*European Psychiatry*

**www.cambridge.org/epa**

## Research Article

**Cite this article:** Bernstorff M, Hansen L, Olesen KKW, Danielsen AA, Østergaard SD (2025). Predicting cardiovascular disease in patients with mental illness using machine learning. *European Psychiatry*, **68**(1), e12, 1–10 https://doi.org/10.1192/j.eurpsy.2025.1

Received: 23 August 2024
Revised: 15 December 2024
Accepted: 30 December 2024

**Keywords:**
artificial intelligence; cardiovascular diseases; psychiatry; precision medicine

**Corresponding author:**
Martin Bernstorff;
Email: manber@rm.dk

# Predicting cardiovascular disease in patients with mental illness using machine learning

Martin Bernstorff[1,2,3], Lasse Hansen[1,2,3], Kevin Kris Warnakula Olesen[4], Andreas Aalkjær Danielsen[1,2] and Søren Dinesen Østergaard[1,2]

[1]Department of Affective Disorders, Aarhus University Hospital – Psychiatry, Aarhus, Denmark; [2]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark; [3]Center for Humanities Computing, Aarhus University, Aarhus, Denmark and [4]Department of Cardiology, Aarhus University Hospital, Aarhus, Denmark

## Abstract

**Background.** Cardiovascular disease (CVD) is twice as prevalent among individuals with mental illness compared to the general population. Prevention strategies exist but require accurate risk prediction. This study aimed to develop and validate a machine learning model for predicting incident CVD among patients with mental illness using routine clinical data from electronic health records.

**Methods.** A cohort study was conducted using data from 74,880 patients with 1.6 million psychiatric service contacts in the Central Denmark Region from 2013 to 2021. Two machine learning models (XGBoost and regularised logistic regression) were trained on 85% of the data from six hospitals using 234 potential predictors. The best-performing model was externally validated on the remaining 15% of patients from another three hospitals. CVD was defined as myocardial infarction, stroke, or peripheral arterial disease.

**Results.** The best-performing model (hyperparameter-tuned XGBoost) demonstrated acceptable discrimination, with an area under the receiver operating characteristic curve of 0.84 on the training set and 0.74 on the validation set. It identified high-risk individuals 2.5 years before CVD events. For the psychiatric service contacts in the top 5% of predicted risk, the positive predictive value was 5%, and the negative predictive value was 99%. The model issued at least one positive prediction for 39% of patients who developed CVD.

**Conclusions.** A machine learning model can accurately predict CVD risk among patients with mental illness using routinely collected electronic health record data. A decision support system building on this approach may aid primary CVD prevention in this high-risk population.

## Introduction

CVD not only diminishes quality of life but also contributes substantially to premature mortality [1,2]. Individuals with mental illness are twice as likely to develop CVD compared to the background population [3, 4], and are at elevated risk of premature death due to CVD [2]. This elevated risk can likely be attributed to higher prevalence of unhealthy lifestyle, such as poor diet, sedentary behaviour, and excessive alcohol consumption [5]. Additionally, psychopharmacological treatment, antipsychotics in particular, acts as a double-edged sword in the context of CVD, increasing risk due to weight gain and dysmetabolism [6], while being associated with lower risk of cardiovascular disease in observational studies [7], likely via the beneficial effect on the underlying mental disorder.

Unfortunately, the elevated risk of CVD among those with mental illness is not reflected in the administration of preventive measures, with screening for CVD occurring at 25% lower rates among individuals with mental illness [3,8], and up to 88% of individuals with schizophrenia with dyslipidaemia not receiving adequate treatment for the latter [9]. Consequently, identifying individuals with mental illness at elevated risk of CVD is a crucial initial step towards implementing effective preventive strategies. However, to the best of our knowledge, there is a paucity of tools designed for predicting CVD risk among patients receiving treatment in psychiatric service systems.

Accurately assessing CVD risk is a multifaceted challenge. Machine learning models are particularly well-suited for this task, given the presence of numerous interacting factors increasing CVD risk [10], and the model's ability to capture complex relationships while mitigating the impact of data idiosyncrasies [11]. Previous research has demonstrated the efficacy of machine learning models in accurately predicting clinical outcomes for patients with mental disorders when trained on electronic health record data. Specifically, it has been possible to predict, for example, mechanical restraint [12], progression from prediabetes to type 2 diabetes [13], and incidence of type 2 diabetes [14]. In line with these achievements, to aid the identification of patients with mental illness who may benefit from targeted intervention to prevent CVD, we aimed to develop and validate a machine learning model trained on electronic health record data to predict the development of CVD among patients with mental illness.

## Methods

The methods are illustrated by panels A-I in Figure 1.

### *Data and cohort extraction*

This study is based on electronic health record data from the PSYchiatric Clinical Outcome Prediction (PSYCOP) cohort, which encompasses all individuals with at least one contact with the Psychiatric Services of the Central Denmark Region in the period from January 1, 2011, and November 22, 2021. The dataset includes information from routine clinical practice (i.e., there was no specific data collection for the purpose of this study) on service contacts, diagnoses, medications, procedures, and laboratory results from all public hospitals (psychiatric as well as general hospitals) in the Central Denmark Region (Figure 1A). Denmark has a tax-financed universal public healthcare system.

A flowchart illustrating the definition of the patient cohort is available in eFigure 1. For this study, we restricted the cohort to patients with contacts to the Psychiatric Services of the Central Denmark Region after January 1, 2013, due to data instability prior to this date caused by the implementation of a new electronic health record system [15, 16]. Only patients aged 18 years or older were included, as the probability of developing CVD is very low in those below the age of 18. Patients with known CVD, defined by meeting one of the outcome criteria (see below) between January 1, 2011, and December 31, 2013, were excluded to minimise issuing of predictions for prevalent cases.

### *Outcome definition (cardiovascular disease)*

The outcome definition had three elements. First, to align with prior research, we took inspiration from the outcome definition from the Systematic Coronary Risk Evaluation 2 (SCORE2) [17]. Specifically, we defined incident CVD as the first occurrence of a diagnosis of myocardial infarction (MI) (International Classification of Diseases, 10th revision (ICD-10): I21-I23 or a diagnosis of stroke (ICD-10: I6, (Figure 1B). Second, we included interventions/procedures which are highly indicative of vascular disease (procedure codes are available in eTable 1) to the outcome definition, namely percutaneous coronary intervention (PCI), coronary artery bypass grafting (CABG), intracranial endovascular thrombolysis and other intracranial endovascular surgery. Third, given the large morbidity and disability burden due to peripheral arterial disease, its increasing incidence, and the potential for prevention [18], we included diagnoses (ICD-10: I70.2, I73.9) and procedures (procedure codes are available in eTable 1) for iliac, femoral, popliteal and distal arterial disease to the outcome definition.

### *Data splitting*

The data were divided into two subsets: a training dataset (85% of the data) and a test dataset (15% of the data). Specifically, all visits to the Psychiatric Services in either the western or eastern part of the Central Denmark Region (Aarhus, Gødstrup, Herning, Holstebro, Horsens, and Randers) were used for the training set, and the central part (Viborg, Silkeborg, and Skive) for the test-set (see Figure 1C). If a patient first had visits in one of the splits (i.e. the training set or the test set), any subsequent visits in the other split were removed. This guaranteed that no patient appeared in both the training and test datasets. After this point, the test dataset was left aside and only used for the final evaluation of the best-performing model obtained during the training phase. This geographical split assessed the generalizability across geography, for example, to which extent the model could be applied without modification if a new hospital was added to the region.

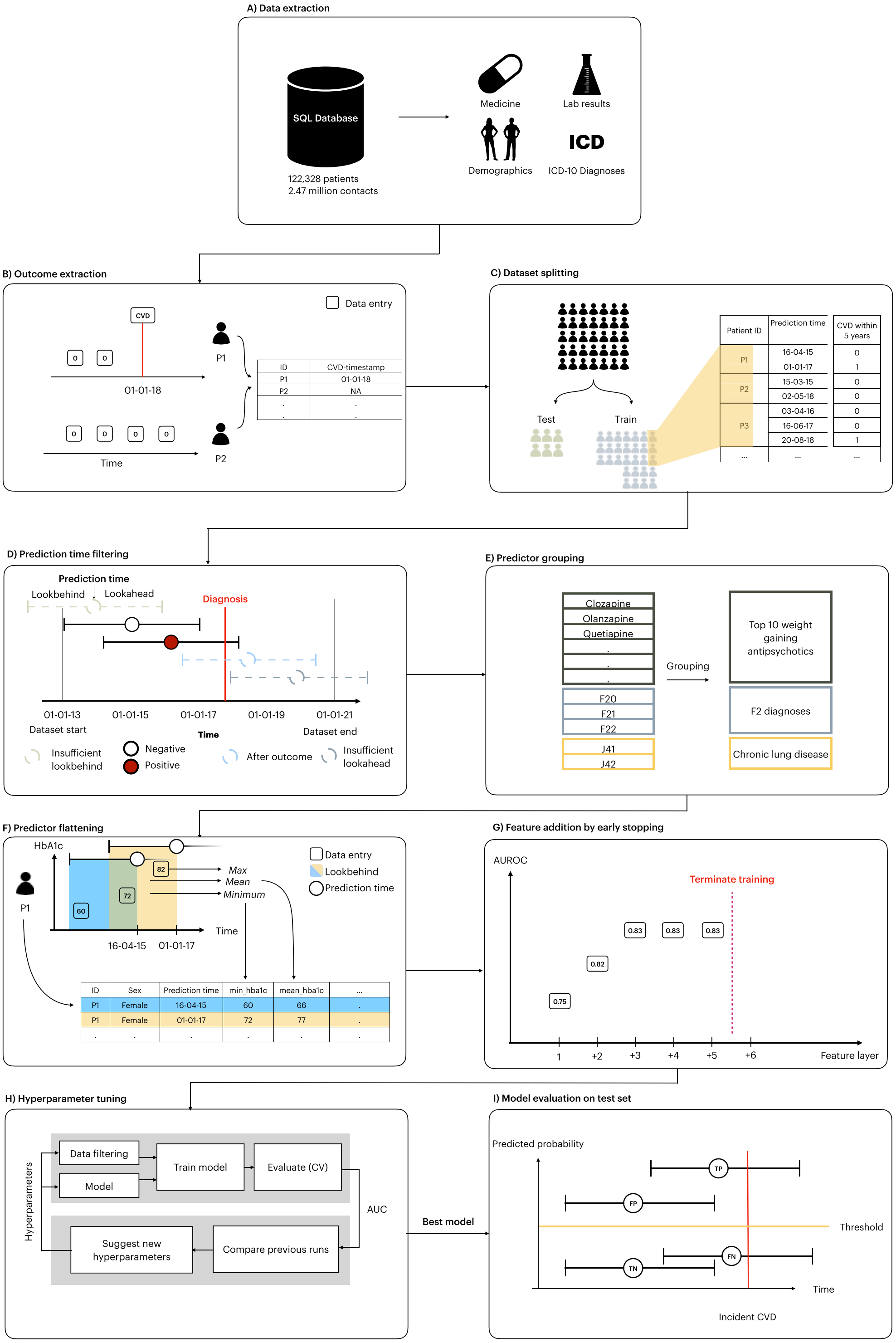
### *Prediction time filtering*

We defined prediction times as the time of any in- or out-patient contact with the Psychiatric Services (service contacts). Consequently, each patient could have multiple prediction times – corresponding to their number of service contacts. We excluded prevalent cases by not issuing a prediction if that patient had already met the CVD outcome criteria at the time of a service contact (Figure 1D). Moreover, no prediction was made if the lookbehind window (the time used for extracting predictors) included time before follow-up started on January 1, 2013, or if the lookahead window (the time within which to detect the outcome) of 2 years extended beyond the end of follow-up, the date of moving out of the Central Denmark Region, or the patient's death. These "truncations" are artefacts caused by data collection. If not accounted for, they could cause the model to learn patterns that do not exist during implementation, leading to discrepancies between the model's test performance and actual implemented performance. In the case of a patient moving into the region, we did not issue predictions for two years after the move, mirroring the wash-in for existing patients.

### *Predictor grouping and flattening*

Predictors were chosen based on a recent meta-analysis of prediction models for CVD in non-psychiatric settings and included demographics, laboratory results, diagnoses, antipsychotics, and mood stabilisers [19]. Specifically, the following predictors were included, all operationalised using routine clinical electronic health record data from the Central Denmark Region: age, sex, smoking status, high- and low-density lipoprotein (HDL and LDL), haemoglobin A1c (HbA1c), systolic blood pressure, diagnosis of chronic lung disease (ICD-10: J40–J44*), diagnoses from all psychiatric subchapters individually (F0–F9), as well as the use of any one of the top 10 weight gaining antipsychotics during inpatient treatment (Anatomical Therapeutic Chemical classification codes in parentheses): clozapine (N05AH02), zotepine (N05AX11), olanzapine (N05AH03), sertindole (N05AE03), chlorpromazine (N05AA01), iloperidone (N05AX14), quetiapine (N05AH04), paliperidone (N05AX13), trifluoperazine (N05AB06), and risperidone (N05AX08), resulting in 26 eligible features (Figure 1E) [20,21]. These predictors were aggregated over the lookbehind windows (90, 365, and 730) days, to incorporate different temporal contexts, and with different aggregation methods (min, mean, and max) using the timeseriesflattener python package [22], resulting in a total of 234 potential predictors (Figure 1F). For further elaboration, see the Supplementary Material.

The dataset includes numerous predictors lacking values within the lookbehind window. However, these absent values do not constitute missing data in the conventional sense, as they are not a result of omitted data entry. Instead, the absence of data reflects the reality of clinical practice. Since this absence aligns



| ID | CVD-timestamp |
|---|---|
| P1 | 01-01-18 |
| P2 | NA |
| . | . |
| . | . |

| Patient ID | Prediction time | CVD within 5 years |
|---|---|---|
| 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 time | min_hba1c | mean_hba1c | ... |
|---|---|---|---|---|---|
| P1 | Female | 16-04-15 | 60 | 66 | . |
| P1 | Female | 01-01-17 | 72 | 77 | . |
| . | . | . | . | . | . |

**Figure 1.** Extraction of data and outcome, dataset splitting, prediction time filtering, specification of predictors and flattening, model training, testing, and evaluation. (A) Data were extracted from the electronic health records. (B) Potential CVD was identified. (C) The dataset obtained is split geographically into an independent training dataset (85%) and test dataset (15%) with no patient being present in both groups. (D) Prediction times were removed if their lookbehind window extended beyond the start of the dataset or their lookahead extended beyond the end of the dataset. Prediction times were also removed after a patient developed CVD. (E) Predictors were grouped. (F) Predictors for each prediction time were extracted by aggregating the variables within the lookbehind with multiple aggregation functions. As a result, each row in the dataset represents a specific prediction time with a column for each predictor. (G) Predictor layers were added until model performance no longer improved. (H) Models were trained and optimised on the training set using five-fold cross-validation. Hyperparameters were tuned to optimise AUROC. (I) The best candidate model was evaluated on the independent test set. True positive predictions were those with predicted probabilities above the decision threshold and the patient having a CVD event within the lookahead window. False positive predictions were those where the model's predicted probability was above the decision threshold, but the patient did not have a CVD event within the lookahead window. False negatives had predicted probabilities below the threshold, but the patient had a CVD event within the lookahead window. True negatives had predicted probabilities below the threshold, and the patient did not have a CVD event within the lookahead window.

with the data available for implementation, patients exhibiting such an absence should be retained in the dataset. During model training, these absent values are either passed on directly (XGBoost) or imputed using the population median (logistic regression).

### *Predictor addition by early stopping*

The predictors were rank ordered into eight layers (see eTable 2). Models were trained incrementally, adding layers until discrimination stabilised ($\Delta$AUROC < 0.01) for the last two layers. The best-performing layer with the fewest features was further refined by incorporating additional aggregation methods (min, max, and mean) and lookbehind windows (90, 365, and 730 days). See the Supplementary Material for further details.

### *Model selection and hyperparameter tuning*

We focused on two models: XGBoost and elastic net regularised logistic regression, due to the large number of possible model configurations (Figure 1G). XGBoost was selected for its, fast training, and ability to handle numerical, categorical, and missing values internally, and due to the fact that gradient boosting methods generally outperform other machine learning approaches on tabular data [23, 24]. As simpler models are more interpretable and easier to implement, logistic regression with elastic net regularisation was included as a benchmark model. Logistic regression requires missing value imputation as part of pre-processing, and we imputed using the median. For the elastic net penalisation to not be affected by predictor units, we Z-score standardised all predictors for the logistic regression. All predictors listed under "Predictor grouping and flattening" were considered for the XGBoost and elastic net regularised logistic regression. As a sensitivity analysis, we trained an elastic net regularised logistic regression using only predictors that mimic those from SCORE2 as closely as possible with the available data (see Supplementary Table 1 for the specific predictors). All models were trained using five-fold cross-validation, with hyperparameter optimisation to maximise the area under the receiver operating characteristic curve (AUROC) using the tree-structured Parzen estimator algorithm in Optuna v2.10.1 (Figure 1H). Additional details, including which hyperparameters were explored, are provided in Supplementary Material.

### *Model evaluation*

The model that achieved the best AUROC on the training dataset was evaluated on the geographically independent (external) test dataset (Figure 1I). Performance metrics, including AUROC, sensitivity, specificity, positive predictive value, and negative predictive value, were calculated. Since healthcare systems are limited by available resources, and can accommodate different amounts of interventions, performance metrics were calculated for different predicted positive rates [25]. The predicted positive rate is the proportion of all prediction times which are marked as "positive." The mean time from the first positive prediction until a patient met the definition of CVD was also determined. Predictor importance was estimated using information gain.

### *Robustness analyses*

The stability of model prediction was assessed across patient sex, age, as well as time from first visit, and month of year.

### *Post-hoc analyses*

A model using the best performing hyperparameters was re-evaluated on a random split of the entire dataset. All patients were randomly allocated (85–15%) to either the training (85%) or test set (15%), ensuring no patient overlap between the splits. This analysis assessed the performance in the case where all application sites were included in the training data.

### *Ethics*

The use of electronic health record data for this study was approved by the Legal Office of the Central Denmark Region in accordance with the Danish Health Care Act §46, Section 2. According to the Danish Committee Act, ethical review board approval is not required for studies based solely on data from electronic health records (waiver for this project: 1-10-72-1-22). 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 data steward.

### *Data and code sharing*

The code for all analyses is available on GitHub: https://github.com/Aarhus-Psychiatry-Research/psycop-common/tree/319b3ade23ce7eb52af5c9689b2a755ee3f9449e/psycop/projects/cvd

## Results

The eligible cohort consisted of 27,954 patients with a total of 364,791 psychiatric service contacts (prediction times). Demographic and clinical information on the cohort is reported in Table 1. Patients in the train- and test data were broadly similar, with median ages of 35.2 and 35.9 years, and proportions of females of 54.9 and 58.0%, respectively. Among the 27,954 patients, 524 (2.0%) experienced a CVD event. The incidence of CVD was slightly higher in the test data compared to the training data (2.2% vs. 1.8%). The incidence of CVD spiked around the end of the wash-out period, after which it declined (eFigure 2). For each predictor, the proportion of prediction times using the fallback value is described in eTable 3.

Figure 2A presents the results of the model training. The XGBoost model using only predictor layers 1 + 2 (sex, age, LDL, systolic blood pressure, smoking [pack-years], and smoking [daily/occasionally/prior/never]) achieved an AUROC of 0.84 (95% CI: 0.83; 0.84). Incorporating additional lookbehinds or aggregation methods did not enhance model performance. Furthermore, the inclusion of further predictor layers did not increase the AUROC materially or statistically significantly (see eTable 4). The SCORE2-like elastic net regularised logistic regression model performed comparably, with an AUROC of 0.83 (95% CI: 0.83; 0.83).

Figure 2B shows the results for the XGBoost model with a 5-year lookahead window applied to the test data. It achieved an AUROC of 0.74 (95% CI: 0.73; 0.75). Figure 2C shows the resulting confusion matrix at a predicted positive rate of 5% with a positive predictive value of 5% and a negative predictive value of 99%, reflecting that for every twenty positive predictions, one prediction was followed by CVD within 5 years. At this predicted positive rate, the sensitivity at the level of prediction times (contacts to the Psychiatric Services) was 19%, and 39% of all patients who developed CVD were predicted positive at least once (Table 2). Figure 2C

**Table 1.** Descriptive statistics for service contacts (A) and patients (B) that were eligible for prediction

| A. Service contacts | | | |
|---|---|---|---|
| | | Train | Test |
| Service contacts, *n* | | 310,127 | 54,664 |
| ***Demographics*** | | | |
| Age, median [Q1,Q3] | | 35.2 [25.9,46.7] | 35.9 [25.1,47.3] |
| Female, *n* (%) | | 185,681 (59.9) | 34,579 (63.3) |
| Smoking (pack-years), mean (SD) | | 30.5 (75.3) | 25.1 (92.8) |
| Smoking (daily/ occasionally/prior/ never), median [Q1,Q3] | | 2.0 [1.0,4.0] | 3.0 [1.0,4.0] |
| BMI, median [Q1,Q3] | | 25.6 [22.1,30.2] | 25.7 [22.0,30.2] |
| Height (cm), median [Q1,Q3] | | 171.0 [165.0,178.5] | 170.8 [165.0,178.0] |
| Weight (kg), median [Q1,Q3] | | 77.0 [64.5,91.4] | 76.5 [63.9,91.2] |
| ***Diagnoses*** | | | |
| Angina, *n* (%) | | 2,355 (0.8) | 355 (0.6) |
| Atrial fibrillation, *n* (%) | | 1,822 (0.6) | 453 (0.8) |
| Chronic kidney failure, *n* (%) | | 805 (0.3) | 149 (0.3) |
| Chronic lung disease, *n* (%) | | 2,307 (0.7) | 819 (1.5) |
| F0 – Organic disorders, *n* (%) | | 8,357 (2.7) | 1,245 (2.3) |
| F1 – Substance abuse, *n* (%) | | 32,767 (10.6) | 4,387 (8.0) |
| F2 – Psychotic disorders, *n* (%) | | 49,889 (16.1) | 6,171 (11.3) |
| F3 – Mood disorders, *n* (%) | | 115,999 (37.4) | 20,048 (36.7) |
| F4 – Neurotic and stress-related, *n* (%) | | 94,095 (30.3) | 13,865 (25.4) |
| F5 – Eating and sleeping disorders, *n* (%) | | 13,689 (4.4) | 2,068 (3.8) |
| F6 – Personality disorders, *n* (%) | | 47,249 (15.2) | 7,185 (13.1) |
| F7 – Mental retardation, *n* (%) | | 5,778 (1.9) | 320 (0.6) |
| F8 – Developmental disorders, *n* (%) | | 9,584 (3.1) | 1,687 (3.1) |
| F9 – Child and adolescent disorders, *n* (%) | | 45,151 (14.6) | 11,018 (20.2) |
| Type 1 diabetes, *n* (%) | | 1,865 (0.6) | 308 (0.6) |
| Type 2 diabetes, *n* (%) | | 6,291 (2.0) | 1,009 (1.8) |
| ***Lab results*** | | | |
| HDL, mean (SD) | | 1.4 (0.4) | 1.4 (0.4) |
| HbA1c, mean (SD) | | 35.7 (7.0) | 35.2 (6.9) |
| LDL, mean (SD) | | 2.9 (0.9) | 2.9 (0.9) |
| Systolic blood pressure, median [Q1,Q3] | | 126.8 [117.5,137.8] | 125.2 [117.0,136.0] |
| Total cholesterol, mean (SD) | | 4.9 (1.0) | 4.8 (1.0) |
| ***Medications*** | | | |
| Antihypertensives, *n* (%) | | 692 (0.2) | 70 (0.1) |
| Top 10 weight-gaining antipsychotics, *n* (%) | | 74,900 (24.2) | 10,709 (19.6) |
| ***Outcomes*** | | | |
| Incident CVD, *n* (%) | | 2,885 (0.9) | 721 (1.3) |
| By subtype, *n* (group-%) | CABG | 15 (0.5) | 8 (1.0) |
| | MI | 608 (18.8) | 75 (9.3) |
| | PAD | 82 (2.5) | 70 (8.7) |
| | PCI | 626 (19.3) | 37 (4.6) |
| | Stroke | 1,909 (58.9) | 618 (76.5) |

Continued

**Table 1.** *Continued*

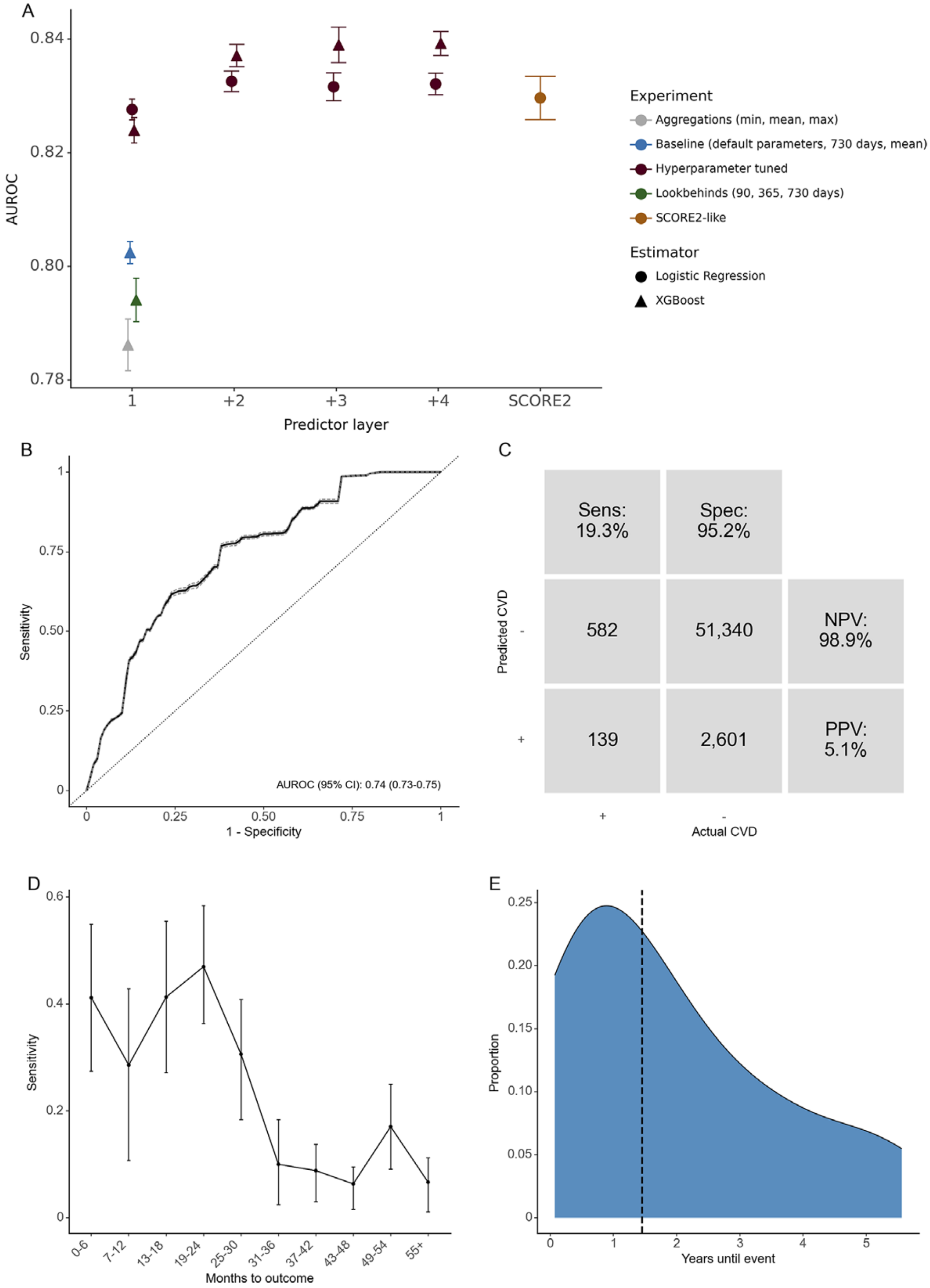
| B. Patients | | | |
|---|---|---|---|
| | | Train | Test |
| Patients, *n* | | 23,584 | 4,370 |
| Female, *n* (%) | | 12,946 (54.9) | 2,535 (58.0) |
| Incident CVD, *n* (%) | | 430 (1.8) | 94 (2.2) |
| By subtype, *n* (group-%) | CABG | 6 (1.4) | <5 |
| | MI | 70 (16.1) | 14 (13.7) |
| | PAD | 13 (3.0) | 8 (7.8) |
| | PCI | 66 (15.2) | 6 (5.9) |
| | Stroke | 280 (64.4) | 73 (71.6) |

*Note*: Cohort demographics by split after preprocessing. For filtering steps, see eFigure 1. Definitions are available in eTable 3. Note that <5 is required by Danish Data Legislation. Abbreviations: CVD, cardiovascular disease; MI, myocardial infarction; PCI, percutaneous coronary intervention; PAD, peripheral artery disease; CABG, coronary artery bypass grafting.

shows that, for patients experiencing a CVD event, the model's probability of flagging them as positive (high risk) increases as the prediction time approaches the CVD event. Figure 2D shows the time from a patient's first positive prediction until they experienced the CVD event. The model marked patients as being at high risk an average of 1.4 years before the CVD event.

Supplementary Table 3 lists prediction by information gain for the best-performing XGBoost model (layers 1 + 2). The most important predictor was age, followed by smoking (daily/occasionally/prior/never), sex, systolic blood pressure, smoking (pack-years), and LDL cholesterol.

Figure 3 highlights that the model was stable across sex, age, and month of year. When calculating model performance within specific age bins, it dropped markedly, which is expected given the relative importance of increasing age for prediction. The model performance also dropped somewhat for patients having been in the system for longer, perhaps indicating a decreasing predictor-sampling frequency over time (most diagnostic workups in the initial hospital contacts).



| | | | |
|---|---|---|---|
| | Sens: 19.3% | Spec: 95.2% | |
| Predicted CVD - | 582 | 51,340 | NPV: 98.9% |
| + | 139 | 2,601 | PPV: 5.1% |
| | + | - | |
| | Actual CVD | | |

**Figure 2.** Results from model training of all models (A) and on geographically independent (external/test) data (B–E). (A) Results of experiments across aggregation methods (mean vs. min, mean, and max), lookbehinds (730 days vs. 90, 365, and 730 days), predictor layers (1, +2, +3, +4), and hyperparameter tuning. Note that results for each layer also include the features of the prior layers. (B) Receiver operating characteristics (ROC) curve. (C) Confusion matrix. PPV, positive predictive value; NPV, negative predictive value. (D) Sensitivity by months from prediction time to event, stratified by desired predicted positive rate (PPR). Note that the numbers do not match those in Table 1, since all prediction times with insufficient lookahead distance have been dropped. (E) Time (months) from the first positive prediction to the patient developing CVD at a 5% predicted positive rate (PPR).

**Table 2.** Performance by predicted positive rate for the best performing model (XGBoost) with 5 years of lookahead on the test set

| Predicted positive rate | True prevalence | PPV | NPV | Sensitivity | Specificity | FPR | FNR | Accuracy | TP | TN | FP | FN | % of all patients with CVD captured | Median years from first positive to CVD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.0% | 1.3% | 5.6% | 98.7% | 1.0% | 95.7% | 4.3% | 99.0% | 97.8% | 31 | 53,417 | 524 | 690 | 7.4% | 2.7 |
| 5.0% | | 5.1% | 98.9% | 4.8% | 80.7% | 19.3% | 95.2% | 94.2% | 139 | 51,340 | 2,601 | 582 | 39.4% | 2.5 |
| 10.0% | | 3.3% | 98.9% | 9.8% | 75.2% | 24.8% | 90.2% | 89.3% | 179 | 48,647 | 5,294 | 542 | 48.9% | 2.6 |
| 20.0% | | 3.6% | 99.2% | 19.6% | 45.6% | 54.4% | 80.4% | 80.1% | 392 | 43,383 | 10,558 | 329 | 70.2% | 2.8 |

*Note*: *Predicted positive rate*: 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 contacts that qualified for CVD within the lookahead window. Numbers are service contacts. *% of all patients with CVD captured*: Percentage of all patients who developed CVD, who had at least one positive prediction. *Median years from first positive to CVD*: For all patients with at least one true positive, the number of years from their first positive prediction to having developed CVD.

Abbreviations: PPV, positive predictive value; NPV, negative predictive value; FPR, false positive rate; FNR, false negative rate; TP, true positives; TN, true negatives; FP, false positives; FN, false negatives.

### Post-hoc analyses

When training (85% split) and evaluating (15% split) the model on a random split of the entire dataset, it obtained an AUROC of 0.84 on the test data, identical to the cross-validated performance in the training data.

## Discussion

In this study, we explored the feasibility of developing a machine learning model trained on routine clinical data from electronic health records to predict the development of CVD in patients with mental illness. An XGBoost model based only on layers 1 + 2 (sex, age, LDL, systolic blood pressure, smoking [pack-years], and smoking [daily/occasionally/prior/never]) achieved an AUROC of 0.74 in the test set at the level of individual service contacts, with a PPV of 5% and an NPV of 99%. For patients who developed CVD and were identified by the model, the median time from initial positive prediction to CVD diagnosis was 1.4 years. This relatively simple model, in which the predictors overlap substantially with those from SCORE2, offers easy implementation in psychiatric services with less comprehensive electronic health record systems [26]. Notably, in spite of the theoretical improvements stemming from the use of machine learning, logistic regression with elastic net penalisation performed as well as the more complex XGBoost. This implies that, for prediction of CVD with a well-established aetiology, simpler models may be sufficient.

A substantial decline in model performance was observed when evaluating the test set (from an AUROC of 0.84 during cross-validation on the training set to an AUROC of 0.74 on the test set). Of note, the training and test sets comprised data from different psychiatric hospitals within the Psychiatric Services of the Central Denmark region. This suggests that substantial distribution shifts can occur even within a relatively homogeneous population sharing geographical proximity, healthcare infrastructure, and clinical protocols, which is further supported by the relative lack of performance difference between training and test when performing a random split of the data (from an AUROC of 0.84 during cross-validation on the training set to an AUROC of 0.84 on the test set). These shifts may be due to variations in patient demographics and/or in data collection between hospitals – despite geographical proximity. More broadly, this lends credence to the argument that external validation should not be considered an absolute prerequisite for scientific publication or model evaluation. Instead, it is proposed that models should undergo rigorous testing within the specific population which they are targeting [27].

Adding information on psychiatric diagnoses by subchapter and antipsychotics (predictor layer 4) did not improve predictive performance. We hypothesise that this is either due to the relatively crude granularity with which these predictors were included, or that their effects are mediated by predictors were already included in the model (e.g., LDL, systolic blood pressure, HbA1c). If diagnoses and antipsychotics affect CVD risk mostly through these variables, they will add no further information. Moreover, the use of antipsychotics results in better treatment of the underlying disease, perhaps resulting in more health-promoting behaviour. In observational studies, antipsychotic use is associated with a lower risk of cardiovascular mortality [7].

To the best of our knowledge, this is the first study to predict the onset of CVD specifically in patients with mental illness based on routine clinical EHR data from psychiatric services. Consequently, comparisons can only be made to studies from other settings/populations. Osborn et al. trained a CVD prediction model specifically for patients with severe mental illness in a primary care setting, including diagnoses and use of antipsychotics as potential predictors [10]. The final model (PRIMROSE) was based on age, gender, height, weight, systolic blood pressure, diabetes, smoking, body mass index, lipid profile, social deprivation, severe mental illness diagnosis, prescriptions of antidepressants, antipsychotics, and reports of heavy alcohol use. It achieved a C-statistic of 0.78, compared to 0.76 of the Framingham risk score (including weights from age, sex, current smoking, total cholesterol, HDL cholesterol, systolic blood pressure, and blood pressure medications). Quadackers et al. compared multiple model's absolute risk estimates for psychiatric inpatient populations, namely SCORE (blood pressure, age, sex, smoking, total cholesterol, and geographical region), the Framingham risk score, and PRIMROSE (described above) [28]. They found very low agreement between the methods, with the Framingham risk score estimating risks 5–10 times higher than SCORE, arguing that it overestimates risk because the risk of CVD was higher at the time of model development than it is now. This indicates the need for re-calibrating models if they are used in markedly different populations than those in which they were developed – one example being patients with mental illness.

Outside the context of patients with mental illness/psychiatric services, a recent meta-analysis found 16 studies comparing machine learning models to traditional statistical models for prediction of CVD [19]. In aggregate, the point estimate of the machine learning methods was marginally better, with a C-statistic of

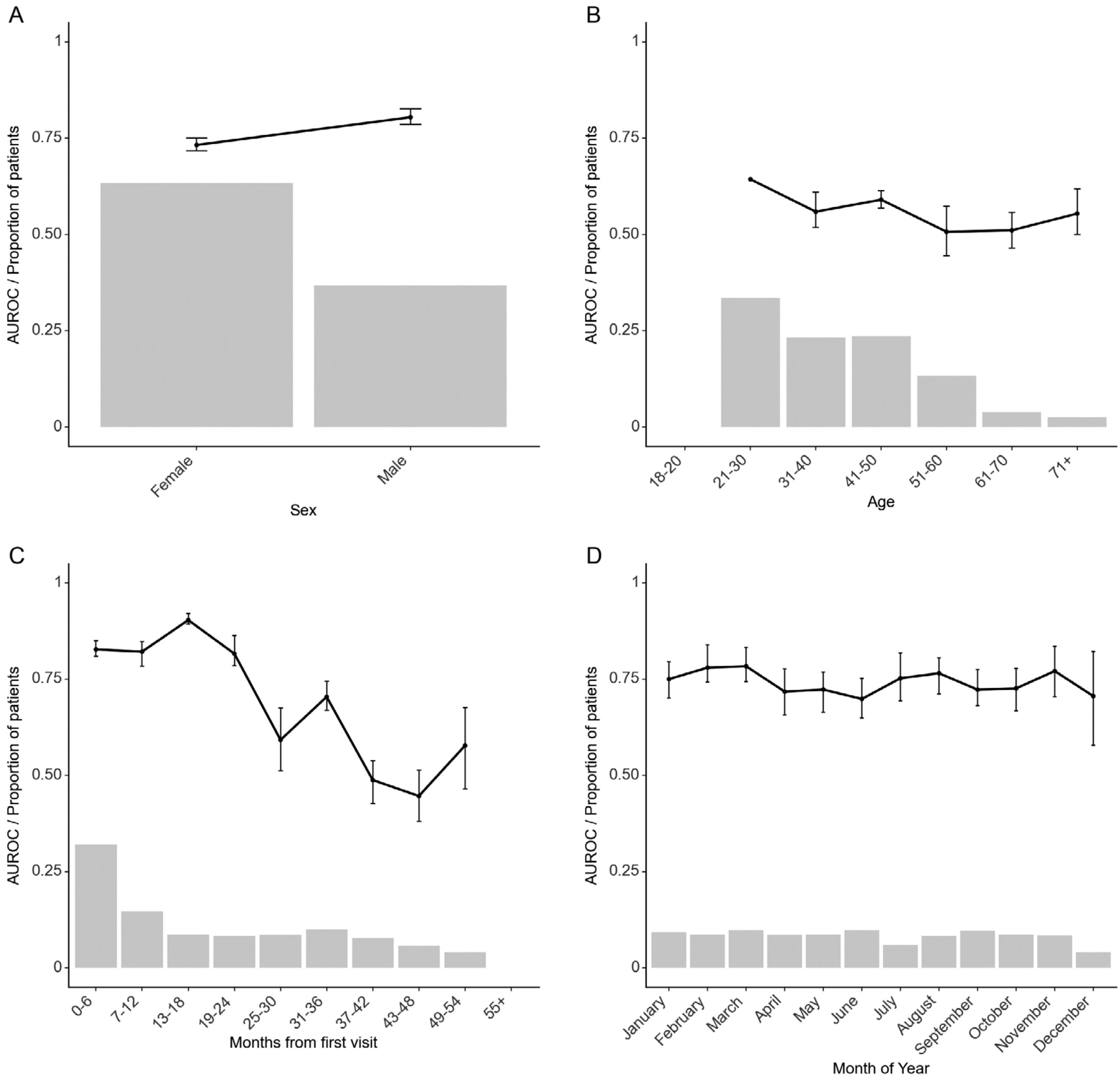


**Figure 3.** Robustness of the best performing model on geographically independent (external/test) data. Robustness of the model across stratifications. The line is the area under the receiver operating characteristics curve. Bars represent the proportion of prediction times in each bin. Error bars are 95%-confidence intervals from 100-fold bootstrap.

0.77 (0.74–0.81) versus 0.76 (0.73–0.79) for traditional statistical models. However, they also find that their implementation is rare and uncertain, arguing that “the impact of missing or unavailable variables and different baseline characteristics on model performance when applied cross-institutionally is unclear.” Indeed, implementing a model based on research cohorts can be challenging, because information on predictors is often not collected as part of routine clinical care, and/or the model assumes that all predictors are available at the time(s) of prediction. We intentionally used only readily available routine clinical data from electronic health records.

If the model developed in this study were to be implemented in the Psychiatric Services of the Central Denmark Region, positive CVD predictions could be automatically presented to healthcare staff via the EHR system, enabling them to initiate appropriate interventions at the level of the individual patient. The specific interventions will depend on the situation. As a first step, more information should typically be gathered, including blood pressure, and a full cardiovascular risk profile. Based on these measurements, patients should be treated according to guidelines [29]. Notably, lifestyle interventions do not appear to be cost-effective in this population, with a large randomised trial of patients with schizophrenia finding no effect [30, 31], and a meta-analysis of trials finding only a clinically insignificant change to BMI ($-0.63$ kg/m$^2$) [32]. Pharmacological interventions, such as statins and antihypertensive drugs, may be more successful, as they require smaller changes to daily life. Another candidate, smoking cessation medication (e.g., bupropion), is as effective among patients with severe mental illness as in the general population, but underutilised [29, 33].

There are limitations to this study that should be considered by the reader. First, prevalent cases of CVD can be misclassified as incident, leading to a false spike in incidence at the beginning of the follow-up period. We mitigated this by employing a 2-year wash-in period. We found that, for most CVD events, incidence was decreasing after the wash-in period. There are multiple potential reasons for this finding. Specifically, it may reflect a true drop in incidence as studies show decreasing incidence rates of CVD in Denmark, but these drops are insufficient to fully explain the trend [34, 35]. As such, it cannot be ruled out that some part of the events we detect are prevalent cases. This is, however, unlikely to cause harm to patients, as prevalent cases also need prevention of further events, but it may have inflated the prediction estimates. Second, this study does not address potential effects of implementing the developed model. When prediction models are implemented, they should affect behaviour, for example by inducing further testing or treatment. Specifically, implementing a CVD prediction model would likely induce more relevant LDL- and blood-pressure measurements. These model-induced measurements should improve the next prediction issued by the model, meaning that predictions following a positive prediction are likely less accurate in the present dataset than they would be following implementation. Third, many important variables for CVD, such as physical activity, dietary habits, or waist circumference, are not collected with sufficient regularity as part of current clinical practice and could not be included in the model. If they had been available, the model would likely perform with greater accuracy. Fourth, since, most patients who experienced an event in the test set had a stroke (71.6%) the model is less likely to generalise to cohorts where stroke is less prevalent. However, given that the important features of the model are very general CVD features, we would expect meaningful generalisation. Finally, machine learning models vary markedly in their generalisability. We used routine clinical data from a system with universal healthcare and observed performance differences between departments within the same regional Psychiatric Services. Therefore, direct transfer of the model to other healthcare systems would probably yield suboptimal predictions. However, the approach is likely to be generalisable, and retraining the model on data from other settings using the same architecture may allow for transferability.

In conclusion, a machine learning model trained on routine clinical data from electronic health records can predict the development of CVD among patients with mental illness at a level that may make clinical implementation as a decision support tool feasible. Specifically, the model may help clinicians identifying which patients will benefit from primary preventative initiatives. Moving forward, we see two main tasks arising from this work. First, we will work towards testing the feasibility of implementing the model as a clinical decision support tool in the Psychiatric Services of the Central Denmark Region. Second, as we believe the model may hold potential for broader application, we aim to conduct external validation in independent samples.

**Supplementary material.** The supplementary material for this article can be found at http://doi.org/10.1192/j.eurpsy.2025.1.

**Data availability statement.** According to Danish law, the personally sensitive data used in this study is only available for research projects conducted by employees in the Central Denmark Region following 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 assistance with the extraction of data and Mathias Brønd Sørensen from the Business Intelligence Office, Central Denmark Region, for assistance in building the infrastructure for model training.

**Author contribution.** The study was conceptualised and designed by MB, KKWO, AAD, and SDØ. The coding and statistical analyses were carried out by MB with assistance from LH. All authors contributed to the interpretation of the results. MB 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.

**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), the Danish Agency for Digitisation Investment Fund for New Technologies (grant number 2020-6720), and by CAAIR; Collaboration on Applied Al Research and Digital Innovation in Healthcare. SDØ 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 numbers: 7016-00048B and 2096-00055A). KKWO is supported by a grant from the Danish Cardiovascular Academy (grant no. CPD5Y-2022001-HF), which is funded by the Danish Heart Association and the Novo Nordisk Foundation. The funders played no role in study design, collection, analysis, or interpretation of data, the writing of the report, or the decision to submit the paper for publication.

**Competing interest.** Danielsen has received a speaker honorarium from Otsuka Pharmaceutical. SDØ received the 2020 Lundbeck Foundation Young Investigator Prize. SDØ 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, USPY, EXH2, 2B76, IS4S, OM3X, and EUNL. The remaining authors report no conflicts of interest.

## References

[1] Roth Gregory A, Mensah George A, Johnson Catherine O, Giovanni A, Enrico A, Baddour Larry M, et al.. Global burden of cardiovascular diseases and risk factors, 1990–2019. J Am Coll Cardiol. 2020;76(25): 2982–3021.

[2] Erlangsen A, Andersen PK, Toender A, Laursen TM, Nordentoft M, Canudas-Romo V Cause-specific life-years lost in people with mental disorders: a nationwide, register-based cohort study. Lancet Psychiatry. 2017;4(12):937–45.

[3] Solmi M, Fiedorowicz J, Poddighe L, Delogu M, Miola A, Høye A, et al. Disparities in screening and treatment of cardiovascular diseases in patients with mental disorders across the world: systematic review and meta-analysis of 47 observational studies. Am J Psychiatry. 2021;178(9): 793–803.

[4] Rødevand L, Steen NE, Elvsåshagen T, Quintana DS, Reponen EJ, Mørch RH, et al. Cardiovascular risk remains high in schizophrenia with modest improvements in bipolar disorder during past decade. Acta Psychiatr Scand. 2019;139(4):348– 60.

[5] Scott D, Happell B The high prevalence of poor physical health and unhealthy lifestyle behaviours in individuals with severe mental illness. Issues Ment Health Nurs. 2011;32(9):589–97.

[6] Rohde C, Köhler-Forsberg O, Nierenberg AA, Østergaard SD. Pharmacological treatment of bipolar disorder and risk of diabetes mellitus: a nationwide study of 30,451 patients. Bipolar Disord. 2023 Jun;25(4):323–334.

[7] Taipale H, Tanskanen A, Mehtälä J, Vattulainen P, Correll CU, Tiihonen J. 20-year follow-up study of physical morbidity and mortality in relationship to antipsychotic treatment in a nationwide cohort of 62,250 patients with schizophrenia (FIN20). World Psychiatry. 2020;19(1):61–8.

[8] Mitchell AJ, Delaffon V, Vancampfort D, Correll CU, Hert MD. Guideline concordant monitoring of metabolic risk in people treated with antipsychotic medication: systematic review and meta-analysis of screening practices. Psychol Med. 2012;42(1):125–47.

[9] Nasrallah HA, Meyer JM, Goff DC, McEvoy JP, Davis SM, Stroup TS, et al. Low rates of treatment for hypertension, dyslipidemia and diabetes in schizophrenia: Data from the CATIE schizophrenia trial sample at baseline. Schizophr Res. 2006;86(1–3):15–22.

[10] Osborn DPJ, Hardoon S, Omar RZ, Holt RIG, King M, Larsen J, et al.. Cardiovascular risk prediction models for people with severe mental illness: results from the prediction and management of cardiovascular risk in people with severe mental illnesses (PRIMROSE) research program. JAMA Psychiatry. 2015;72(2):143–51.

[11] Song X, Mitnitski A, Cox J, Rockwood K. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. MEDINFO 2004. 2004;107(pt 1):736–40.

[12] Danielsen AA, Fenger MHJ, Østergaard SD, Nielbo KL, Mors O. Predicting mechanical restraint of psychiatric inpatients by applying machine learning on electronic health data. Acta Psychiatr Scand. 2019;140(2):147–57.

[13] Cahn A, Shoshan A, Sagiv T, Yesharim R, Goshen R, Shalev V, et al. Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. Diabetes Metab Res Rev. 2020; 36(2):e3252.

[14] Bernstorff M, Hansen L, Enevoldsen K, Damgaard J, Hæstrup F, Perfalk E, et al.. Development and validation of a machine learning model for prediction of type 2 diabetes in patients with mental illness. Acta Psychiatr Scand. 2024;acps.13687.

[15] Hansen L, Enevoldsen K, Bernstorff M, Perfalk E, Danielsen AA, Nielbo KL, et al.. Lexical stability of psychiatric clinical notes from electronic health records over a decade. Acta Neuropsychiatr. 2023;1–11.

[16] Bernstorff M, Hansen L, Perfalk E, Danielsen AA, Østergaard SD. 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 Psychiatr Scand. 2022;146(3):272–83.

[17] SCORE2 Working Group and ESC Cardiovascular Risk Collaboration, Hageman S, Pennells L, Ojeda F, Kaptoge S, Kuulasmaa K, et al.. SCORE2 risk prediction algorithms: New models to estimate 10-Year risk of cardiovascular disease in Europe. Eur Heart J. 2021;42(25):2439–54.

[18] Kim MS, Hwang J, Yon DK, Lee SW, Jung SY, Park S, et al.. Global burden of peripheral artery disease and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Glob Health. 2023;11(10):e1553–65.

[19] Liu W, Laranjo L, Klimis H, Chiang J, Yue J, Marschner S, et al.. Machine-learning versus traditional approaches for atherosclerotic cardiovascular risk prognostication in primary prevention cohorts: a systematic review and meta-analysis. Eur Heart J – Qual Care Clin Outcomes. 2023;9(4):310–22.

[20] Rotella F, Cassioli E, Calderani E, Lazzeretti L, Ragghianti B, Ricca V, et al.. Long-term metabolic and cardiovascular effects of antipsychotic drugs. A meta-analysis of randomized controlled trials. Eur Neuropsychopharmacol. 2020;32:56–65.

[21] WHOCC - ATC/DDD Index [Internet]. [cited 2023 Apr 12]. Available from: https://www.whocc.no/atc_ddd_index/

[22] Bernstorff M, Enevoldsen K, Damgaard J, Danielsen A, Hansen L. timeseriesflattener: A Python package for summarizing features from (medical) time series. J Open Source Softw. 2023;8(83):5197.

[23] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining [Internet]. 2016 [cited 2023 Feb 17]. p. 785–94. Available from: http://arxiv.org/abs/1603.02754

[24] Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? [Internet]. arXiv; 2022 [cited 2023 Feb 17]. Available from: http://arxiv.org/abs/2207.08815

[25] Wornow M, Xu Y, Thapa R et al. The shaky foundations of large language models and foundation models for electronic health records. npj Digit. Med. 2023; 6: 135.

[26] Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. Machine learning: the high interest credit card of technical debt. In: SE4ML: software engineering for machine learning (NIPS 2014 Workshop). 2014.

[27] Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. Diagn Progn Res. 2022;6(1):24.

[28] Quadackers D, Liemburg E, Bos F, Doornbos B, Risselada A, PHAMOUS investigators, et al.. Cardiovascular risk assessment methods yield unequal risk predictions: a large cross-sectional study in psychiatric secondary care outpatients. BMC Psychiatry. 2023;23(1):536.

[29] Polcwiartek C, O'Gallagher K, Friedman DJ, Correll CU, Solmi M, Jensen SE, et al.. Severe mental illness: cardiovascular risk assessment and management. Eur Heart J. 2024;45(12):987–97.

[30] Speyer H, Christian Brix Nørgaard H, Birk M, Karlsen M, Storch Jakobsen A, Pedersen K, et al. The CHANGE trial: no superiority of lifestyle coaching plus care coordination plus treatment as usual compared to treatment as usual alone in reducing risk of cardiovascular disease in adults with schizophrenia spectrum disorders and abdominal obesity. World Psychiatry Off J World Psychiatr Assoc WPA. 2016;15(2):155–65.

[31] Jakobsen AS, Speyer H, Nørgaard HCB, Karlsen M, Birk M, Hjorthøj C, et al.. Effect of lifestyle coaching versus care coordination versus treatment as usual in people with severe mental illness and overweight: two-years follow-up of the randomized CHANGE trial. PLOS One. 2017;12(10): e0185881.

[32] Speyer H, Jakobsen AS, Westergaard C, Nørgaard HCB, Pisinger C, Krogh J, et al.. Lifestyle interventions for weight management in people with serious mental illness: a systematic review with meta-analysis, trial sequential analysis, and meta-regression analysis exploring the mediators and moderators of treatment effects. Psychother Psychosom. 2019;88(6):350–62.

[33] Tsoi DT yin, Porwal M, Webster AC. Efficacy and safety of bupropion for smoking cessation and reduction in schizophrenia: systematic review and meta-analysis. Br J Psychiatry. 2010;196(5):346–53.

[34] Skajaa N, Adelborg K, Horváth-Puhó E, Rothman KJ, Henderson VW, Casper Thygesen L, et al.. Nationwide trends in incidence and mortality of stroke among younger and older adults in Denmark. Neurology. 2021; 96(13):e1711–23.

[35] Schmidt M, Andersen LV, Friis S, Juel K, Gislason G. Data resource profile: Danish heart statistics. Int J Epidemiol. 2017;46(5):1368–9g.