





Lexical stability of psychiatric clinical notes from electronic health records over a decade

Lasse Hansen^{1,2,3} , Kenneth Enevoldsen^{2,3}, Martin Bernstorff^{1,2}, Erik Perfalk^{1,2}, Andreas A. Danielsen^{2,4}, Kristoffer L. Nielbo³ and Søren D. Østergaard^{1,2} 

Original Article

Cite this article: Hansen L, Enevoldsen K, Bernstorff M, Perfalk E, Danielsen AA, Nielbo KL, and Østergaard SD. (2023) Lexical stability of psychiatric clinical notes from electronic health records over a decade. *Acta Neuropsychiatrica* 1–11. doi: [10.1017/neu.2023.46](https://doi.org/10.1017/neu.2023.46)

Received: 20 June 2023
Revised: 15 August 2023
Accepted: 18 August 2023

Keywords:

Natural language processing; Data mining; Electronic health records; Mental disorders; Psychiatry

Corresponding author:

Lasse Hansen; Email: lasse.hansen@clin.au.dk

¹Department of Affective Disorders, Aarhus University Hospital – Psychiatry, Aarhus, Denmark; ²Department of Clinical Medicine, Aarhus University, Aarhus, Denmark; ³Center for Humanities Computing, Aarhus University, Aarhus, Denmark and ⁴Psychosis Research Unit, Aarhus University Hospital - Psychiatry, Aarhus, Denmark

Abstract

Objective: Natural language processing (NLP) methods hold promise for improving clinical prediction by utilising information otherwise hidden in the clinical notes of electronic health records. However, clinical practice – as well as the systems and databases in which clinical notes are recorded and stored – change over time. As a consequence, the content of clinical notes may also change over time, which could degrade the performance of prediction models. Despite its importance, the stability of clinical notes over time has rarely been tested. **Methods:** The lexical stability of clinical notes from the Psychiatric Services of the Central Denmark Region in the period from January 1, 2011, to November 22, 2021 (a total of 14,811,551 clinical notes describing 129,570 patients) was assessed by quantifying sentence length, readability, syntactic complexity and clinical content. Changepoint detection models were used to estimate potential changes in these metrics. **Results:** We find lexical stability of the clinical notes over time, with minor deviations during the COVID-19 pandemic. Out of 2988 data points, 17 possible changepoints (corresponding to 0.6%) were detected. The majority of these were related to the discontinuation of a specific note type. **Conclusion:** We find lexical and syntactic stability of clinical notes from psychiatric services over time, which bodes well for the use of NLP for predictive modelling in clinical psychiatry.

Significant outcomes

- Clinical notes in the electronic health records (EHRs) from a large psychiatric service in Denmark were found to be lexically stable over time, with no major changes observed on the aggregate level in the usage of terms describing psychopathology.
- The discontinuation of a specific note type and an increase in the use of virtual communication during the COVID-19 pandemic were the main factors contributing to the identified changepoints in the EHR content.
- While there were some changes in the usage of terms describing mood disorders during the COVID-19 pandemic, the overall novelty in the use of words related to psychopathology was relatively low, suggesting consistency in the content of the clinical notes over time. However, gradual changes in note length and other characteristics may have implications for the performance of prediction models over time, emphasising the need for continuous monitoring and evaluation.

Limitations

- The calculation of novelty based on word frequency may not identify emerging terms or shifts in writing style, potentially missing important changes in the clinical notes.
- The analysis is sensitive to relative changes in rare terms describing psychopathology, which might increase the risk of finding artefacts due to changes in the patient population.
- The study's changepoint detection models may not be sensitive to slow distribution shifts and drifts, requiring manual analysis and inspection to identify such changes in the clinical notes.

© The Author(s), 2023. Published by Cambridge University Press on behalf of Scandinavian College of Neuropsychopharmacology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use and/or adaptation of the article.



Introduction

In medical care, electronic health record (EHR) systems are used to document all patient-related information. The type of recorded information is highly heterogeneous and covers lab results, diagnoses, administered medication, as well as subjective (reported by the patient) and objective (findings made by the healthcare professionals) descriptions of behaviour and symptoms in the form of text in clinical notes (Häyrynen *et al.*, 2008). Notably, the clinical notes are not only a site for collection – but also of aggregation and synthesis – of clinical information (Rosenbloom *et al.*, 2010). Especially in psychiatry, important clinical information is often documented as text (Abbe *et al.*, 2016).

As the volume of information increases in EHR systems, it may become increasingly harder for healthcare professionals to make meaningful decisions based on this information (Cimino, 2013). In these situations, decision support based on analysis of the EHR data can guide clinicians towards relevant information and suggest courses of action (Sutton *et al.*, 2020). Whereas structured numerical data can relatively easily be aggregated with conventional statistical descriptors (e.g. minimum and maximum values, means, medians etc.), extracting relevant information from free-form clinical text requires a different approach (Wang *et al.*, 2018). In recent years, however, the field of natural language processing (NLP) has made marked progress (Vaswani *et al.*, 2017; Devlin *et al.*, 2019; Otter *et al.*, 2020), and textual information is no longer outside the purview of, for example, machine learning algorithms. Indeed, the utility of written text for prediction models in healthcare has been investigated in multiple studies and has been found to improve performance (Goldstein *et al.*, 2017; Boag *et al.*, 2018; Rajkomar *et al.*, 2018; Danielsen *et al.*, 2019; Huang *et al.*, 2019; Huang *et al.*, 2019). However, the stability of this predictive performance relies on the EHR source's stability, including the text in the clinical notes. Specifically, the performance of prediction models tends to degrade markedly the more the data used for prediction differs from that used for training (Marcus, 2018). Formally, in the case of distribution shift or drift, the assumption of identically distributed data no longer holds. This can lead to spurious or highly time-dependent patterns in prediction, which are unlikely to generalise (Shen *et al.*, 2021).

Hospital services using EHR are highly dynamic; for instance, clinical procedures, treatments, diagnostic criteria, the distribution of tasks between healthcare professions, healthcare staff composition and EHR systems tend to change substantially over time. The EHR information contained in written clinical notes may be particularly prone to changes over time, but – to our knowledge – the stability of the written clinical notes from EHR is never directly tested. Therefore, we aimed to assess the lexical stability over time of written clinical notes from a large psychiatric hospital system. More specifically, we aimed to answer the question of lexical stability using data from the Psychiatric Clinical Outcome Prediction (PSYCOP) cohort (Hansen *et al.*, 2021), which contains EHR clinical notes covering almost 130,000 patients and more than 14,000,000 clinical notes over the 10 years from 2011 to 2021. We employed changepoint detection models to examine whether (i) the use of words describing psychopathology (symptoms) in clinical notes in the EHR is stable over time and (ii) if sentence length, readability, syntactic complexity, and usage of different types of clinical notes are stable over time. The methods employed and the results obtained may serve as a benchmark for future studies testing the lexical stability of clinical notes from other hospital systems.

Table 1. Demographic characteristics of the cohort

Age	<18		≥18	
	Females	Males	Females	Males
Last ICD-10 diagnosis				
F00–F09 Organic mental disorders	<10*	<10*	5235	3696
F10–F19 Mental and behavioural disorders due to psychoactive substance use	40	113	967	2501
F20–F29 Schizophrenia, schizotypal and delusional disorders	361	245	3275	4459
F30–F39 Mood [affective] disorders	850	407	12,355	7621
F40–F48 Neurotic, stress-related and somatoform disorders	3184	1654	10,865	8505
F50–F59 Behavioural syndromes associated with physiological disturbances and physical factors	1158	132	1910	553
F60–F69 Disorders of adult personality and behaviour	383	38	3469	1129
F70–F79 Mental retardation	244	408	790	904
F80–F89 Disorders of psychological development	1711	5074	427	852
F90–F98 Behavioural and emotional disorders with onset usually occurring in childhood and adolescence	3773	7827	2713	3714
Other diagnoses	3168	3432	9693	9723

The table shows the number of unique patients grouped by their last ICD-10 diagnosis, age group at the first visit and sex. * If less than 10 patients in a group, the cell count is set at <10 to avoid the identification of individuals. See Supplementary Table 2 for further age stratification.

Material and methods

Data source

We used data from an updated version of the PSYCOP cohort (Hansen *et al.*, 2021). Specifically, we analysed all EHR clinical notes from all patients with at least one contact with the Psychiatric Services of the Central Denmark Region in the period from January 1, 2011, to November 22, 2021, covering 129,570 patients and 14,811,551 notes. In the EHR, the clinical notes are labelled according to their content, for example, 'current mental state', 'current social state' or 'medication' (all labels are listed in the Supplementary Information). For the present study, we selected 17 note types, which are among the most widely used and rich in text. See Supplementary Table 1 for a description of the note types.

Tables 1 and 2 (and Supplementary Table 2) show dataset characteristics related to the number of patients and note types. The dataset contains data from a total of 129,570 patients, of whom 66,575 (51%) are female, and is representative of all age groups. The dataset contains more than 1 billion tokens and 14.8 million notes across the 17 selected note types, with a mean note length of 71.5 tokens.

Table 2. Number and length of notes by note type sorted in descending order based on total number of tokens

Note type	Number of notes	Total tokens	Mean tokens per note
Observation of patient, psychiatry	2,972,943	202,733,055	68.19
Current mental state	2,227,647	354,778,113	159.26
Medication	1,837,916	70,759,048	38.5
Objective mental state	1,485,799	66,716,343	44.9
Appointments, psychiatry	1,404,610	38,665,226	27.53
Conversation with treatment aim	967,541	77,400,486	80.00
Current social function, psychiatry	710,314	48,003,618	67.58
Current somatic, psychiatry	513,777	19,739,080	38.42
Telephone consultation	426,126	24,772,013	58.13
Plan	424,641	21,544,611	50.74
Reason for contact	421,300	20,679,254	49.08
Patient note	419,942	28,029,332	66.75
Conclusion	398,857	35,871,674	89.94
Telephone note	246,163	16,875,687	68.55
Prescription	199,870	3,852,471	19.27
Objective, somatic	99,915	4,746,572	47.51
Semi-structured diagnostic interview	54,190	24,425,900	450.75
Total	14,811,551	1,059,592,483	71.50

Data processing

For the analysis (i) of the stability of descriptions of psychopathological symptoms over time, a total of 365 terms deemed most important for describing psychopathology were chosen by authors EP and AAD (both registrars in psychiatry). The chosen terms were based on psychopathology described in the Present State Examination (Danish version) (Pedersen and Bertelsen, 2018) and the ICD-10 (World Health Organization, 1993). The 365 terms describing psychopathology were grouped by the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) diagnostic category which the symptom best described. Subsequently, we counted how many times (summed at the quarterly level) each of the terms describing psychopathology – grouped by ICD-10 diagnostic category as well as a category for terms used in a mental state examination – occurred in each of the 17 types of clinical notes. Terms that occurred less than 10 times in 5 or more quarters were removed to reduce the sparsity of the keyword matrix and thereby increase the robustness of the analysis. A total of 56 terms were removed due to this criterion (see Supplementary Table 3 for a full list of included and excluded keywords), resulting in 309 terms used for the analysis. For each diagnostic group (ICD-10: F0–F9 and mental status examination), as well as for all terms describing psychopathology together (aggregated words describing psychopathology), the counts of terms describing psychopathology were transformed to proportions to calculate the *novelty* of each quarter

to the previous two quarters. *Novelty* is a measure of windowed relative entropy that expresses information novelty as a difference in content from the past and is further described (under ‘Metrics’) below.

For the analysis (ii) of the stability of sentence length, readability metrics, syntactic complexity and usage of the clinical notes over time, we extracted the mean note length, the mean automated readability index (ARI) (Kincaid *et al.*, 1975) and mean dependency distance from each note type using the *spaCy* v3.1.1 (Montani *et al.*, 2021) and *textdescriptives* v1.0.6 (Hansen *et al.*, 2022) Python libraries and the *spaCy da_core_news_lg* model. For these analyses, we randomly sampled 10% of the notes to reduce computational costs. The proportion of total notes was calculated by dividing the number of notes of each type by the total number of notes (out of the 17 note types). Each metric was aggregated as the quarterly mean, and the resulting time series of quarterly means were then analysed using changepoint detection algorithms. The aggregation was carried out to reduce noise and computational costs. The metrics, as well as the changepoint detection analysis, are described in more detail below.

Data from before 2013 were found to be highly noisy and therefore excluded from the main analyses (Bernstorff *et al.*, 2022). The noise was largely due to the gradual implementation of a new EHR system in the region starting in 2011 (Hansen *et al.*, 2021; Bernstorff *et al.*, 2022). See Supplementary Figure 1 for more details.

Metrics

Novelty

Novelty measures the average divergence between the present keyword distribution $s^{(j)}$ and past keyword distributions $s^{(j-1)}$, $s^{(j-2)}$, ..., $s^{(j-w)}$ in window w :

$$\text{Novelty}_w(j) = \frac{1}{w} \sum_{d=1}^w \text{JSD}(s^{(j)} | s^{(j-d)})$$

where JSD is the Jensen–Shannon divergence:

$$\text{JSD}(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$$

where P and Q are probability distributions, $M = \frac{1}{2}(P + Q)$ and D is the Kullback–Leibler divergence.

Novelty can be interpreted as the degree of relative entropy or ‘surprise’ in the distribution based on the recent past. The novelty will rise if word usage changes, but if the change becomes more permanent, the novelty will fall again as it becomes the new normal. Many continual changes, that is, low predictability, cause high novelty, and conversely, few changes cause novelty to be low. A visual representation of how changes affect novelty is shown in Fig. 1. Code for calculating novelty can be found on the following Github repository: <https://github.com/Aarhus-Psychiatry-Research/lexical-stability>

Mean note length

We calculated the mean note length as the mean number of tokens in each document. A token is a meaningful segment of text and includes whitespace-separated words, punctuation and more (Webster and Kit, 1992). For the present study, the documents were tokenised using *spaCy*’s Danish tokenisation module (Honnibal *et al.*, 2020).

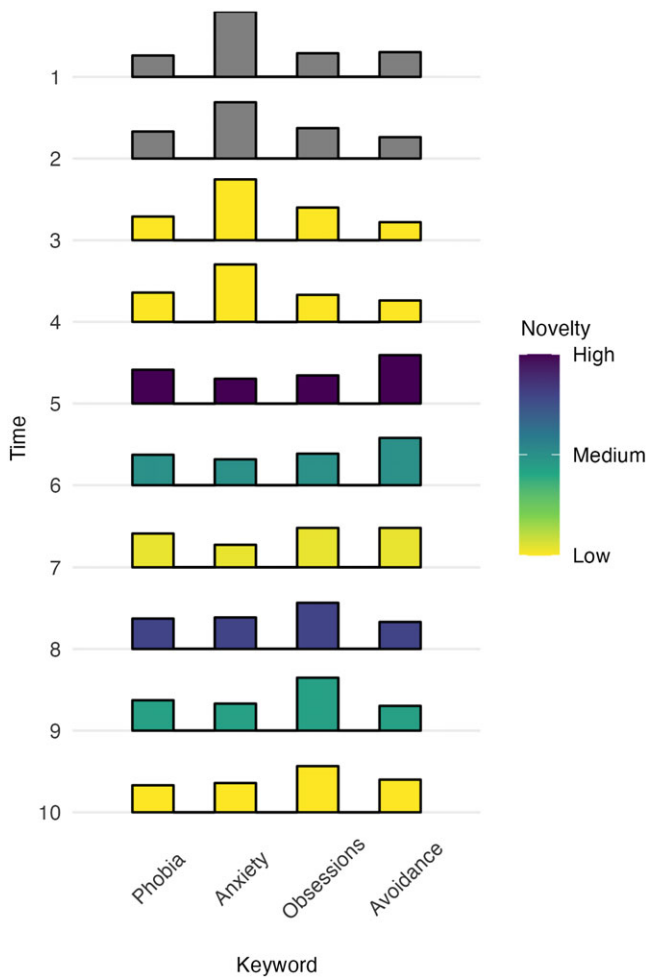


Figure 1. Example of novelty calculation with a window size of 2. The figure shows the distribution shift over time of four keywords from the F4 category. The distribution of the four keywords at each time point is represented with a histogram, with the height of the bars indicating word counts and the colour indicating novelty levels. Major shifts such as from time 4 to 5 lead to high novelty, which gradually decreases as the distribution becomes stable. The larger the shift in distribution, the greater the novelty. Novelty cannot be calculated for the first two points as we have defined the window size to be 2.

Automated readability index

The ARI is a measure of the readability of a text, designed to estimate the grade level required to comprehend the text (Kincaid et al., 1975). The equation below shows the formula for calculating the ARI:

$$ARI = 4.71 \left(\frac{\text{characters}}{\text{words}} \right) + 0.5 \left(\frac{\text{words}}{\text{sentences}} \right) - 21.43$$

Dependency distance

We define dependency distance as the number of words between a word and its headword. The mean dependency distance is the average dependency distance of a document or sentence, as visualised in Fig. 2. The dependency tree is obtained by performing a dependency parse using spaCy, and dependency distance was calculated using *textdescriptives* (Hansen et al., 2023).

Changepoint detection

All time series (83) were analysed for multiple changepoints in both mean and variance using the Pruned Exact Linear Time (PELT) algorithm (Killick et al., 2012). Specifically, the time series were (1) one for each of the 17 note types for each for the 4 descriptive statistics (mean number of tokens, mean dependency distance, ARI and proportion of notes (68 in total)), (2) three time series for the 17 note types combined (mean number of tokens, mean dependency distance and ARI) and (3) a time series for the novelty of each of the 10 ICD-10 diagnostic categories of words describing psychopathology, one for the novelty of words used in the mental state examination, and one for the novelty of all words (aggregate) describing psychopathology (12 in total). PELT automatically finds the appropriate number of changepoints by iteratively searching for new changepoints at each data split. This makes PELT particularly suitable for tasks without prior information on the location and number of changepoints. Variables were detrended by differencing with a lag of one time step and z-scored before input to PELT using the default parameters of the *cpt.meanvar* function (penalty = 'MBIC', minseqlen = 2). Changepoint detection was conducted in R v3.6.1 (R Core Team, 2019) using the *tidyverse v1.3.1* (Wickham et al., 2019) and *changepoint v2.2.2* (Killick and Eckley, 2014) packages. Given the 83 time series with a data point for each quarter for nine years (2013 Q1–2021 Q4), the number of data points (and hence the number of possible changepoints) was $4 \times 9 \times 83 = 2988$.

Stability of the use of words describing psychopathology

The analyses of the stability in the use of the 309 words describing psychopathology (Fig. 5) were conducted by identifying the words that had the largest relative change from the mean of the two preceding time points to the time point of interest. The relative change in word usage was quantified as the ratio between the current novelty value and the mean of the two previous points. The divergence metric in the novelty equation is sensitive to relative

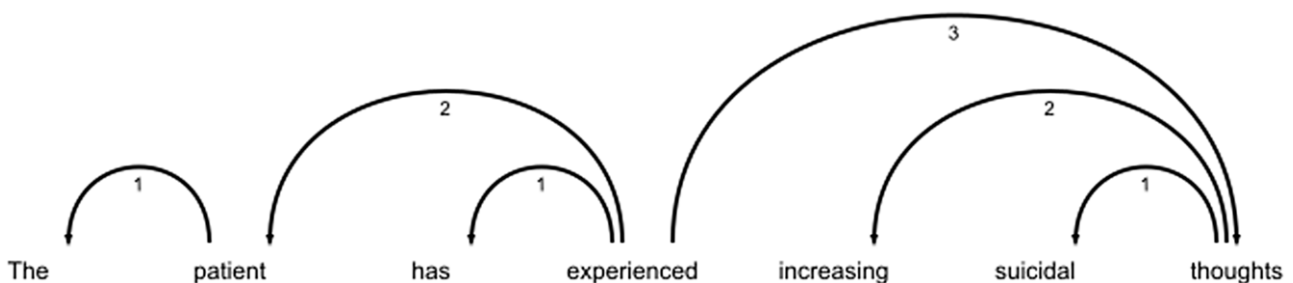


Figure 2. Calculation of dependency distance. Numbers on arrows indicate the distance from the head to the word. Note that ‘experienced’ has no arrows pointing to it, as it is the root of the sentence. The dependency distance of ‘experienced’ is therefore zero. The mean dependency distance of the sentence is $(1 + 2 + 1 + 0 + 2 + 1 + 3) / 7 = 1.43$.

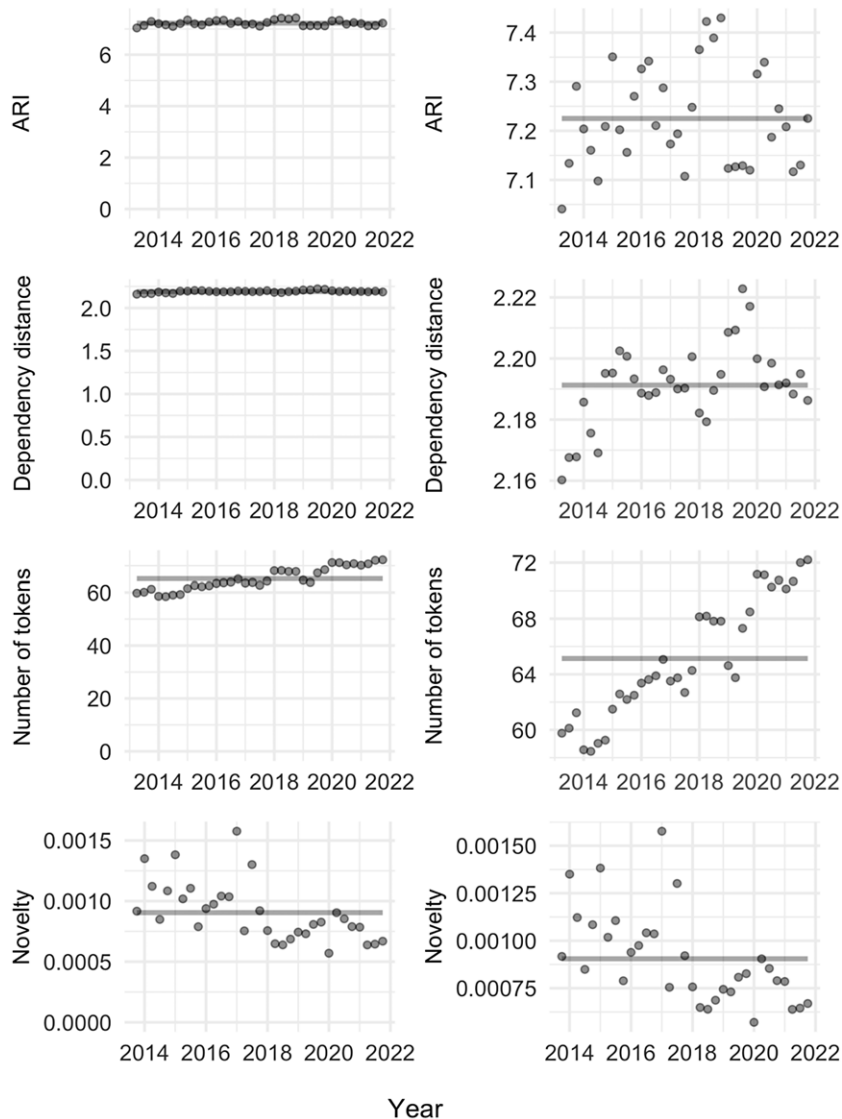


Figure 3. Time series of text stability. The figure panels illustrate the automated readability index (ARI), mean dependency distance, the mean number of tokens for the aggregated clinical notes (top three rows) and novelty across all psychopathological keywords. The grey line indicates the estimated changepoint segments (none found for any groups). The left column shows the data with the y-axis going to 0, and the right column shows the data with the y-axes allowed to vary.

changes in proportions as opposed to absolute changes, that is, a change in proportional usage of a keyword from 0.4 to 0.2 will have the same impact on novelty as a change from 0.1 to 0.05. Hence, words that are used rarely are more likely to have large fluctuations in relative proportions and are therefore stronger drivers of novelty than more commonly used words.

Sensitivity analyses

To assess the effect of different window sizes for calculating novelty, we reran the analysis on the stability of descriptions of psychopathological symptoms over time with window sizes of 2, 4, 6 and 8 quarters, respectively. To investigate the robustness of the analysis on the stability of sentence length, readability metrics, syntactic complexity and usage of the clinical notes over time, we reran the analysis on a different 10% random sample of the data. Finally, upon a suggestion from a reviewer, we also employed an alternative model for dependency parsing (DaCy large (Enevoldsen *et al.*, 2021)).

Code availability

All code used for extracting and analysing the data can be found on the following Github repository: <https://github.com/Aarhus-Psychiatry-Research/lexical-stability>. While not all steps can be reproduced due to data privacy, we have provided the aggregated data files used for the changepoint analyses.

Results

Stability across all notes and words describing psychopathology

As seen in Fig. 3, the mean number of tokens, mean dependency distance, ARI and novelty on the aggregate level (across all clinical notes or all words describing psychopathology) were relatively stable over the course of the study. No changepoints were found in any of the groups. However, the mean number of tokens increased steadily over time.

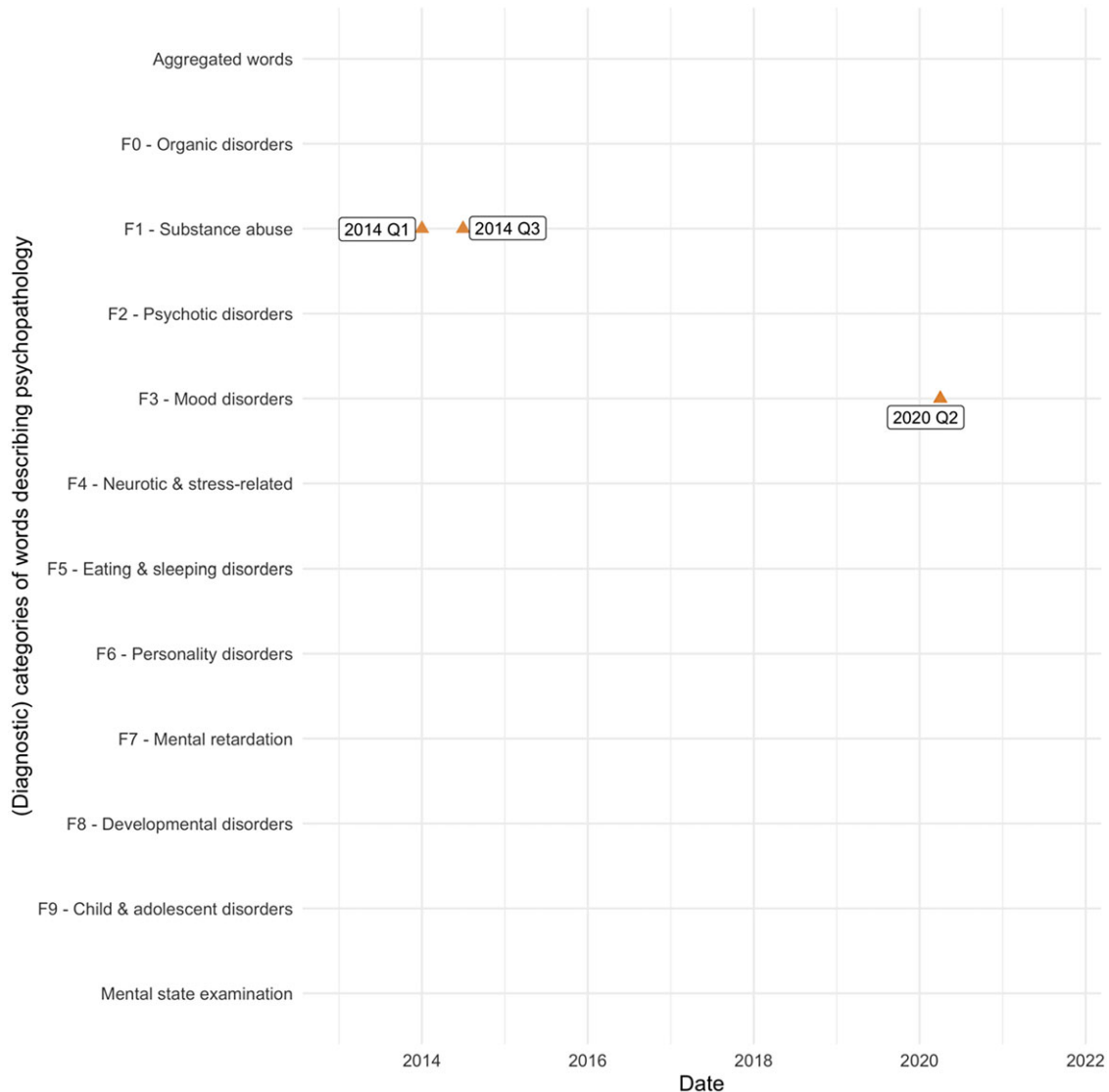


Figure 4. Location of changepoints on novelty for terms across diagnostic categories.

Analysis i: Stability in the use of terms describing psychopathology

As shown in Fig. 4, a total of three changes in novelty were found in 2 out of the 12 (diagnostic) categories of terms describing psychopathology (ICD-10 codes F0–F9, objective description and aggregated words), namely F1 (substance abuse) and F3 (mood disorders). The novelty was low across the categories, which indicates that the distributions are generally stable over time. Visual inspection of the identified changepoints in F1 – substance abuse (Supplementary Figure 2) suggests that these are artefacts rather than actual changepoints as they do not map onto any visually salient changes. The novelty for words describing psychopathology related to substance abuse was extremely low, which suggests that the distribution is particularly stable. For F3 – mood disorders, the changepoint reveals a slight rise in novelty in 2020 Q2, followed by a consistent drop (see Supplementary Figure 2). Figure 5 shows that this rise was mainly driven by decreased use of ‘nihilistic delusions’ and ‘grandiose delusions’ as well as increased use of ‘dysthymia’.

Analysis ii: Stability of sentence length, readability metrics and syntactic complexity of clinical notes

Figure 6 shows the changepoints for the mean number of tokens, mean dependency distance, ARI and proportion of total notes for each note type. Across the 17 note types and 4 categories of variables, a single changepoint was identified for the number of tokens, two for dependency distance, four for ARI and seven for note-type proportion (a total of 14 changepoints). The changepoints were concentrated mainly around the year 2020; however, most of the estimated changepoints were due to deviations from a linear trend or represented artefacts caused by reduced usage of the note type due to administrative changes (see the discussion for further explanation). No changepoints were found for eight of the note types, nor when analysing all notes together (aggregate).

Supplementary Figures 3–6 show the time series for the mean number of tokens, dependency distance, ARI and proportion of total notes for all note types with the output from the changepoint model overlaid. Half of the estimated changepoints (7 out of 14) were related to changes in how large a proportion of the total notes

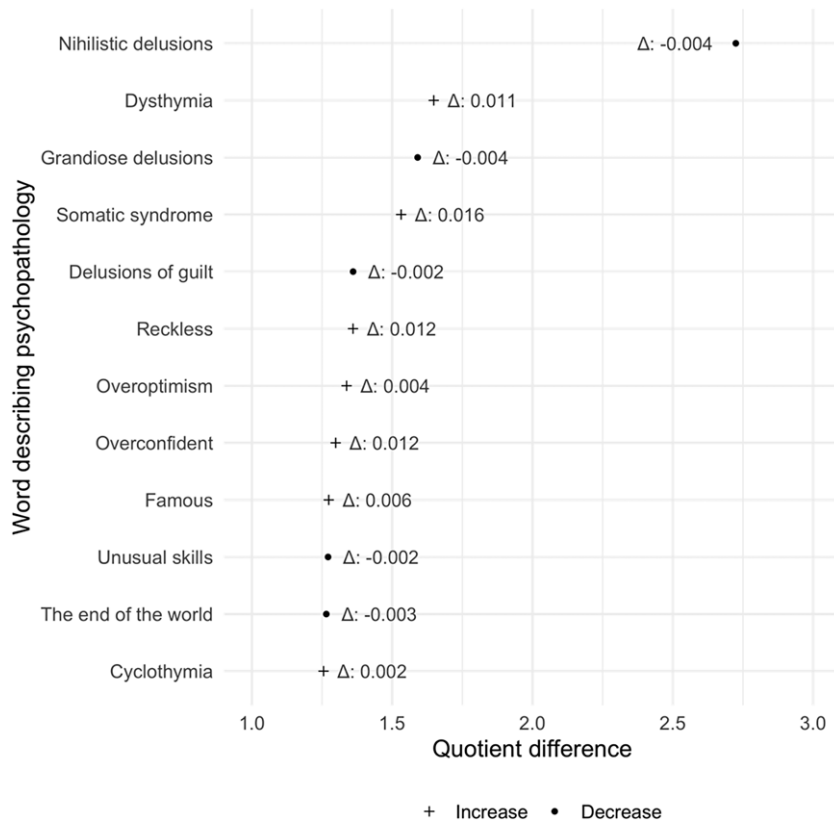


Figure 5. The 12 words describing psychopathology with the largest relative change at the 2020 Q2 changepoint for F3 – mood disorders. The x-axis shows the quotient difference between the mean of the two previous time points (novelty window) to the time point of interest (1 = no difference). The shape indicates whether there was an increase or decrease in word use. The text next to the label denotes the actual difference in means, that is, a change of 0.012 means that the use of a word rose to comprise an additional 0.012% of all words in the F3 – mood disorder category (the 68 words in the F3 – mood disorder category are available in Supplementary Table 3) in 2020 Q2 compared to the previous two quarters.

the note type made up. ‘Conclusion’ had the most significant change by practically falling out of use after 2020 Q3, going from comprising a mean of 2.7% of notes from 2013 to 2020 Q2 to only comprising a mean of 0.3% afterwards. Therefore, the changes in ARI and dependency distance in 2020 Q4 and 2021 Q1, respectively, are likely to be artefacts driven by a very low number of examples (see the discussion for further explanation). The proportion of notes stemming from ‘Phone consultations’ rose from comprising 2.5% of notes on average in 2014–2019 Q2 (pre-COVID-19 pandemic) to 3.9% afterwards (during the COVID-19 pandemic). The proportion of ‘Telephone notes’ also rose during the COVID-19 pandemic. ‘Conversation with treatment aim’ saw a drop in usage in 2019, from comprising approximately 6.5% of notes to 5%. The change in ‘Current mental state’ seems to reflect a minor deviation from a linearly decreasing trend after 2018. ‘Objective, somatic’ saw a halving in usage from comprising a mean of 1.2% of notes before 2014 Q4 to a mean of 0.5% afterwards.

The changepoints in the ARI (4 out of 14) include a slight increase in trend for ARI for ‘Appointments, Psychiatry’ in 2020, and a slight outlier in 2014 Q2 for ‘Patient note’, with a mean ARI of 9.5 compared to a mean of 8.5 in the previous year. The changepoint in ARI for ‘Conclusion’ in 2020 Q3 is due to reduced usage of the ‘Conclusion’ note type. All dependency distance-related changepoints (2 out of 14) seem to be related to reductions in the use of the note type. As with ARI, the 2020 Q4 ‘Conclusion’ changepoint is an artefact of reduced usage. The 2021 Q1 changepoint in ‘Medicine’ appears to reflect a degree of variability over time rather than a salient changepoint. Only a single changepoint (1 out of 14) was identified in the mean number of tokens per note in the ‘Conclusion’ note type in 2013 Q2. Visual

inspection of Supplementary Figure 3 revealed the changepoint to be a minor deviation from a linearly increasing trend.

Sensitivity analyses

Overall, the sensitivity analyses yielded results highly similar to those from the main analyses reported above. Using larger window sizes for calculating novelty led to an estimated changepoint for aggregated notes in 2017 Q3, but values of novelty were highly similar (see Supplementary Figures 7 and 8). The sensitivity analysis of sentence length, readability metrics and syntactic complexity of clinical notes found the extracted metrics to be extremely similar to those from the main analysis (see Supplementary Figures 9–13). Additional changepoints for mean dependency distance in the semi-structured diagnostic interview note type in 2015 Q2 and Q4, and for Appointment, Psychiatry in 2020 Q4 were identified by using the DaCy large model for dependency parsing. The location of estimated changepoints were, however, highly similar to those identified by the main analyses, and deviations seem to mainly stem from the sensitivity of the PELT algorithm rather than major changes in the actual metrics. See the supplementary material for further discussion of the results from the sensitivity analyses.

Discussion

This study investigated the lexical stability of clinical notes in EHR from a large cohort of patients attending psychiatric services in Denmark. Analyses of changepoints in (i) the use of words describing psychopathology – as well as in (ii) note length, readability, syntactic complexity and usage of the clinical notes

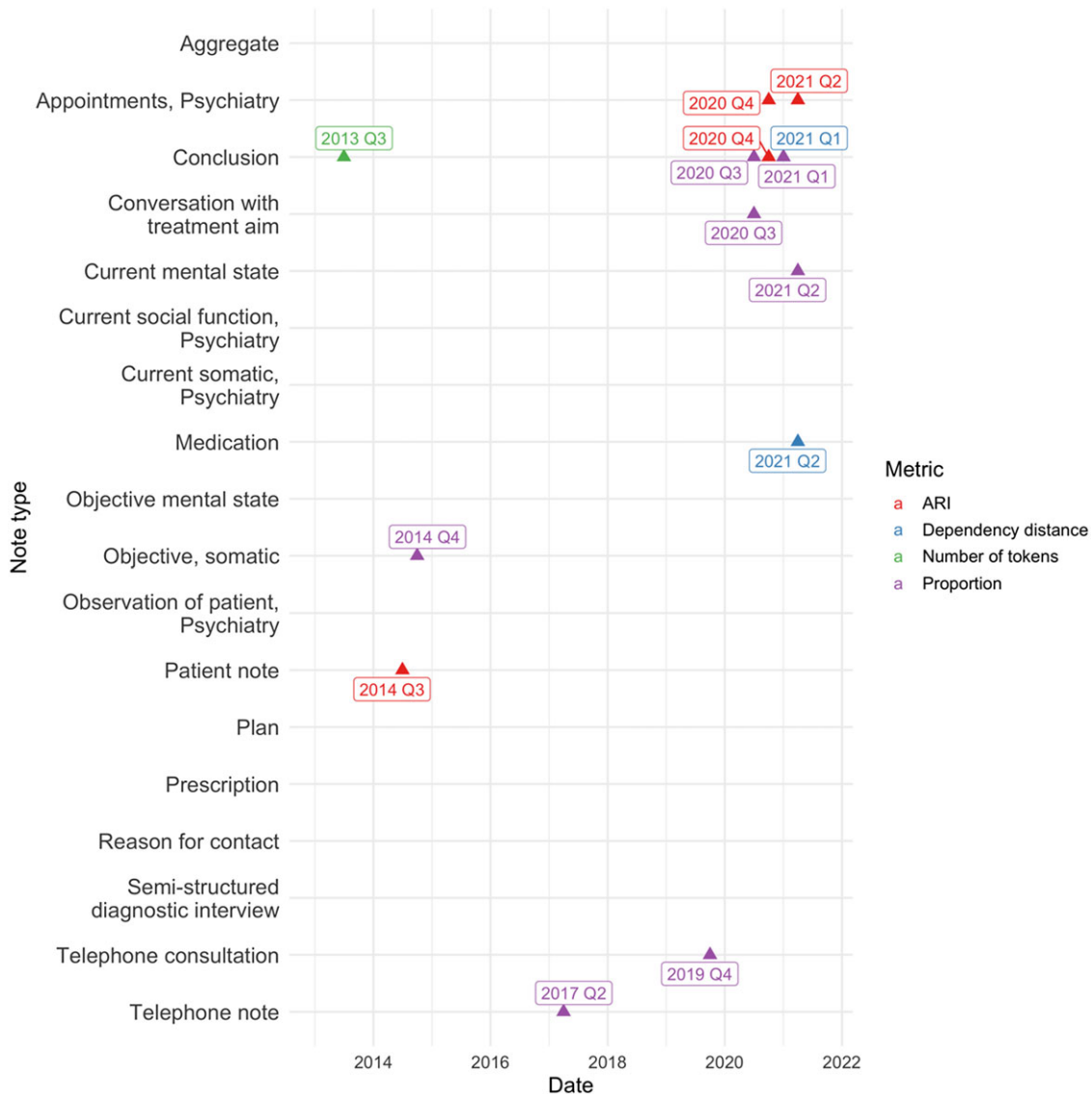


Figure 6. Location of the 14 changepoints in the mean number of tokens, dependency distance, automated readability index (ARI) and proportion of total notes for each note type. The colours indicate which metric the changepoint occurred in.

found the EHR content to generally be stable over time. Notably, no changepoints were found on the aggregate level, that is, when analysing all notes and all words describing psychopathology together. Out of 2988 data points, 17 (0.6%) possible changepoints were identified. We found most changepoints to be related to the discontinuation of a specific note type ('Conclusion') or from increases in the use of virtual communication during the COVID-19 pandemic. Changepoints for words describing psychopathology were only found in the ICD-10 F3 – mood disorders category and were likely caused by a change in the patient presentation during COVID-19. A slow distribution shift was, however, observed in the note length of some specific note types ('Current mental state', 'Current social, Psychiatry', 'Current somatic, Psychiatry', 'Patient note', 'Medication', 'Plan' and 'Semi-structured diagnostic interview'), which was also apparent on the aggregate level. These changes were not detected by changepoint analyses yet might gradually degrade the performance of prediction models over time. This underlines the importance of thoroughly evaluating the performance of clinical prediction models over time and

continuously monitoring them after deployment to ensure model quality.

The most prominent change was found in the use of the 'Conclusion' note type, which dropped from comprising a mean of 2.7% of notes before 2020 Q3 to only comprise a mean of 0.3% afterwards. Conversations with officials from the Business Intelligence Office of the Central Denmark Region, which administers the EHR, revealed the large drop in usage of 'Conclusion' to be caused by the introduction of a new note type in December 2020, namely 'Evaluation/conclusion', which was intended to replace 'Conclusion'. This change was not otherwise documented and implies that one should be cautious if using 'Conclusion' as a point of analysis and highlights the importance of evaluating the stability of EHR data records before use for research purposes.

Furthermore, the analysis revealed a rise in the proportion of 'Phone consultations' and 'Telephone notes' during the COVID-19 pandemic, which is compatible with the regional policy on replacing physical meetings with patients with telephone or video

calls to avoid the spread of the coronavirus. In fact, the build-up of changepoints around the year 2020 is likely influenced by the COVID-19 pandemic. Previous research from the Central Denmark Region has identified changes to language use in the clinical notes during the pandemic (Enevoldsen *et al.*, 2022), exacerbation of psychopathology amongst patients (Jefsen *et al.*, 2020; Kølbaek *et al.*, 2022) and large fluctuations in the number of referrals to the psychiatric services (Kølbaek *et al.*, 2021), which has likely led to a different patient mix during the pandemic. These factors are likely contributors to the relative instability from 2020 to 2021.

The analysis of the use of words describing psychopathology suggests that there are changes in the use of terms related to mood disorders in 2020 Q2, after which novelty immediately dropped. Due to the timing of this changepoint, it may be caused by the COVID-19 pandemic – as a consequence of either the logistic changes described in the paragraph above (e.g. physical meetings replaced by phone or video consultation) or of changes in psychopathology due to the COVID-19 pandemic (Kølbaek *et al.*, 2021, 2022). However, as described in the Methods section, the calculation of novelty is sensitive to relative changes in the use of words describing psychopathology rather than absolute changes. Therefore, due to the rarity of the words driving the change in novelty for F3 – mood disorders, a small increase in the number of patients with these symptoms can lead to relatively large changes in novelty. These changes in novelty are likely of minor importance, supported by the fact that the overall novelty in the present dataset is low compared to other sources, such as literature or Danish news (Jing *et al.*, 2019; Baglini *et al.*, 2021). For instance, for the group of words describing psychotic and mood disorders, the mean novelty was around 0.001, whereas the novelty of literature or Danish news rarely drops below 0.2. Although different representations were used to calculate novelty (topic model for news and word frequencies in the present case), we argue that the novelty in the use of words describing psychopathology is particularly low.

Limitations

There are several limitations to this study. First, using frequencies of words describing psychopathology as the basis for novelty calculation will not identify the emergence of new terms or find shifts in writing style. Novelty can be calculated on any distribution and is often based on topic modelling (e.g. Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003)), which might provide a higher-level representation of the data. Because LDA and other latent variable models are known to perform badly on short texts (Wu *et al.*, 2020), we chose to use the relative words usage describing psychopathology over time to calculate novelty, as this is invariant to the length of the clinical notes. Another option for short texts is to use contemporary neural embedding techniques such as transformer-based topic modelling (Grootendorst, 2022), but these are not feasible to use without GPU access, which was not available to us on the server of the Central Denmark Region at the time of the study. Second, novelty is based on relative changes, not absolute changes. This means that our analysis is sensitive to changes in rare psychopathological keywords as they are likely to fluctuate more than commonly used words. This might increase the risk of finding artefacts due to changes in the patient population. Third, to identify changes in writing style and writing conventions, one could use methods for out-of-distribution detection such as perplexity of the texts over time (Arora *et al.*, 2021). This, however, requires the use of a GPU as well as a

generative language model, neither of which were available to us. Fourth, changepoint detection models are not sensitive to slow distribution shifts and drifts but rather to changes in means or variance that persist over some time. As a consequence, our method will identify major breaks, such as the discontinuation of the 'Conclusion' note type but will not find distribution shifts, such as the increase in note length, without manual analysis and inspection. Fifth, there are several different formulas for calculating readability. We chose to use ARI as it is simple to implement, coupled with the fact that it had the highest mean correlation (0.93) to the other readability metrics available in the *textdescriptives* python package (Hansen *et al.*, 2023). Given the high correlation, it is reasonable to expect other readability metrics to produce very similar results. Sixth, the sensitivity analysis revealed that the PELT model with default parameters might be too sensitive, that is, it has a false-positive rate that is too high. Future studies would be advised to adjust the penalty to avoid finding spurious changepoints. Lastly, the degree to which these changepoints will impact other tasks such as prediction modelling remains to be tested. Future studies on the PSYCOP cohort will investigate this matter.

In conclusion, in a large body of clinical notes from psychiatric services spanning almost a decade, the mean length of notes, readability, syntactic complexity, note type distribution and usage of words describing psychopathology were generally stable over time. The discontinuation of the 'Conclusion' note type was the cause of most of the changepoints detected. There were changes in novelty in some groups of words describing psychopathology, but the overall novelty was relatively low, suggesting consistency in the content of the clinical notes over time. Taken together, these findings suggest that prediction models can be trained on the content of clinical notes without fitting to idiosyncrasies of a specific time period. However, one should be cautious of gradual changes in the data distribution, such as an increase in the average note length. To ensure reliable results of future studies based on text from clinical notes, we encourage that checks of stability similar to those presented here are conducted ahead of analysis.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/neu.2023.46>

Data availability statement. In accordance with Danish law for the protection of privacy, the data used for this study are only available for research projects conducted by employees in the Central Denmark Region upon approval from the Legal Office under the Central Denmark Region (in accordance with the Danish Health Care Act §46, Section 2).

Acknowledgements. The authors thank Bettina Nørremark from Aarhus University Hospital – Psychiatry for her assistance with data extraction.

Author contribution. Conception and design: all authors. Funding obtainment: S.D.Ø. Provision of study data: S.D.Ø. and A.A.D. Data analysis: L.H. and K.C.E. Interpretation: All authors. Manuscript writing: L.H., E.P., M.B. and S.D.Ø. Revision of manuscript for important intellectual content: K.C.E., A.A.D. and K.L.N. Final approval of the manuscript: all authors.

Financial support. The study is supported by grants from the Lundbeck Foundation (grant number: R344-2020-1073), the Danish Cancer Society (grant number: R283-A16461), the Central Denmark Region Fund for Strengthening of Health Science (grant number: 1-36-72-4-20) and the Danish Agency for Digitisation Investment Fund for New Technologies (grant number 2020-6720) to Østergaard, who reports further funding from the Lundbeck Foundation (grant number: R358-2020-2341), the Novo Nordisk Foundation (grant number: NNF20SA0062874) and Independent Research Fund Denmark (grant number: 7016-00048B). The funders played no role in the study design,

collection, analysis or interpretation of data, the writing of the report or the decision to submit the paper for publication.

Competing interests. Danielsen has received a speaker honorarium from Otsuka Pharmaceutical. SDØ received the 2020 Lundbeck Foundation Young Investigator Prize. Furthermore, SDØ owns/has owned units of mutual funds with stock tickers DKIGI, IAIMWC and WEKAFKI and has owned units of exchange-traded funds with stock tickers BATE, TRET, QDV5, QDVH, QDVE, SADM, IQQH, USPY, EXH2, 2B76 and EUNL. The remaining authors declare no conflicts of interest.

Ethics. This study was carried out to ensure the validity/stability of the data used for studies based on the PSYCOP cohort (Hansen *et al.*, 2021). The use of EHRs from the Central Denmark Region was approved by the Central Denmark Region Legal Office per the Danish Health Care Act §46, Section 2. We submitted the study to the Central Denmark Region ethical review board, which waived informed consent (waiver for this project: 1-10-72-1-22). All data were processed and stored in accordance with the European Union General Data Protection Regulation, and the project is registered on the internal list of research projects having the Central Denmark Region as the data steward. All experiments were performed in accordance with regulations.

References

- Abbe A, Grouin C, Zweigenbaum P and Falissard B (2016) Text mining applications in psychiatry: a systematic literature review. *International Journal of Methods in Psychiatric Research* 25(2), 86–100.
- Arora U, Huang W and He H (2021) Types of out-of-distribution texts and how to detect them. *arXiv:2109.06827 [cs]*. Available at <http://arxiv.org/abs/2109.06827> (accessed 3 November 2021).
- Baglini RB, Nielbo KL, Hæstrup F, Enevoldsen K, Vahlstrup PB and Roepstorff A (2021) When no news is bad news: detection of negative events from news media content. In Digital Humanities Benelux, Leiden, Belgium. Available at <https://2021.dhbenelux.org/>
- Bernstorff M, Hansen L, Perfalk E, Danielsen AA and Østergaard SD (2022) Stability of diagnostic coding of psychiatric outpatient visits across the transition from the second to the third version of the Danish National Patient Registry. *Acta Psychiatrica Scandinavica* 146(3), 272–283.
- Blei DM, Ng AY and Edu JB (2003) Latent Dirichlet Allocation. Michael I. Jordan. Available at <http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- Boag W, Doss D, Naumann T and Szolovits P (2018) What's in a note? Unpacking predictive value in clinical note representations. *AMIA Summits on Translational Science Proceedings* 26, 2018–2034.
- Boag W, Kovaleva O, McCoy TH, Rumshisky A, Szolovits P and Perlis RH (2021) Hard for humans, hard for machines: predicting readmission after psychiatric hospitalization using narrative notes. *Translational Psychiatry* 11(1), 1–6.
- Cimino JJ (2013) Improving the electronic health record: getting what we wished for. *JAMA: The Journal of the American Medical Association* 309(10), 991. DOI: 10.1001/jama.2013.890.
- Danielsen AA, Fenger MHJ, Østergaard SD, Nielbo KL and Mors O (2019) Predicting mechanical restraint of psychiatric inpatients by applying machine learning on electronic health data. *Acta Psychiatrica Scandinavica* 140(2), 147–157.
- Devlin J, Chang M-W, Lee K and Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*. Available at <http://arxiv.org/abs/1810.04805> (accessed 16 Sep 2020).
- Enevoldsen K, Hansen L and Nielbo KL (2021) DaCy: a unified framework for Danish NLP. In: CEUR Workshop Proceedings, pp. 206–216.
- Enevoldsen KC, Danielsen AA, Rohde C, Jepsen OH, Nielbo KL and Østergaard SD (2022) Monitoring of COVID-19 pandemic-related psychopathology using machine learning. *Acta Neuropsychiatrica* 34(3), 148–152.
- Goldstein BA, Navar AM, Pencina MJ and Ioannidis JPA (2017) Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* 24(1), 198–208.
- Grootendorst M (2022) BERTopic: neural topic modeling with a class-based TF-IDF procedure. Available at <http://arxiv.org/abs/2203.05794> (accessed 16 Aug 2022).
- Hansen L, Enevoldsen KC, Bernstorff M, Nielbo KL, Danielsen AA and Østergaard SD (2021) The PSYchiatric clinical outcome prediction (PSYCOP) cohort: leveraging the potential of electronic health records in the treatment of mental disorders. *Acta Neuropsychiatrica* 33, 323–330.
- Hansen L, Olsen LR, Bernstorff M, Enevoldsen K and Hæstrup F (2022) HLaSse/TextDescriptives: Version 1.0.6. Available at <https://zenodo.org/record/6326976> (accessed 4 Mar 2022).
- Hansen L, Olsen LR and Enevoldsen K (2023) TextDescriptives: a Python package for calculating a large variety of metrics from text. *Journal of Open Source Software* 8(84), 5153.
- Häyriäinen K, Saranto K and Nykänen P (2008) Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International Journal of Medical Informatics* 77(5), 291–304.
- Honnibal M, Montani I, Van Landeghem S and Boyd A (2020) spaCy: Industrial-strength Natural Language Processing in Python. DOI: 10.5281/zenodo.1212303.
- Huang K, Altsaar J and Ranganath R (2019) ClinicalBERT: modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342 [cs]*. Available at <http://arxiv.org/abs/1904.05342> (accessed 17 Sep 2020).
- Jepsen OH, Rohde C, Nørremark B and Østergaard SD (2020) COVID-19-related self-harm and suicidality among individuals with mental disorders. *Acta Psychiatrica Scandinavica* 142(2), 152–153.
- Jing E, DeDeo S and Ahn Y-Y (2019) Sameness Attracts, Novelty Disturbs, but Outliers Flourish in Fanfiction Online. Available at <http://arxiv.org/abs/1904.07741> (accessed 16 Aug 2022).
- Killick R and Eckley IA (2014) Changepoint: an R package for changepoint analysis. *Journal of Statistical Software* 58(3), 1–19.
- Killick R, Fearnhead P and Eckley IA (2012) Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* 107(500), 1590–1598.
- Kincaid JP, Fishburne Jr. RP, Rogers RL and Chissom BS (1975) Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Naval Technical Training Command Millington TN Research Branch.
- Kølbæk P, Jepsen OH, Speed M and Østergaard SD (2022) Mental health of patients with mental illness during the COVID-19 pandemic lockdown: a questionnaire-based survey weighted for attrition. *Nordic Journal of Psychiatry* 76(5), 338–347.
- Kølbæk P, Nørremark B and Østergaard SD (2021) Forty percent reduction in referrals to psychiatric services during the COVID-19 pandemic. *Psychotherapy and Psychosomatics* 90(1), 67–68.
- Marcus G (2018) Deep Learning: A Critical Appraisal. Available at <http://arxiv.org/abs/1801.00631> (accessed 16 Aug 2022).
- Montani I, Honnibal M, Honnibal B, Landeghem SV, Boyd A, Peters H, Samsonov M, McCann PO, Geovedi J, Regan J, Orosz G, Altinok D, Kristiansen SL, Roman, Fiedler L, Bot E, Howard G, Phatthiyaphaibun W, Tamura Y, Bozek S, Amery M, Böing B, Tippa PK, Vogelsang LU, Balakrishnan R, Mazaev V, GregDubbin, jeannefukumaru and Henry W (2021) explosion/spaCy: 3.1.1: support for ancient Greek and various bug fixes. Available at <https://zenodo.org/record/5115698> (accessed 4 Mar 2022).
- Otter DW, Medina JR and Kalita JK (2020) A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* 32(2), 604–624.
- Pedersen K and Bertelsen A (2018) Present State Examination: Kort version til klinisk brug, 5th edn. København: Munksgaard.
- R Core Team (2019) R: A Language and Environment for Statistical Computing [online]. Vienna, Austria: R Foundation for Statistical Computing. Available at <https://www.R-project.org/>
- Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, Sundberg P, Yee H, Zhang K, Zhang Y, Flores G, Duggan GE, Irvine J, Le Q, Litsch K, Mossin A, Tansuwan J, Wang D,

- Wexler J, Wilson J, Ludwig D, Volchenboum SL, Chou K, Pearson M, Madabushi S, Shah NH, Butte AJ, Howell MD, Cui C, Corrado GS and Dean J (2018) Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1), 1–10.
- Rosenbloom ST, Stead WW, Denny JC, Giuse D, Lorenzi NM, Brown SH and Johnson KB (2010) Generating clinical notes for electronic health record systems. *Applied Clinical Informatics* 1(3), 232–243.
- Shen Z, Liu J, He Y, Zhang X, Xu R, Yu H and Cui P (2021) Towards Out-Of-Distribution Generalization: A Survey. Available at <https://arxiv.org/abs/2108.13624v1> (accessed 24 Nov 2021).
- Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN and Kroeker KI (2020) An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine* 3(1), 1–10.
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L and Polosukhin I (2017) Attention is all you need. *arXiv:1706.03762 [cs]*. Available at <http://arxiv.org/abs/1706.03762> (accessed 17 Sep 2020).
- Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S and Liu H (2018) Clinical information extraction applications: a literature review. *Journal of Biomedical Informatics* 77, 34–49.
- Webster JJ and Kit C (1992) Tokenization as the initial phase in NLP. In COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics.
- Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L and Hester J (2019) Welcome to the Tidyverse. *Journal of Open Source Software* 4(43), 1686.
- World Health Organization (1993) The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research [online]. World Health Organization. Available at <https://apps.who.int/iris/handle/10665/37108> (accessed 20 May 2021).
- Wu X, Li C, Zhu Y and Miao Y (2020) Short text topic modeling with topic distribution quantization and negative sampling decoder. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1772–1782.