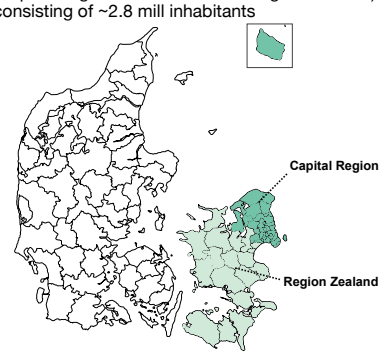
Figure 1. Overview of prior language models pretrained on clinical notes

	Bio+ClinicalBERT[13]	BlueBERT[15]	Clinical Llama-LoRA[18]	Clinical-kb-bert[24]	Clinical-Longformer[12]	Clinical-T5 base (Lehman)[7]	Clinical-T5 large (Lehman)[7]	Clinical-T5 base (Lu)[25]	Clinical-T5 large (Lu)[25]	ClinicalBERT[8]	Clinical BERT embeddings[13]	ClinicalDistilBERT[16]	ClinicalXLNet[26]	CORe[41]	CXR-BERT[27]	EntityBERT[28]	GatorTron small[10]	GatorTron medium[10]	GatorTron large[10]	Medical Domain BERT[29]	Medical Domain CharacterBERT[29]	medRoberta[7]	NYUTron[9]	RadBERT[30]	UCSFBert[31]	UmlsBERT[14]	Bert-for-Radiology[32]	Biomedical-clinical language model for Spanish[33]	Clinical ELECTRA[5]	MedRobBERTa.nl[34]	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.

Figure 2. Graphical abstract with data description and study overview

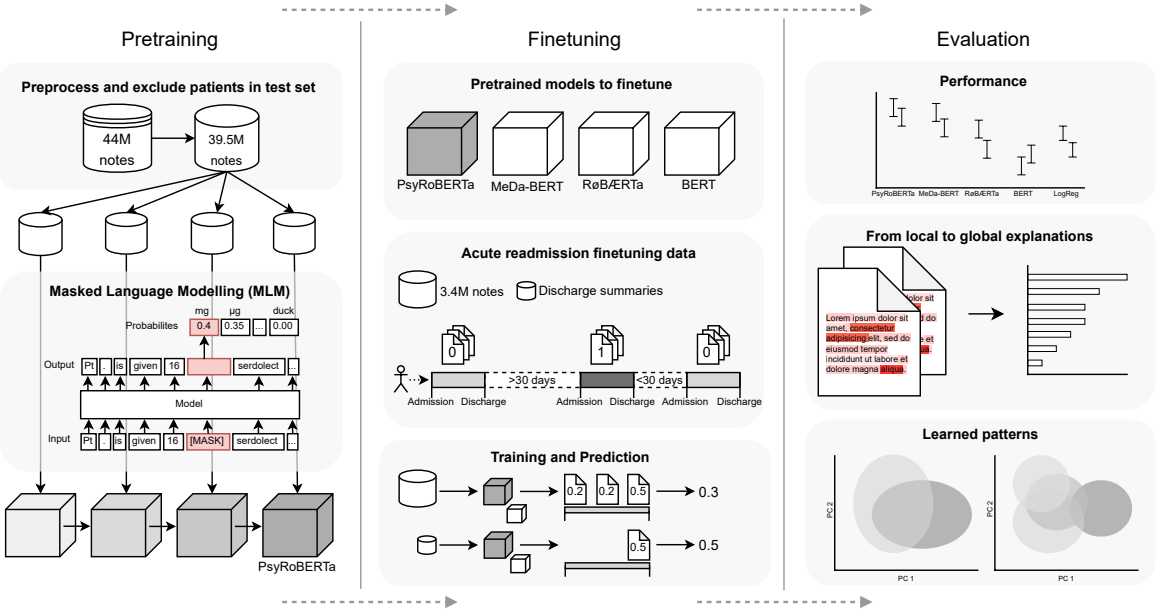
A EHR data from the eastern part of Denmark (Capital Region of Denmark and Region Zealand) consisting of ~2.8 mill inhabitants



B Summary of data

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

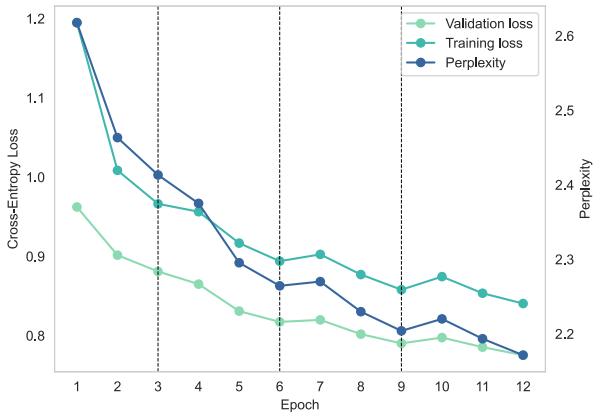
C Study overview



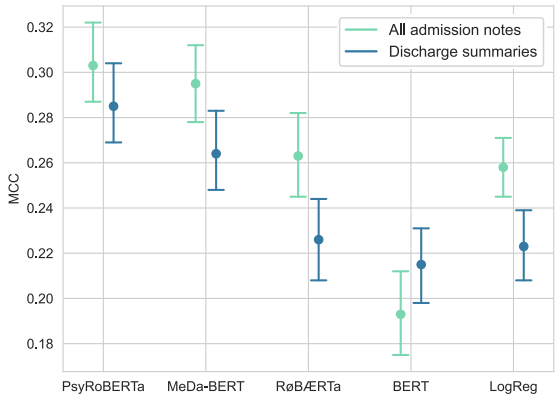
*A. Origin of EHR data. **B.** Size of pretraining and finetuning datasets. Raw GB refers to the size of a text file containing only the notes. Tokens were counted after tokenization with the RoBERTa tokenizer (byte-level BPE). Cases and controls refer to the admissions that lead or did not lead to acute readmissions. **C.** Study overview.*

Figure 3. Model performance

A Masked Language Modelling performance of PsyRoBERTa



B Predictive performance of all models by MCC

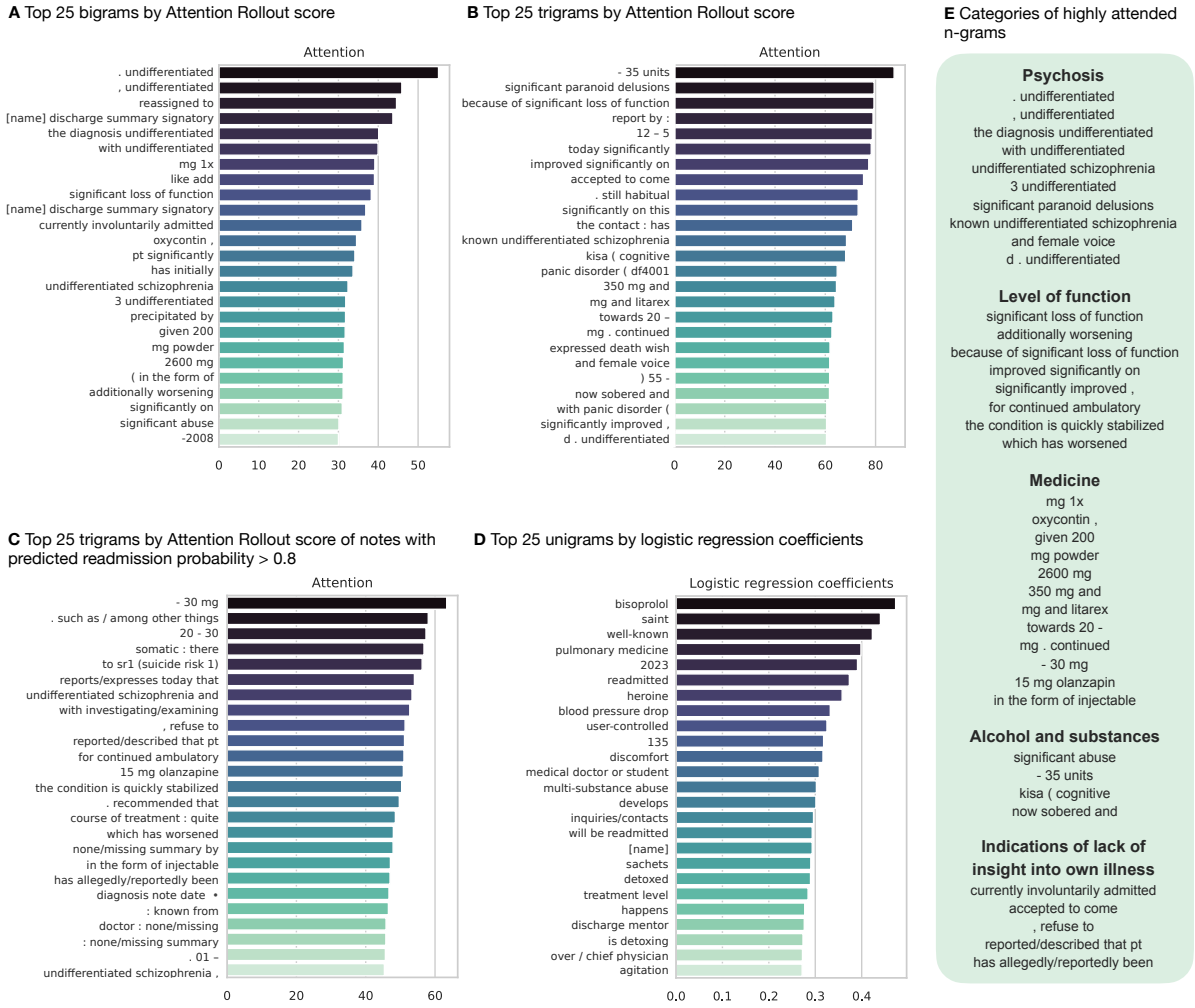


C All performance results of models finetuned for predicting psychiatric acute readmission

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]

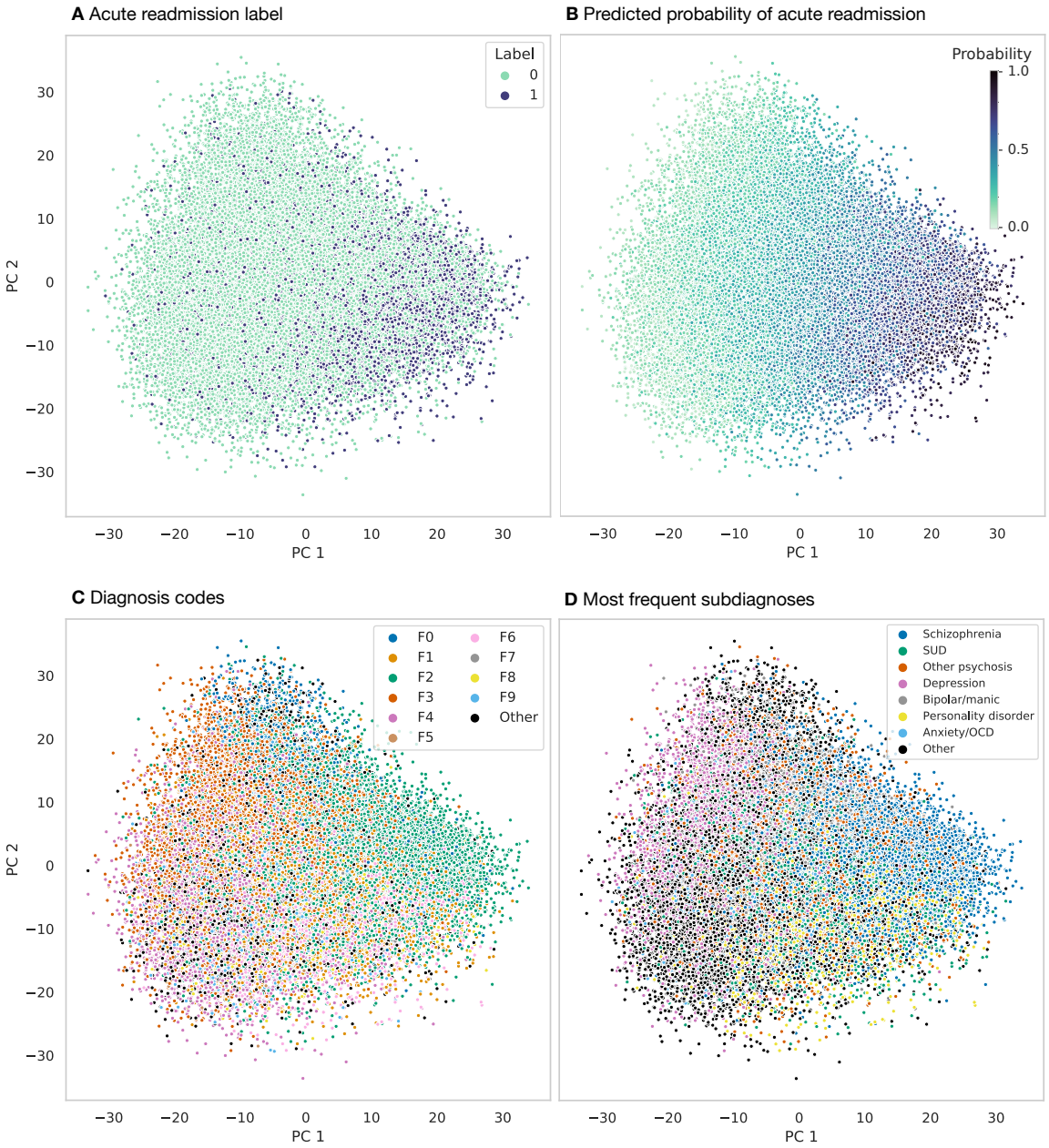
A. *Performance of the masked language modelling objective in pretraining. The striped line indicates the first epoch of the next model in our four-stage continuous pretraining.* ***B.*** *MCC of PsyRoBERTa, MeDa-BERT, RøBÆRTa, BERT and logistic regression. For each model on the x-axis, the left-most 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.*

Figure 4. Top 25 most highly weighted bigrams and trigrams by Attention Rollout score



*Top 25 most highly weighted bigrams (**A**) and trigrams (**B**) sorted by their max AR scores. N-grams were translated from Danish (**eFigure 6-8**). 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.*