Psychological Medicine

cambridge.org/psm

Original Article

Predicting involuntary admission following inpatient psychiatric treatment using machine learning trained on electronic health record data

Erik Perfalk[1,2], Jakob Grøhn Damgaard[1,2], Martin Bernstorff[1,2], Lasse Hansen[1,2], Andreas Aalkjær Danielsen[1,2] and Søren Dinesen Østergaard[1,2]

[1]Department of Affective Disorders, Aarhus University Hospital – Psychiatry, Aarhus, Denmark and [2]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

Abstract

Background. Involuntary admissions to psychiatric hospitals are on the rise. If patients at elevated risk of involuntary admission could be identified, prevention may be possible. Our aim was to develop and validate a prediction model for involuntary admission of patients receiving care within a psychiatric service system using machine learning trained on routine clinical data from electronic health records (EHRs).

Methods. EHR data from all adult patients who had been in contact with the Psychiatric Services of the Central Denmark Region between 2013 and 2021 were retrieved. We derived 694 patient predictors (covering e.g. diagnoses, medication, and coercive measures) and 1134 predictors from free text using term frequency-inverse document frequency and sentence transformers. At every voluntary inpatient discharge (prediction time), without an involuntary admission in the 2 years prior, we predicted involuntary admission 180 days ahead. XGBoost and elastic net models were trained on 85% of the dataset. The models with the highest area under the receiver operating characteristic curve (AUROC) were tested on the remaining 15% of the data.

Results. The model was trained on 50 634 voluntary inpatient discharges among 17 968 patients. The cohort comprised of 1672 voluntary inpatient discharges followed by an involuntary admission. The best XGBoost and elastic net model from the training phase obtained an AUROC of 0.84 and 0.83, respectively, in the test phase.

Conclusion. A machine learning model using routine clinical EHR data can accurately predict involuntary admission. If implemented as a clinical decision support tool, this model may guide interventions aimed at reducing the risk of involuntary admission.

Introduction

The incidence of involuntary admissions is on the rise worldwide (Sheridan Rains et al., 2019). Involuntary admissions are used when patients are in urgent need of psychiatric inpatient treatment, but are too ill (typically psychotic) to consent (Salize & Dressing, 2004). Involuntary admissions can be traumatic for patients and are costly for society (Katsakou & Priebe, 2007). Therefore, various interventions to reduce the need for involuntary admissions have been investigated (de Jong et al., 2016), but none are currently systematically applied in the Central Denmark Region. To ensure cost-effectiveness of implementation, these interventions should preferably target patients at high risk of involuntary admission. However, such individual risk assessments are complex.

Several risk factors for involuntary admission have been identified in large patient populations, e.g. prior involuntary admission, and psychotic or bipolar disorders (Walker et al., 2019). However, assessing risk at the level of the individual patient is challenging due to potential interactions between risk factors, waxing and waning of risk factors, and irregular/noisy clinical data on risk factors (Bzdok & Meyer-Lindenberg, 2018). This resonates well with the complexity of patient-level risk assessment in clinical practice. Recently, however, machine learning methods have been demonstrated to handle this level of complexity well in some cases – with notable exceptions (Christodoulou et al., 2019). Unlike standard statistical analyses, machine learning inherently accommodates complex interactions and idiosyncrasies and also handles large amounts of predictors (Cerqueira, Torgo, & Soares, 2019; Song, Mitnitski, Cox, & Rockwood, 2004).

We are aware of two prior machine learning studies having examined involuntary admission via routine clinical data (Karasch, Schmitz-Buhl, Mennicken, Zielasek, &

Cite this article: Perfalk E, Damgaard JG, Bernstorff M, Hansen L, Danielsen AA, Østergaard SD (2024). Predicting involuntary admission following inpatient psychiatric treatment using machine learning trained on electronic health record data. *Psychological Medicine* **54**, 4348–4361. https://doi.org/10.1017/S0033291724002642

Received: 18 June 2024
Revised: 24 September 2024
Accepted: 30 September 2024
First published online: 18 November 2024

Keywords:
artificial intelligence; involuntary admission; machine learning; prediction; psychiatry

Abbreviations:
AUROC: area under the receiver operating characteristic curve; CI: confidence interval; ECT: electroconvulsive therapy; EHR: electronic health record; ICD-10: International Classification of Disease, Tenth Revision; NPV: negative predictive value; PPR: predicted positive rate; PPV: positive predictive value; TF-IDF: term frequency-inverse document frequency

Corresponding author:
Erik Perfalk;
Email: erperf@rm.dk

Gouzoulis-Mayfrank, 2020; Silva, Gholam, Golay, Bonsack, & Morandi, 2021). Both, however, fail to construct a relevant prediction task as they do not issue predictions, which is a prerequisite for clinical relevance, but merely utilize machine learning methods for identification of risk factors for involuntary admission. Additionally, both studies only consider patients with complete data in their primary analysis, which could potentially decrease the generalizability as data from real-world practice are typically not missing at random (Bzdok & Meyer-Lindenberg, 2018). We have previously shown that a machine learning model trained on routine clinical data from electronic health records (EHRs) can accurately predict mechanical restraint (Danielsen, Fenger, Østergaard, Nielbo, & Mors, 2019) and are currently in the process of implementing a decision support (risk reduction) tool based on this model in clinical practice. To our knowledge, no studies have used machine learning to predict involuntary admissions at the level of the individual patient using EHR data. Therefore, the aim of this study was to fill this gap in the literature.

Methods

An illustration of the methods used in this study is shown in Fig. 1.

Reporting guidelines

This study adhered to the reporting guidelines of TRIPOD+AI (Collins et al., 2024) and the TRIPOD+AI checklist is available in the online Supplementary materials (Table S7).

Data source

The study is based on data from the PSYchiatric Clinical Outcome Prediction cohort, encompassing routine clinical EHR data from all individuals with at least one contact to the Psychiatric Services of the Central Denmark Region in the period from January 1, 2011 to November 22, 2021 (Hansen et al., 2021). The Central Denmark Region is one of the five Danish Regions and has a catchment area of approximately 1.3 million people. The dataset includes records from all service contacts to the public hospitals in the Central Denmark Region (both psychiatric and general hospitals). A service contact can be either an inpatient admission, outpatient visit, home visit, or consultation by phone, and each is labeled with a timestamp and diagnosis. Due to the universal healthcare system in Denmark, the large majority of hospital contacts are to public hospitals (there are no private psychiatric hospitals in Denmark) and, thus, covered by these data. Importantly, the dataset also includes blood samples from general practitioners as they are analyzed at public hospitals and, as a result, are included in this dataset (Bernstorff et al., 2024).

Data extraction

All EHR data from patients with at least one contact with the Psychiatric Services of the Central Denmark Region in the period from 2013 to 2021 were extracted (Fig. 1a). To ensure the feasibility of subsequent implementation of a predictive machine learning model potentially developed in this study, only data collected routinely as part of standard clinical practice and recorded in the EHR system were used (i.e. there was no data collection for the purpose of this study) (Hansen et al., 2021).

Cohort definition

Figure 1b illustrates the cohort definition. The cohort consisted of all adult patients with at least one contact to the Psychiatric Services of the Central Denmark Region in the time period from 2011 to 2021. Data prior to 2013 were dropped due to data instability, primarily due to the gradual implementation of a new EHR system in 2011 (Bernstorff, Hansen, Perfalk, Danielsen, & Østergaard, 2022; Hansen et al., 2023). However, data on involuntary admissions from 2012 were used to establish incidence of involuntary admissions since these data were registered via an alternative digital system and, therefore, unaffected by the implementation of the new EHR system (*Sundhedsdatastyrelsen:Register over Anvendelse af Tvang i Psykiatrien, 2024*).

Dataset splitting

The data were randomly split into a training (85%) and a test (15%) set by sampling unique patients, stratified by whether they had an involuntary admission within the follow-up (see Fig. 1c). This ensured a balanced proportion of patients with involuntary admission in the training and test sets and prevented leakage of learnt subject-specific patterns (due to repeated observations) to the test set. The test set was not examined until the final stage of model evaluation, where no additional changes were made to the model.

Prediction times and exclusion criteria

Prediction times were defined as the last day of a voluntary psychiatric admission. A prediction at this timepoint would enable outpatient clinics to initiate targeted intervention/monitoring to reduce the risk of involuntary admission. Additionally, an exclusion criterion stipulating that patients should not have had an involuntary admission in the 2 years prior the prediction time was implemented. This prevented predictions in cases where clinicians were already aware of the patient's risk of involuntary admission, thus proactively reducing the risk of alert fatigue. Additionally, if a prediction time did not have a long enough lookbehind- (for predictors) or lookahead window (for outcomes), that prediction time was dropped (Fig. 1d). For definition of lookbehind- and lookahead windows, see the following two sections.

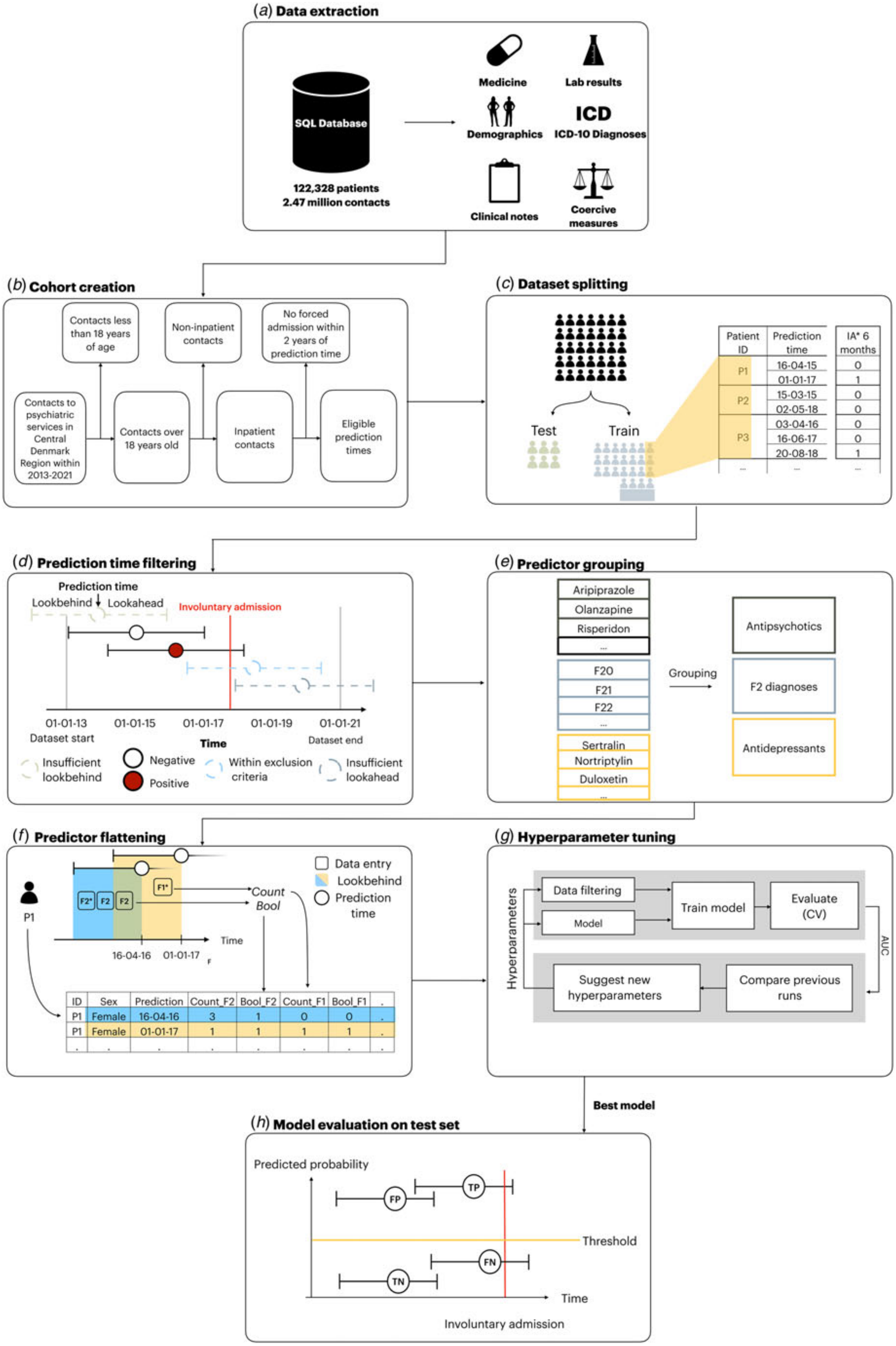
Outcome definition and lookahead window

The outcome was defined as the start of an involuntary admission. The lookahead window (the period following the prediction time in which the outcome could occur) was 180 days. Hence, all prediction times for which an involuntary admission occurred within 180 days were deemed to be positive outcomes (Fig. 1d).

Predictor engineering and lookbehind window

A full list of the predictors (a total of 1828) and their definitions is available in online Supplementary Table S1. The predictors were chosen based on the literature on risk factors for involuntary admissions (Walker et al., 2019) supplemented with clinical domain knowledge. Predictors were engineered by aggregating the values for the variable of interest within a specified lookbehind window (10, 30, 180, and 365 days leading up to a prediction time) using different predictor aggregation functions (mean, max,

https://doi.org/10.1017/S0033291724002642 Published online by Cambridge University Press



Patient ID	Prediction time	IA* 6 months
P1	16-04-15	0
	01-01-17	1
P2	15-03-15	0
	02-05-18	0
P3	03-04-16	0
	16-06-17	0
	20-08-18	1
...	...	...

ID	Sex	Prediction	Count_F2	Bool_F2	Count_F1	Bool_F1	.
P1	Female	16-04-16	3	1	0	0	.
P1	Female	01-01-17	1	1	1	1	.
.	.	.	.	.	.	.	.

Figure 1. Extraction of data and outcome, dataset splitting, prediction time filtering, specification of predictors and flattening, model training, testing, and evaluation. This figure was modified to this project based on Bernstorff et al. (2024). IA, involuntary admissions; F1 and F2, ICD-10 diagnoses within the group of diagnoses included in F1 and F2 chapters; CV, cross-validation; TP, true positive; FP, false positive; TN, true negative; FN, false negative.

https://doi.org/10.1017/S0033291724002642 Published online by Cambridge University Press

bool, etc.). The specific aggregation methods for each variable can be found in online Supplementary Table S1. This processing was performed using the timeseriesflattener v2.0.1 package (Fig. 1f) (Bernstorff, Enevoldsen, Damgaard, Danielsen, & Hansen, 2023). If a predictor was not present in the lookbehind period from a prediction time, it was labeled as 'missing'. However, these instances do not indicate missing values in the conventional sense, as they stem from a genuine lack of data, rather than, e.g. a missed visit in a clinical trial. This absence reflects real-world clinical practice, and, therefore, patients with such missing data should not be excluded, as it aligns with the available data for potential implementation.

The predictors can be grouped into nine strata: age and sex, hospital contacts, psychiatric diagnoses, medications, lab results, coercive measures, psychometric rating scales, suicide risk assessment, and free-text predictors from EHR clinical notes (extracted via natural language processing). Specifically, hospital contacts included both inpatient and outpatient contacts with linked diagnoses. Diagnoses included all psychiatric subchapters (F0–F9) from the International Classification of Disease, Tenth Revision (ICD-10) (World Health Organization, n.d.-b) with specific predictors for schizophrenia (F20), bipolar disorders (F30–F31), and cluster b-personality disorders (F60.2–F60.4) (dissocial-, borderline-, and histrionic personality disorder). Medication predictors were based on structured anatomical therapeutic chemical classification system codes (World Health Organization, n.d.-a) and grouped as follows (Fig. 1e): antipsychotics, first-generation antipsychotics, second-generation antipsychotics, depot antipsychotics, antidepressants, anxiolytics, hypnotics/sedatives, stimulants, analgesics, and drugs for alcohol abstinence/opioid dependence. Finally, lithium, clozapine, and olanzapine were included as individual predictors. Predictors based on laboratory tests included plasma levels of antipsychotics, antidepressants, paracetamol, and ethanol. Coercive measures included involuntary medication, manual restraint, chemical restraint, and mechanical (belt) restraint. Scores from psychometric rating scales included the Brøset violence checklist (Woods & Almvik, 2002), the 17-item Hamilton depression rating scale (HAM-D17) (Hamilton, 1960) and a simplified version of the Bech Rafaelsen mania rating scale (MAS-M) (Bech, Rafaelsen, Kramp, & Bolwig, 1978). Data on suicide risk assessment were based on a scoring system used in the Central Denmark Region with the following risk levels: 1 (no increased risk), 2 (increased risk), and 3 (acutely increased risk).

Predictors from free text stemmed from the subset of EHR clinical note types deemed to be most informative and stable over time, e.g. 'Subjective Mental State' and 'Current Objective Mental State' (for the full list of clinical note types, see online Supplementary Table S2) (Bernstorff et al., 2022). Two different algorithms were applied to create predictors from the free text: term frequency-inverse document frequency (TF-IDF) (Pedregosa et al., 2011) and sentence transformers (Reimers & Gurevych, 2019). For the TF-IDF model, the unstructured free text was first preprocessed by lower-casing all words and removing stop words and symbols. Subsequently, the model generated all uni- and bi-grams. Second, top 10% by document frequency were removed (due to assumed low predictive value). Lastly, the top 750 uni- or bi-grams were included in the model. For each patient, all clinical notes within the 180 days lookbehind prior to a prediction time were concatenated into a single document from which the TF-IDF predictors were constructed. A pre-trained multilingual sentence transformer model (Reimers & Gurevych, 2019) was applied to extract sentence embeddings (model: 'paraphrase-multilingual-MiniLM-L12-v2'). This model is bound by a maximum input sequence length of 512 tokens. For each patient, the first 512 tokens from each clinical note within the 180 days lookbehind prior to a prediction time were extracted and input to the model, yielding a contextualized embedding of the text with 384 dimensions. Subsequently, the embeddings from each note within the lookbehind window were averaged to obtain a single aggregated embedding, which was included as a predictor in the model.

Hyperparameter tuning and model training

Two types of machine learning models were trained: XGBoost and elastic net-regularized logistic regression (using Scikit-learn, version 1.2.1) (Pedregosa et al., 2011). XGBoost was chosen because gradient boosting techniques typically excel in predictive accuracy for structured data, offer rapid training, and intrinsically handle missing values (Grinsztajn, Oyallon, & Varoquaux, 2022). Elastic net-regularized logistic regression served as a benchmark model (Desai, Wang, Vaduganathan, Evers, & Schneeweiss, 2020; Nusinovici et al., 2020). A five-fold stratified cross-validation was adopted for training with no patient occurring in more than one-fold. Fine-tuning of hyperparameters (see online Supplementary Table S3 for details) was performed over 300 runs to optimize the area under the receiver operating characteristic curve (AUROC) through the tree-structured parzen estimator method in Optuna v2.10.1.33 (see Fig. 1g) (Lundberg & Lee, 2017). All analyses were performed using Python (version 3.10.9).

Model evaluation on test data

The XGBoost and elastic net model which achieved the highest AUROC following cross-validation on the training set were evaluated on the test set (see Fig. 1h). Apart from the AUROC, we also calculated the sensitivity, specificity, positive predictive values (PPVs), and negative predictive values (NPVs) at predicted positive rates (PPRs) of 1%, 2%, 3%, 4%, 5%, 10%, 20%, and 50%, respectively. The PPR is the proportion of all prediction times that are marked as positive. To test the robustness of the best performing model, its performance was examined across sex, age, months since the first visit, month of year, and day of week strata. Furthermore, a time-to-outcome robustness analysis was conducted to assess how the model behaved at different time-to-outcome thresholds.

Additionally, the calibration of the model was visualized with calibration plots with adjoining distribution plots of the predicted probabilities for the best XGBoost and elastic net model. Clinical usefulness was assessed by decision curve analysis (Vickers & Elkin, 2006). Plots were limited to an upper bound of 0.20 which represents a one-in-five chance of having an involuntary admission within 6 months should nothing change, and it is unlikely that risk thresholds greater would be tolerated. Net benefit is calculated as the additional percentage of cases that could be intervened upon with use of our models with no increase in false-positives.

To address the temporal stability of the best performing models, we performed temporal cross-validation using gradually increasing alternating endpoints for the training set (2016–2020) with validation on the available data from the subsequent remaining years after the training set endpoint (2016–2021). This analysis replicates a case where a trained

https://doi.org/10.1017/S0033291724002642 Published online by Cambridge University Press

model is implemented at a given timepoint, and the performance of the model is evaluated at a later timepoint.

Estimation of predictor importance

To interpret which predictors informed the predictions in the models, we calculated predictor importance metrics. For XGBoost models, predictor importance was estimated via information gain (Chen & Guestrin, 2016). In this case the gain of a predictor is calculated as the average improvement in loss when generalizing to the training data accomplished by the predictor across all node splits in the model that handle the predictor. For elastic net models, predictor importance was analyzed by obtaining the standardized model coefficients (Zou & Hastie, 2005). Standardized coefficients represent the change in log-odds for a one standard deviation increase in the predictor values. Hence, the magnitude of a coefficient specifies the strength of the relationship between the predictor and the outcome while controlling for the other predictors and this measure, therefore, allows for easy comparison of the relative importance of predictors. Importantly, the elastic net model coefficients are directed, meaning they convey whether an increase in the predictor value pushes the model toward a positive or a negative prediction. This is not the case for the information gain estimations for the XGBoost models which only inform about general predictive importance regardless of direction.

Secondary analyses of alternative model designs

As secondary analyses, we performed cross-validated model training using alternative model designs. First, we removed the implemented exclusion criterion of having an involuntary admission in the 2 years preceding a prediction time. Second, we assessed the importance of the number of predictors, by using only subsets of the full predictor set in the model training. Specifically, three distinct predictor sets were considered (all including sex and age): only diagnoses, only patient descriptors (all predictors except for text predictors), and only text predictors. Third, models with lookahead windows of 90 and 365 days, respectively, were trained. The performance metrics of the alternative model designs are derived from the cross-validation on the training set and were not tested on the hold-out test set.

Ethics

The study was approved by the Legal Office of the Central Denmark Region in accordance with the Danish Health Care Act §46, Section 2. The Danish Committee Act exempts studies based only on EHR data from ethical review board assessment (waiver for this project: 1-10-72-1-22). Handling and storage of data complied with the European Union General Data Protection Regulation. The project is registered on the list of research projects having the Central Denmark Region as data steward. There was no patient nor public involvement in this study.

Results

The full dataset consisted of 52 600 voluntary admissions distributed among 19 252 unique patients. A total of 1672 of the voluntary admissions were followed by an involuntary admission within 180 days after discharge (positive outcome), distributed across 806 unique patients (an involuntary admission can be included in multiple positive outcomes as a patient can have multiple voluntary admissions [prediction times] in the 180 days prior to an involuntary admission [positive outcome]).

Table 1 lists clinical and demographic patient data for the prediction times included in the training and evaluation of the main model. The main model included predictors with a lookbehind window of up to 365 days. After filtering away all prediction times where the lookbehind or lookahead windows extended beyond the available data for a patient, a total of 50 364 prediction times remained. These prediction times were distributed across 17 968 unique patients (49.4% females [training set = 49.3% and test set = 50.0%], median age = 40.2 years [training set = 40.5 and test set = 39.2]).

Hyperparameters and model training

The cross-validation on the training set for model tuning showed that XGBoost (full predictor set AUROC = 0.79) outperformed elastic net (full predictor set AUROC = 0.78) across all model variations (see Table 2). The hyperparameters used for the best XGBoost and elastic net models are listed in online Supplementary Table S4.

Model evaluation on test data

After the model selection in the training phase, the best performing XGBoost and elastic net models yielded an AUROC of 0.84 and 0.83, respectively, on the test set (see Figs 2a and 3a).

Table 3 lists the performance metrics from the XGBoost (3A) and elastic net (3B) model on the test set based on different PPRs. At a PPR of 5%, the XGBoost model has a sensitivity of 39% and a PPV of 36%. Thus, approximately two out of five of all true positive outcomes are correctly predicted, and for every three positive predictions, more than one prediction time is truly followed by an involuntary admission within 180 days. At this PPR, 36% of the unique involuntary admissions that underlie the positive outcomes are correctly detected (predicted positive) at least once. In comparison, at a PPR of 5%, the elastic net model has a sensitivity of 36% and a PPV of 33%. At this PPR, 27% of the unique involuntary admissions that underlie the positive outcomes are correctly detected (predicted positive) at least once. Decision curve analysis of the models showed that both yield a universally greater net benefit than competing strategies in a sensible threshold probability range of 0.02–0.20 (see online Supplementary Fig. S5).

Figures 2C and 3C show the sensitivity of the models for prediction times with varying time to the outcome at different PPRs. The sensitivity curves appear to remain stable as the time to outcome increases for both models. The median time from the first-positive prediction to the involuntary admission was 70 days for the XGBoost model and 64 days for the elastic net model.

Online Supplementary Figs S1 and S2 show the performance of the models across different patient characteristics and calendar time subgroups. The models appear robust across all characteristics and the minor fluctuations, such as the variation in performance between the sexes, can likely be attributed to similar minor differences in sample distributions. The calibration curves (see online Supplementary Figs S3 and S4) indicate that both models are sufficiently calibrated with both models, however, slightly undershooting on patients with higher predicted probabilities. The plots are cut-off at predicted probabilities above 20% as there are too few patients with higher probabilities to make stable calibration estimates above this threshold.

Table 1. Descriptive statistics for prediction times

	Overall		Train		Test	
Prediction times, *n*	50 364		43 188		7176	
	Median	**Q1, Q3**	**Median**	**Q1, Q3**	**Median**	**Q1, Q3**
Age	40.2	28.3, 53.5	40.5	28.3, 53.7	39.2	28.0, 52.2
	n	**%**	***n***	**%**	***n***	**%**
Age 18–20 years	998	2.0	861	2.0	137	1.9
Age 20–29 years	12 462	24.7	10 630	24.6	1832	25.5
Age 30–39 years	10 560	21.0	8965	20.8	1595	22.2
Age 40–49 years	9657	19.2	8209	19.0	1448	20.2
Age 50–59 years	8281	16.4	7157	16.6	1124	15.7
Age 60–69 years	4715	9.4	4091	9.5	624	8.7
Age 70–79 years	2568	5.1	2294	5.3	274	3.8
Age 80–89 years	941	1.9	808	1.9	133	1.9
Age 90+ years	182	0.4	173	0.4	9	0.1
Female	25 219	50.1	21 495	49.8	3724	51.9
F0[a] Organic mental disorder	2506	5.0	2186	5.1	320	4.5
F1 Substance use disorders	13 018	25.9	11 164	25.8	1854	25.8
F2 Psychotic disorders	17 090	33.9	14 640	33.9	2450	34.1
F3 Affective disorders	18 231	36.2	15 556	36.0	2675	37.3
F4 Neurotic disorders	14 397	28.6	12 371	28.6	2026	28.2
F5 Eating, sleeping, and sexual disorders	1653	3.3	1380	3.2	273	3.8
F6 Personality disorders	6395	12.7	5644	13.1	751	10.5
F7 Mental retardation disorders	1631	3.2	1402	3.2	229	3.2
F8 Disorders of psychological development	2070	4.1	1703	3.9	367	5.1
F9 Child and adolescent disorders	5653 (11.2)	11.2	4719 (10.9)	10.9	934	13.0

[a](F*) indicates the ICD-10 chapter.

Online Supplementary Tables S5 and S6 list the 30 predictors with the highest information gain (XGBoost) and standard coefficients (elastic net). For the best XGBoost model 14 out of the 30 top predictors were text predictors – both represented by TF-IDF and sentence transformers. The TF-IDF predictors were based on the following terms from free text: 'ECT', 'police', 'social psychiatric institution', 'self-harm', and 'woman'. The 16 remaining predictors were distributed on the following patient descriptors: detention (sectioning during an inpatient stay after being admitted voluntarily), coercion due to danger to self or others, lab test of plasma-paracetamol, Brøset violence checklist score, diagnosis of child and adolescent disorder/unspecified mental disorder (ICD F9-chapter), visit due to a physical disorder, suicide risk assessment score, and diagnosis of personality disorder (ICD F6-chapter). As elastic net coefficients provide direction, the top predictors are divided into 15 positive (increases risk of outcome) and 15 negative coefficients (decreases risk of outcome). For the 15 top predictors with positive coefficients, six were free-text predictors only including TF-IDF predictors: 'Self-harm', 'Social Psychiatric Institution', 'Contact person', 'Eat', 'Mother', and 'Simultaneously'. The remaining nine top predictors covered the following patient descriptors: suicide risk score, detention, plasma paracetamol, visits due to a physical disorder, involuntary medication, alcohol abstinence medication, diagnosis of child and adolescent disorder/unspecified mental disorder (ICD-10, chapter F9), and Brøset violence checklist score. For the 15 top predictors with negative coefficients, 10 were free-text predictors including both TF-IDF and sentence transformers. The TF-IDF predictors were based on the following terms from free text: 'Energy', 'Looking forward to', 'Thursday', 'Receive', 'Daughter', and 'Interest'. The remaining five top predictors covered the following patient descriptors: clozapine, plasma clozapine (occurred twice with different lookbehind), plasma lithium, and coercion due to a physical disease.

The temporal stability of the elastic net and XGBoost models is visualized in online Supplementary Figs S6 and S7. Most model configurations show a gradual decline in performance as a function of time since the end of model training. However, models trained on data from a longer time period (and, thus, more data) display better temporal stability.

Secondary analyses on alternative model designs

Performance of the cross-validated models using different subsets of the full predictor set (Table 2A) different lookaheads (Table 2B), and models without the exclusion criterion of having an involuntary admission in the 2 years preceding the prediction time (Table 2B) are shown in Table 2. Among those trained on

Table 2. Model performance after cross-validation hyperparameter tuning for XGBoost and elastic net models trained on different subsets of the predictors (2A) and different lookaheads (2B)

Predictor set	Number of predictors	Number of training samples	Number of outcomes in training data	Internal AUROC	Internal 95% CI	Apparent AUROC	Model optimism
2A							
XGBoost							
Full predictor set	1828	36 611	1395	0.79	0.756–0.828	0.87	0.08
Patient descriptors only	694	36 611	1395	0.78	0.755–0.813	0.94	0.15
TF-IDF features only	752	36 611	1395	0.76	0.733–0.787	0.95	0.19
Sentence transformer embeddings only	386	36 611	1395	0.74	0.708–0.763	0.98	0.24
Diagnoses only	12	36 611	1395	0.64	0.618–0.666	0.66	0.02
Elastic net							
Full predictor set	1828	36 611	1395	0.78	0.749–0.813	0.83	0.053
Patient descriptors only	694	36 611	1395	0.76	0.731–0.786	0.79	0.032
TF-IDF features only	752	36 611	1395	0.74	0.717–0.772	0.80	0.055
Sentence transformer embeddings only	386	36 611	1395	0.73	0.697–0.761	0.76	0.035
Diagnoses only	12	36 611	1395	0.60	0.574–0.629	0.60	0.0021
2B							
XGBoost							
365 days lookahead	1828	36 805	2051	0.79	0.767–0.809	0.95	0.17
90 days lookahead	1828	40 867	991	0.80	0.777–0.826	0.93	0.13
Without exclusion criteria[a]	1828	49 331	6003	0.91	0.898–0.913	0.93	0.03
Elastic net							
365 days lookahead	1828	33 908	1916	0.78	0.753–0.799	0.82	0.048
90 days lookahead	1828	37 970	931	0.78	0.738–0.829	0.83	0.051
Without exclusion criteria[a]	1828	49 331	6003	0.90	0.890–0.906	0.91	0.011

All models included demographics (age/sex). The models with different lookahead window were trained on the full predictor set. Details on predictor description can be found in online Supplementary Table S1. Apparent AUROC represents the performance on the training data and the internal AUROC represents the performance on the test folds during cross-validation. Model optimism is calculated from the apparent AUROC – Interval AUROC.

[a]Models trained without the exclusion criterion of having an involuntary admission in the 2 years preceding the prediction time.

different subsets of predictors, the best performing model was the XGBoost model trained on only patient descriptors (no text). In the models trained on different outcome lookaheads, the models with a 90 day lookahead performed better than the ones with 180 and 365 day lookaheads. Both models trained without the exclusion criteria significantly outperformed any of the other model configurations.

Discussion

This study investigated if involuntary admission can be predicted using machine learning models trained on EHR data. When issuing a prediction at the discharge from a voluntary inpatient admission, based on both structured and text predictors, the best model (XGBoost) performed with an AUROC of 0.84. The model was generally stable across different patient characteristics, calendar times, and with varying times from prediction to outcome.

To our knowledge, this is the first study to develop and validate a prediction model for involuntary admission using routine clinical data from EHRs. We can, therefore, not offer a direct comparison of our results to those from other studies. However, a crude comparison to other prediction studies in psychiatry shows that our results are within the performance ranges that have previously been published (Meehan et al., 2022). Many of these studies have, however, not been developed on routine clinical data, but rather on data collected for the purpose of the studies, which complicates clinical implementation.

On the independent test set, the prediction model performed with an AUROC above the upper boundary of the confidence interval (CI) estimated from the five-fold cross-validation in the training phase for both XGBoost and elastic net models. While this suggests that the models did not overfit to the cross-validation training folds, it does expose a high degree of uncertainty in the precision of the performance measure. The variation in model performance might be attributable to both the limited number of positive cases in the test set and the general heterogeneity of the outcome and its underlying causes. The overall discrimination and calibration between the two models on the hold-out test set were similar. However, a main performance difference is observed in the number of unique predicted outcomes where the XGBoost

https://doi.org/10.1017/S0033291724002642 Published online by Cambridge University Press

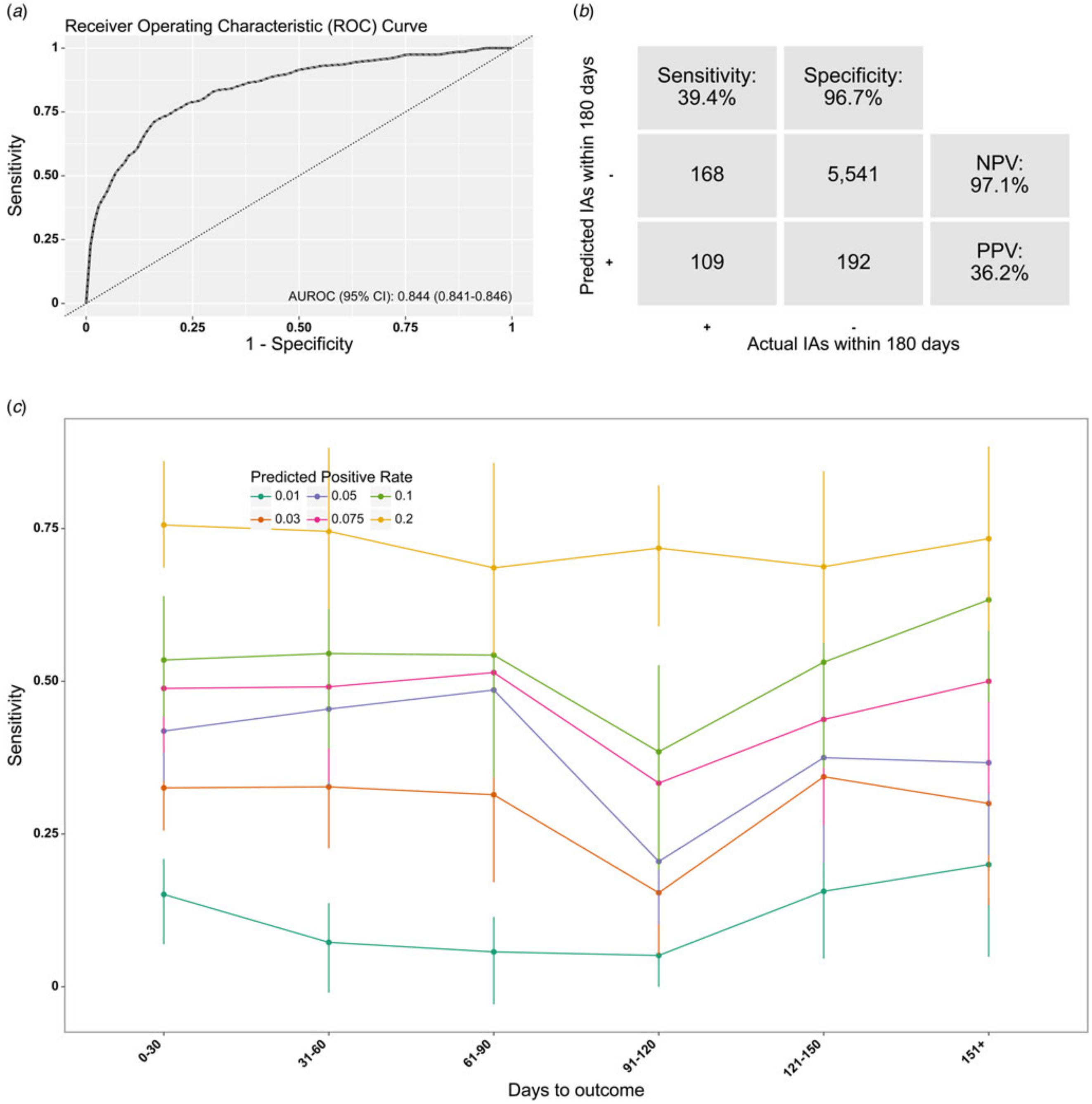
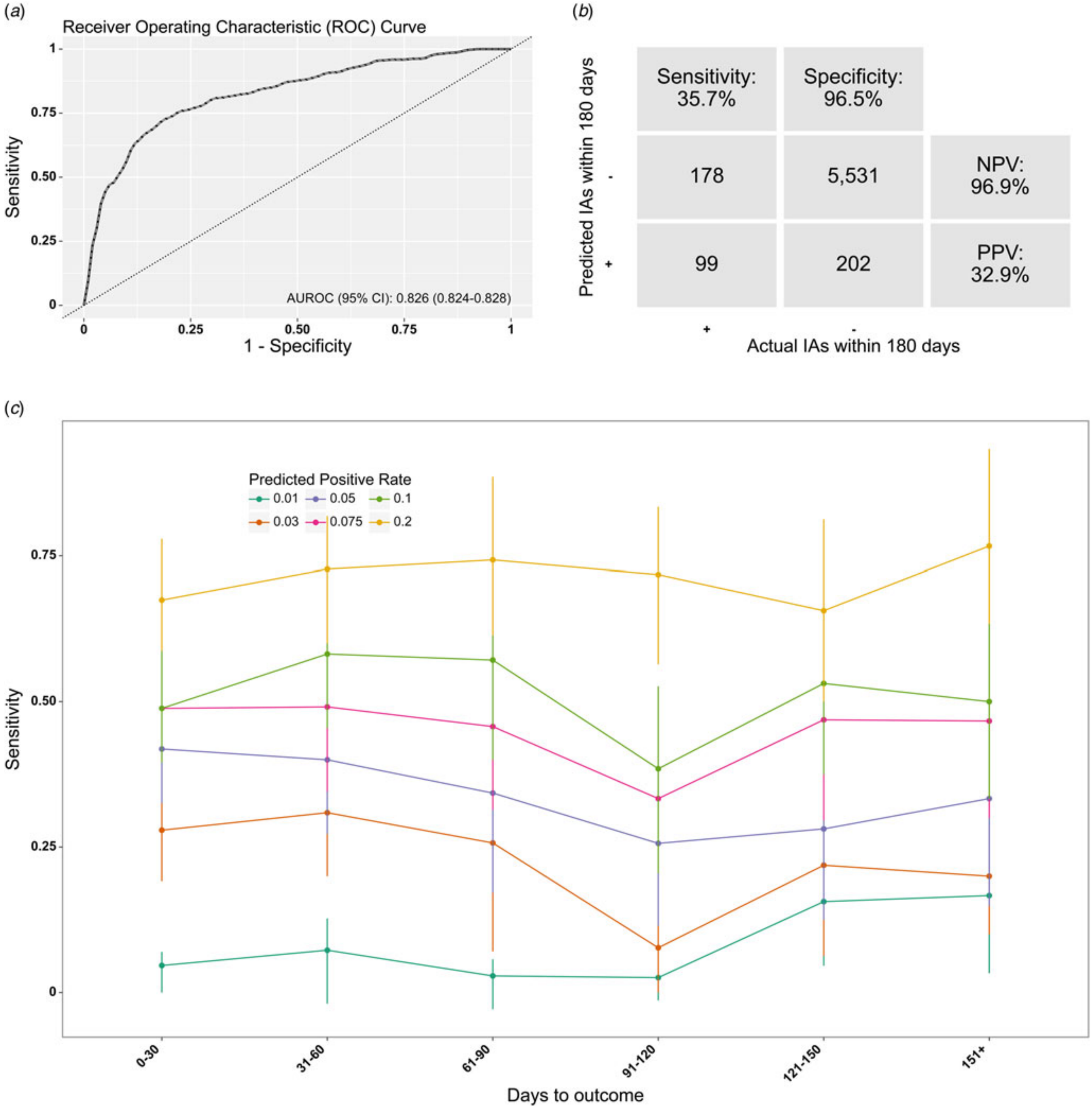


Figure 2. Model performance of the XGBoost model in the test set. (a) Receiver operating characteristics curve. AUROC, area under the receiver operating characteristics curve. (b) Confusion matrix. PPR, positive predictive rate; NPV, negative predictive value; IA, involuntary admission. The decision threshold is defined based on a PPR of 5%. (c) Sensitivity (at the same specificity) by months from prediction time to event, stratified by desired PPR.

model consistently outperformed the elastic net model across varying positive prediction rates. This metric is important when analyzing dynamic prediction models because multiple true positive predictions for the same outcome event do not necessarily lead to multiple interventions and, thus, increased probability of preventing the outcome event; once a patient has already been 'flagged' as high-risk, subsequent flaggings are not equally important. At a PPR of 5%, the XGBoost model correctly 'flagged' 36% of all unique involuntary admissions at least once. Strikingly, this rate is first achieved for the elastic net model when the PPR is set to 10%, meaning that double the amount of 'flaggings' need to be made by this model to identify the same number of unique outcome events.

When considering additional performance metrics, both models demonstrated relatively stable sensitivity when increasing time from prediction to outcome (up to several months), highlighting that model performance is not merely driven by prediction of cases where an involuntary admission occurs shortly after discharge from a voluntary admission. Indeed, the median time from the first-positive prediction to the involuntary admission of 70 (XGBoost) and 64 (elastic net) days is sufficient to issue a potentially preventive intervention through, e.g. advance

https://doi.org/10.1017/S0033291724002642 Published online by Cambridge University Press



	+	-	
	Sensitivity: 35.7%	Specificity: 96.5%	
-	178	5,531	NPV: 96.9%
+	99	202	PPV: 32.9%

Figure 3. Model performance of the elastic net model in the test set. (a) Receiver operating characteristics curve. AUROC, area under the receiver operating characteristics curve. (b) Confusion matrix. PPR, positive predictive rate; NPV, negative predictive value; IA, involuntary admission. The decision threshold is defined based on a PPR of 5%. (c) Sensitivity (at the same specificity) by months from prediction time to event, stratified by desired PPR.

statements/crisis plans (de Jong et al., 2016). Both models demonstrated clinical usefulness, showing a positive net benefit at a low threshold probability. This aligns with the fact that potential interventions, such as crisis plans and/or intensified outpatient treatment, would have a low clinical threshold since these interventions carry minimal side effects or risks.

A series of secondary analyses were conducted to explore the impact of various model design decisions. First, the exclusion criterion stipulating that patients could not have had an involuntary admission in the 2 years prior to a prediction time was added to minimize the potential alert fatigue in clinicians. Specifically, this measure aimed to omit scenarios where clinicians are likely already aware of an increased risk of involuntary admission, given that prior involuntary admission is a major risk factor for subsequent involuntary admission (Walker et al., 2019). Indeed, this was confirmed by our results as the model trained without this exclusion criterion performed with an AUROC of 0.91 (XGBoost) and 0.90 (elastic net) (on the cross-validated training set). This highlights the challenging balance between minimizing potential alert fatigue among clinicians and optimizing model performance for prediction models in healthcare. Second, the performance of the models trained on a limited feature set including

https://doi.org/10.1017/S0033291724002642 Published online by Cambridge University Press

Table 3. Performance metrics on test set for model trained on full predictor set at varying positive rates

A: XGBoost

PPR (%)	AUROC	True prevalence (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	FPR (%)	FNR (%)	Acc (%)	TP	TN	FP	FN
50.0	0.84	4.6	8.3	99.1	90.3	51.9	48.1	9.7	53.7	250	2978	2755	27
20.0	0.84	4.6	16.8	98.4	72.9	82.6	17.4	27.1	82.1	202	4733	1000	75
10.0	0.84	4.6	24.3	97.6	52.7	92.1	7.9	47.3	90.2	146	5278	455	131
7.5	0.84	4.6	28.6	97.3	46.6	94.4	5.6	53.4	92.2	129	5411	322	148
5.0	0.84	4.6	36.2	97.1	39.4	96.7	3.3	60.6	94.0	109	5541	192	168
4.0	0.84	4.6	40.2	96.9	35.0	97.5	2.5	65.0	94.6	97	5589	144	180
3.0	0.84	4.6	45.9	96.7	30.0	98.3	1.7	70.0	95.1	83	5635	98	194
2.0	0.84	4.6	50.8	96.3	22.4	99.0	1.0	77.6	95.4	62	5673	60	215
1.0	0.84	4.6	52.5	95.9	11.6	99.5	0.5	88.4	95.4	32	5704	29	245

PPR (%)	F1 (%)	MCC (%)	Total number of unique outcome events	Number of positive outcomes in test set (TP + FN)	Number of unique outcome events detected	Prop. of unique outcome events detected (%)	Median days from first positive to outcome
50.0	15.2	17.7	136	277	124	91.2	78
20.0	27.3	29.1	136	277	96	70.6	82
10.0	33.3	31.3	136	277	70	51.5	81
7.5	35.4	32.6	136	277	60	44.1	70
5.0	37.7	34.6	136	277	49	36.0	70
4.0	37.5	34.7	136	277	40	29.4	70
3.0	36.2	34.7	136	277	36	26.5	72
2.0	31.1	31.7	136	277	24	17.6	67
1.0	18.9	23.1	136	277	13	9.6	53

B: Elastic net

PPR (%)	AUROC	True prevalence (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	FPR (%)	FNR (%)	Acc (%)	TP	TN	FP	FN
50.0	0.83	4.6	8.1	98.8	87.4	51.8	48.2	12.6	53.4	242	2970	2763	35
20.0	0.83	4.6	16.3	98.3	70.8	82.5	17.5	29.2	81.9	196	4727	1006	81
10.0	0.83	4.6	23.5	97.5	50.9	92.0	8.0	49.1	90.1	141	5273	460	136
7.5	0.83	4.6	28.2	97.3	45.8	94.3	5.7	54.2	92.1	127	5409	324	150
5.0	0.83	4.6	32.9	96.9	35.7	96.5	3.5	64.3	93.7	99	5531	202	178
4.0	0.83	4.6	32.8	96.6	28.5	97.2	2.8	71.5	94.0	79	5571	162	198
3.0	0.83	4.6	36.5	96.4	23.8	98.0	2.0	76.2	94.6	66	5618	115	211
2.0	0.83	4.6	35.5	96.0	15.5	98.6	1.4	84.5	94.8	43	5655	78	234
1.0	0.83	4.6	32.8	95.7	7.2	99.3	0.7	92.8	95.0	20	5692	41	257

PPR (%)	F1 (%)	MCC (%)	Total number of unique outcome events	Number of positive outcomes in test set (TP + FN)	Number of unique outcome events detected	Prop. of unique outcome events detected (%)	Median days from first positive to outcome
50.0	14.7	16.4	136	277	86	63.2	79
20.0	26.5	27.9	136	277	69	50.7	78
10.0	32.1	30.0	136	277	49	36.0	82

(*Continued*)

https://doi.org/10.1017/S0033291724002642 Published online by Cambridge University Press

Table 3. (*Continued.*)

PPR (%)	F1 (%)	MCC (%)	Total number of unique outcome events	Number of positive outcomes in test set (TP + FN)	Number of unique outcome events detected	Prop. of unique outcome events detected (%)	Median days from first positive to outcome
7.5	34.9	32.0	136	277	43	31.6	76
5.0	34.3	31.0	136	277	36	26.5	64
4.0	30.5	27.5	136	277	24	17.6	43
3.0	28.8	26.8	136	277	20	14.7	48
2.0	21.6	21.1	136	277	13	9.6	48
1.0	11.8	13.6	136	277	5	3.7	44

Predicted positive rate (PPR): the proportion of contacts predicted positive by the model. Since the model outputs a predicted probability, this is a threshold set during evaluation. True prevalence: the proportion of admissions that qualified for an outcome within the lookahead window. AUROC: area under the receiver operator characteristic curve. PPV: positive predictive value. NPV: negative predictive value. FPR: false positive rate. FNR: false negative rate. TP: true positives. Numbers are based on prediction times (end of psychiatric admission). TN: true negatives. Numbers are based on prediction times (end of psychiatric admission). FP: false positives. Numbers are based on prediction times (end of psychiatric admission). FN: false negatives. Numbers are based on prediction times (end of psychiatric admission). F1: the harmonic mean of the precision and recall. MCC: Matthew's correlation coefficient. Prop. of unique outcome events detected: proportion of the involuntary admissions that are flagged by a least one true positive prediction. Median days from first positive to outcome: for all involuntary admissions with at least one true positive, the number of days of from first-positive prediction to outcome (involuntary admission).

only age, sex, and diagnoses was tested, resulting in an AUROC of 0.64 (XGBoost) and 0.60 (elastic net) (on the cross-validated training set). This demonstrates that using the full predictor set resulted in substantially better predictive performance, underlining the complexity of risk prediction at the level of the individual patient. Third, a lookahead window of 180 days was chosen for the main model as this leaves a reasonable window of opportunity for prevention of an involuntary admission. Models trained with a lookahead window of 90 days achieved an AUROC of 0.80 (XGBoost) and 0.78 (elastic net) and a lookahead window of 365 days achieved an AUROC of 0.79 (XGBoost) and 0.78 (elastic net) (on the cross-validated training set). This further validates the performance-wise stability of the method across different time-to-outcome intervals and justifies determining the optimal lookahead window based on clinical judgment.

With regard to the predictors driving the discriminative abilities of the model, text features comprised of 14 out of top 30 predictors for XGBoost and 16 out of the top 30 predictors for the elastic net, showcasing the importance of including text. This might be especially true for the field of psychiatry where the clinical condition of patients is mainly described in natural language in the EHR rather than in structured variables. The inclusion of predictors based on TF-IDF and sentence transformer features/embeddings of the text also overall indicated an increased performance of the model. This is in line with prior results of both our own (Danielsen et al., 2019) and others (Irving et al., 2021; Tenenbaum & Ranallo, 2021). Among the XGBoost predictors extracted from the free text using TF-IDF, 'ECT' (electroconvulsive therapy), 'police', and 'self-harm' were among the predictors with the highest predictive value. These terms are very likely proxies for severity, as ECT is mainly used for very severe manifestations of unipolar depression, bipolar disorders and schizophrenia (Espinoza & Kellner, 2022), involvement of the police is also suggestive of severe illness (e.g. aggression or suicidality) (Canova Mosele, Chervenski Figueira, Antônio Bertuol Filho, Ferreira De Lima, & Calegaro, 2018; Mortensen, Agerbo, Erikson, Qin, & Westergaard-Nielsen, 2000), and self-harm may refer to a spectrum of behavior from, e.g. superficial cutting to suicide attempts (Skegg, 2005). For the elastic net model, the positive top predictors (increases risk of outcome) using TF-IDF included 'self-harm', 'social psychiatric institution', and the negative top predictors (decreases risk of outcome) included 'energy', 'looking forward to', and 'interest' which makes intuitive sense from a clinical perspective as the first are proxies for severity of illness, while the latter reflect psychological well-being. There are also TF-IDF text predictors that are challenging to interpret without the broader context of the clinical notes, such as terms like 'Thursday', 'receive', and 'eat'. While sentence transformers can capture the contextual meaning of clinical notes, they currently lack interpretability of their embeddings (Reimers & Gurevych, 2019). There were some similarities in the structured top predictors of the two models (e.g. prior detention, Brøset violence checklist score, and suicide risk assessment score), although a direct comparison of predictors from different models should be done with caution. Furthermore, a lab test of plasma-paracetamol – another top predictor of both models – likely indicates that a patient has taken a toxic dose of paracetamol in relation to self-harm or a suicide attempt – also a manifestation of severe mental illness (Reuter Morthorst, Soegaard, Nordentoft, & Erlangsen, 2016). Among the negative top predictors (decreases risk of outcome) of the elastic net model, several were related to clozapine treatment, including a lab test for plasma clozapine levels, suggesting that continuous clozapine treatment, arguably the most effective antipsychotic agent for treatment of schizophrenia (Kane, 1988; McEvoy et al., 2006), may be associated with a lower risk of involuntary admission.

Both information gain estimates and coefficients should be interpreted with caution due to model-dependent handling of predictors in the model training processes. Specifically, due to the structure of a decision tree model, top predictors containing mutual information can be omitted. Similarly, elastic net employs lasso regularization which prunes out highly correlated features. Consequently, this may lead to only one of several mutual information predictors appearing in the predictor importance tables (Chen & Guestrin, 2016). The most important insight from the top predictors of both models may be that the model is not informed by a few dominant predictors, but instead relies on a plethora of predictors. In line with this, the models trained on a limited feature set 'Diagnoses only' performed with an AUROC of, respectively, 0.64 (XGBoost) and 0.60 (elastic net) on the training set. This demonstrates the complexity of the outcome and supports our approach of processing and including a large and

https://doi.org/10.1017/S0033291724002642 Published online by Cambridge University Press

diverse set of predictors, thus, enabling the models to locate the relevant information independently.

There are no set performance thresholds when a prediction algorithm should be considered for clinical implementation. At a PPR of 5%, the best performing model in the present study had a specificity of 97%, a sensitivity of 39%, an NPV of 97%, and a PPV of 36%. Both models demonstrated clinical usefulness, showing a net benefit even at a low threshold probability in the decision curve analysis. Considering that potential preventive measures informed by the model such as advance statements/crisis plans (de Jong et al., 2016) are both cheap and, presumably, without substantial side effects, we would argue that implementation could, indeed, be considered. Successful implementation would rely on the clinical staff being presented with the 'at risk' assessment (flagging) by the model, such that the preventive intervention can be elicited at the right time. In the Psychiatric Services of the Central Denmark Region, our EHR system supports this modality, which is currently being implemented alongside a mechanical restraint prediction model (Danielsen et al., 2019). Furthermore, the clinical staff will have to trust the risk assessment performed by the model. In our experience, this requires targeted information/training of the staff (Perfalk, Bernstorff, Danielsen, & Østergaard, 2024). Investigation of implementation strategies, cost-effectiveness, and clinical utility of the models is beyond the scope of the present study, but should be explored going forward.

There are limitations to this study, which should be taken into account. First, there is a limited number of outcomes (involuntary admissions) in the dataset, and the main model considered a total of 1828 predictors. If not handled properly, this could result in 'curse of dimensionality' (Berisha et al., 2021) and lead to potential overfitting. Furthermore, the limited number of outcomes introduces uncertainty in the robustness of estimates for specific subgroups, such as age and gender. To mitigate this, we employed several strategies: structured predictors were constructed based on findings from prior research and clinical domain knowledge, we used cross-validation during training, and, during hyperparameter tuning, feature selection was adopted. Finally, we used a hold-out test set to ensure that potential overfitting during the training phase is accounted for in the evaluation. Second, the test set was not independent with regard to time or geographic location, i.e. the generalizability of the model across these boundaries has not been tested. Machine learning models inherently vary in their generalizability and reusing our model 1:1 in another hospital setting would probably result in reduced performance. However, the overall approach is likely to be generalizable and, thus, retraining the model on another EHR dataset, while keeping the same architecture, could enable transferability (Curth et al., 2020). The temporal stability analyses (see online Supplementary Figs S6 and S7) showed, as expected, slight decline in model performances as a function of time since the end of model training. Some of this decline may be driven by insufficient data (i.e. few outcomes). Also, the dataset spans several abnormalities, namely a transition to new diagnostic registration practices in March of 2019 (Bernstorff et al., 2022) and the COVID-19 pandemic with onset in 2020. These events likely partially explain the general drops in performance that can be observed in 2019–2021. Ultimately, if implemented in clinical practice, it is crucial to monitor the model's performance over time and continuously recalibrate it if necessary. Third, despite several text predictors demonstrating high predictive value, the methods for obtaining the predictors from the free-text notes were relatively simple. In future studies, we believe it may be possible to unlock vastly more predictive value from the text by applying more advanced language models. Specifically, a future direction could involve a transformer-based model fine-tuned specifically to psychiatric clinical notes and the given prediction task (Huang, Altosaar, & Ranganath, 2020). Fourth, the approach in this project is characterized by fitting a classical binary prediction framework to a task that is inherently sequential in nature. As sequential transformer-based models are gradually adapted from language modeling to the general health care domain, it is likely that such architectures may be better suited to this task and will enhance performance. The adaptation of transformer-based models to the healthcare domain is, however, still in an explorative phase, and, hence, we deem that involving such methods in this study – which was aimed at developing a model for potential clinical implementation – would be premature. Fifth, elastic net and XGBoost do not inherently handle the potential problems with repeated risk predictions on the same individual which could lead to overfitting on individual-specific risk trajectories. However, we ensured that no patient was present in both the train- and the test sets, both during cross-validation, and for the independent hold-out test set. Furthermore, if overfitting on individual-specific risk trajectories had occurred, it would have negatively impacted performance on the hold-out test set. However, no such drop in performance was observed.

Conclusion

A machine learning model using routine clinical data from EHRs can accurately predict involuntary admission. If implemented as a clinical decision support tool, this model may guide interventions aimed at reducing the risk of involuntary admission.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291724002642.

Data availability statement. No pre-registration of the study was carried out. According to Danish law, the patient-level data for this study cannot be shared. The code for all analyses is available at: https://github.com/Aarhus-Psychiatry-Research/psycop-common/tree/main/psycop/projects/forced_admission_inpatient.

Acknowledgments. The authors are grateful to Bettina Nørremark for data management.

Author contributions. The study was conceptualized and designed by all authors. Funding was raised by S. D. Ø. The data were procured by S. D. Ø. The statistical analyses were carried out by E. P. and J. G. D. All authors contributed to the interpretation of the results. E. P. wrote the first draft of the manuscript, which was subsequently revised for important intellectual content by the remaining authors. All authors approved the final version of the manuscript prior to submission.

Funding statement. The study is supported by grants to S. D. Ø. from the Lundbeck Foundation (grant number: R344-2020-1073), the Central Denmark Region Fund for Strengthening of Health Science (grant number: 1-36-72-4-20), the Danish Agency for Digitisation Investment Fund for New Technologies (grant number: 2020-6720), and by a private donation to the Department of Affective Disorders at Aarhus University Hospital – Psychiatry to support research into bipolar disorder. S. D. Ø. has received grants for other purposes from the Novo Nordisk Foundation (grant number: NNF20SA0062874), the Danish Cancer Society (grant number: R283-A16461), the Lundbeck Foundation (grant number: R358-2020-2341), and Independent Research Fund Denmark (grant numbers: 7016-00048B and 2096-00055A). These funders had no role in the study design, data analysis, interpretation of data, or writing of the manuscript.

Competing interests. A. A. D. has received a speaker honorarium from Otsuka Pharmaceutical. S. D. Ø. received the 2020 Lundbeck Foundation Young Investigator Prize and S. D. Ø. owns/has owned units of mutual funds with stock tickers DKIGI, IAIMWC, SPIC25KL, and WEKAFKI, and owns/has owned units of exchange traded funds with stock tickers BATE, TRET, QDV5, QDVH, QDVE, SADM, IQQH, IQQJ, USPY, EXH2, 2B76, IS4S, OM3X, and EUNL. The remaining authors declare no competing interests.

References

Bech, P., Rafaelsen, O. J., Kramp, P., & Bolwig, T. G. (1978). The mania rating scale: Scale construction and inter-observer agreement. *Neuropharmacology*, *17*(6), 430–431. https://doi.org/10.1016/0028-3908(78)90022-9

Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., & Liss, J. (2021). Digital medicine and the curse of dimensionality. *Npj Digital Medicine*, *4*(1), 153. https://doi.org/10.1038/s41746-021-00521-5

Bernstorff, M., Hansen, L., Perfalk, E., Danielsen, A. A., & Østergaard, S. D. (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. https://doi.org/10.1111/acps.13463

Bernstorff, M., Enevoldsen, K., Damgaard, J., Danielsen, A., & Hansen, L. (2023). Timeseriesflattener: A python package for summarizing features from (medical) time series. *Journal of Open Source Software*, *8*(83), 5197. https://doi.org/10.21105/joss.05197

Bernstorff, M., Hansen, L., Enevoldsen, K., Damgaard, J., Hæstrup, F., Perfalk, E., … Østergaard, S. D. (2024). Development and validation of a machine learning model for prediction of type 2 diabetes in patients with mental illness. *Acta Psychiatrica Scandinavica*, acps.13687. https://doi.org/10.1111/acps.13687

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007

Canova Mosele, P. H., Chervenski Figueira, G., Antônio Bertuol Filho, A., Ferreira De Lima, J. A. R., & Calegaro, V. C. (2018). Involuntary psychiatric hospitalization and its relationship to psychopathology and aggression. *Psychiatry Research*, *265*, 13–18. https://doi.org/10.1016/j.psychres.2018.04.031

Cerqueira, V., Torgo, L., & Soares, C. (2019, September 29). Machine learning vs statistical methods for time series forecasting: size matters. arXiv. Retrieved from http://arxiv.org/abs/1909.13316

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Curth, A., Thoral, P., van den Wildenberg, W., Bijlstra, P., de Bruin, D., Elbers, P., & Fornasa, M. (2020). Transferring clinical prediction models across hospitals and electronic health record systems. In P. Cellier & K. Driessens (Eds.), *Machine learning and knowledge discovery in databases* (pp. 605–621). Cham: Springer International Publishing.

Danielsen, A. A., Fenger, M. H. J., Østergaard, S. D., Nielbo, K. L., & Mors, O. (2019). Predicting mechanical restraint of psychiatric inpatients by applying machine learning on electronic health data. *Acta Psychiatrica Scandinavica*, *140*(2), 147–157. https://doi.org/10.1111/acps.13061

de Jong, M. H., Kamperman, A. M., Oorschot, M., Priebe, S., Bramer, W., van de Sande, R., … Mulder, C. L. (2016). Interventions to reduce compulsory psychiatric admissions: A systematic review and meta-analysis. *JAMA Psychiatry*, *73*(7), 657. https://doi.org/10.1001/jamapsychiatry.2016.0501

Desai, R. J., Wang, S. V., Vaduganathan, M., Evers, T., & Schneeweiss, S. (2020). Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Network Open*, *3*(1), e1918962. https://doi.org/10.1001/jamanetworkopen.2019.18962

Espinoza, R. T., & Kellner, C. H. (2022). Electroconvulsive therapy. *New England Journal of Medicine*, *386*(7), 667–672. https://doi.org/10.1056/NEJMra2034954

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022, July 18). Why do tree-based models still outperform deep learning on tabular data? arXiv. Retrieved from http://arxiv.org/abs/2207.08815

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, *23*(1), 56–62.

Hansen, L., Enevoldsen, K. C., Bernstorff, M., Nielbo, K. L., Danielsen, A. A., & Østergaard, S. D. (2021). The PSYchiatric clinical outcome prediction (PSYCOP) cohort: Leveraging the potential of electronic health records in the treatment of mental disorders. *Acta Neuropsychiatrica*, *33*(6), 323–330. https://doi.org/10.1017/neu.2021.22

Hansen, L., Enevoldsen, K., Bernstorff, M., Perfalk, E., Danielsen, A. A., Nielbo, K. L., & Østergaard, S. D. (2023). Lexical stability of psychiatric clinical notes from electronic health records over a decade. *Acta Neuropsychiatrica*, *25*, 1–11. https://doi.org/10.1017/neu.2023.46

Huang, K., Altosaar, J., & Ranganath, R. (2020, November 28). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. arXiv. Retrieved from http://arxiv.org/abs/1904.05342

Irving, J., Patel, R., Oliver, D., Colling, C., Pritchard, M., Broadbent, M., … Fusar-Poli, P. (2021). Using natural language processing on electronic health records to enhance detection and prediction of psychosis risk. *Schizophrenia Bulletin*, *47*(2), 405–414. https://doi.org/10.1093/schbul/sbaa126

Kane, J. (1988). Clozapine for the treatment-resistant schizophrenic: A double-blind comparison with chlorpromazine. *Archives of General Psychiatry*, *45*(9), 789. https://doi.org/10.1001/archpsyc.1988.01800330013001

Karasch, O., Schmitz-Buhl, M., Mennicken, R., Zielasek, J., & Gouzoulis-Mayfrank, E. (2020). Identification of risk factors for involuntary psychiatric hospitalization: Using environmental socioeconomic data and methods of machine learning to improve prediction. *BMC Psychiatry*, *20*(1), 401. https://doi.org/10.1186/s12888-020-02803-w

Katsakou, C., & Priebe, S. (2007). Patient's experiences of involuntary hospital admission and treatment: A review of qualitative studies. *Epidemiologia e Psichiatria Sociale*, *16*(2), 172–178. https://doi.org/10.1017/S1121189X00004802

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. arXiv:1705.07874 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1705.07874

McEvoy, J. P., Lieberman, J. A., Stroup, T. S., Davis, S. M., Meltzer, H. Y., Rosenheck, R. A., … Davis, C. E. (2006). Effectiveness of clozapine versus olanzapine, quetiapine, and risperidone in patients with chronic schizophrenia who did not respond to prior atypical antipsychotic treatment. *American Journal of Psychiatry*, *163*(4), 600–610.

Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., & Danese, A. (2022). Clinical prediction models in psychiatry: A systematic review of two decades of progress and challenges. *Molecular Psychiatry*, *27*(6), 2700–2708. https://doi.org/10.1038/s41380-022-01528-4

Mortensen, P., Agerbo, E., Erikson, T., Qin, P., & Westergaard-Nielsen, N. (2000). Psychiatric illness and risk factors for suicide in Denmark. *The Lancet*, *355*(9197), 9–12. https://doi.org/10.1016/S0140-6736(99)06376-X

Nusinovici, S., Tham, Y. C., Chak Yan, M. Y., Wei Ting, D. S., Li, J., Sabanayagam, C., … Cheng, C.-Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Cournapeau, D. (2011). Scikit-learn: Machine learning in python. *Machine Learning in Python*, *10/2011*, 2825–2830.

Perfalk, E., Bernstorff, M., Danielsen, A. A., & Østergaard, S. D. (2024, September 10). Receiving information on machine learning-based clinical decision support systems in psychiatric services increases staff trust in

https://doi.org/10.1017/S0033291724002642 Published online by Cambridge University Press

these systems: A randomized survey experiment. https://doi.org/10.1101/2024.09.09.24313303

Reimers, N., & Gurevych, I. (2019, August 27). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. arXiv. Retrieved from http://arxiv.org/abs/1908.10084

Reuter Morthorst, B., Soegaard, B., Nordentoft, M., & Erlangsen, A. (2016). Incidence rates of deliberate self-harm in Denmark 1994–2011: A nationwide register study. *Crisis*, *37*(4), 256–264. https://doi.org/10.1027/0227-5910/a000391

Salize, H. J., & Dressing, H. (2004). Epidemiology of involuntary placement of mentally ill people across the European Union. *British Journal of Psychiatry*, *184*(2), 163–168. https://doi.org/10.1192/bjp.184.2.163

Sheridan Rains, L., Zenina, T., Dias, M. C., Jones, R., Jeffreys, S., Branthonne-Foster, S., … Johnson, S. (2019). Variations in patterns of involuntary hospitalisation and in legal frameworks: An international comparative study. *The Lancet Psychiatry*, *6*(5), 403–417. https://doi.org/10.1016/S2215-0366(19)30090-2

Silva, B., Gholam, M., Golay, P., Bonsack, C., & Morandi, S. (2021). Predicting involuntary hospitalization in psychiatry: A machine learning investigation. *European Psychiatry*, *64*(1), e48. https://doi.org/10.1192/j.eurpsy.2021.2220

Skegg, K. (2005). Self-harm. *The Lancet*, *366*(9495), 1471–1483. https://doi.org/10.1016/S0140-6736(05)67600-3

Song, X., Mitnitski, A., Cox, J., & Rockwood, K. (2004). Comparison of machine learning techniques with classical statistical models in predicting health outcomes. *Studies in Health Technology and Informatics*, *107*(Pt 1), 736–740.

Sundhedsdatastyrelsen:Register over Anvendelse af Tvang i Psykiatrien. (2024). Retrieved from http://www.esundhed.dk/.

Tenenbaum, J. D., & Ranallo, P. A. (Eds.). (2021). *Mental health informatics: Enabling a learning mental healthcare system*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-70558-9

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989X06295361

Walker, S., Mackay, E., Barnett, P., Sheridan Rains, L., Leverton, M., Dalton-Locke, C., … Johnson, S. (2019). Clinical and social factors associated with increased risk for involuntary psychiatric hospitalisation: A systematic review, meta-analysis, and narrative synthesis. *The Lancet Psychiatry*, *6*(12), 1039–1053. https://doi.org/10.1016/S2215-0366(19)30406-7

Woods, P., & Almvik, R. (2002). The Brøset violence checklist (BVC). *Acta Psychiatrica Scandinavica*, *106*(s412), 103–105. https://doi.org/10.1034/j.1600-0447.106.s412.22.x

World Health Organization. (n.d.-a). *Anatomical therapeutic chemical (ATC) classification*. World Health Organization. Retrieved from https://www.whocc.no/atc_ddd_index/.

World Health Organization. (n.d.-b). *International statistical classification of diseases and related health problems/World Health Organization* (Vol. 2004). World Health Organization.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

https://doi.org/10.1017/S0033291724002642 Published online by Cambridge University Press