Quality and Reliability Engineering International

WILEY

CASE STUDY OPEN ACCESS

Prediction of Depression Relapse Using Machine Learning With Administrative Data: Balancing Complexity and Simplicity

Paulina von Stackelberg[1] | Rob Goedhart[1] | Leo C. E. Huberts[2] | Joran Lokkerbol[3] | Ş. İlker Birbil[1]

[1]Amsterdam Business School, Business Analytics, University of Amsterdam, Amsterdam, the Netherlands | [2]Centre for Big Data Research in Health, University of New South Wales, Sydney, New South Wales, Australia | [3]Centre of Economic Evaluations & Machine Learning, Trimbos Institute, Netherlands Institute of Mental Health and Addiction, Utrecht, the Netherlands

Correspondence: Paulina von Stackelberg (p.b.vonstackelberg@uva.nl)

Received: 29 March 2025 | **Revised:** 27 November 2025 | **Accepted:** 20 December 2025

Keywords: depression prediction | machine learning | mental health | patient monitoring

ABSTRACT

Depression is a mental disorder with a high lifetime prevalence and one of the leading causes of disability worldwide. As many patients experience another depressive episode after being treated, predictive monitoring for the risk of relapse is essential for healthcare professionals to be able to follow up on patients and intervene early. However, automatically monitoring these large groups requires additional considerations going beyond predictive performance, such as data availability and interpretability. In the present paper, we study the suitability of using readily available administrative data for this prediction task. We contrast a logistic regression model containing only a small number of predictors on demographics, medication, and estimated depression severity with regularized regression and XGBoost models incorporating a large number of predictors describing individual treatment and social information. Our results demonstrate that the inclusion of more detailed input does not result in a significant improvement in performance when compared to simpler regression models. In similar data types, we therefore recommend to primarily focus on a small interpretable model.

1 | Background

The healthcare field has witnessed a significant increase in available data. This opens new opportunities in predictive monitoring, where the goal is to build early warning systems that signal health crises [1–3]. These methods can inform policymakers and enable early interventions on an individual level, therefore containing severe effects as much as possible. To work with large volumes of information, clinical researchers have started testing the feasibility of using machine learning (ML) methods for medical prediction modeling (see, e.g., [4–7]). Despite the significant progress that has been made in research, the application of ML in clinical settings continues to be a subject of debate. It has been suggested that the improvement in predictive performance is marginal when tabular data are used [8], and the lack of interpretability may pose an issue in high-stakes settings [9, 10]. Monitoring psychiatric disorders in such patient populations is especially challenging. First of all, the outcome cannot be observed explicitly; while conditions like hypertension are directly measurable, mental health issues such as depressive disorders require supplemental information. Second, the group to be monitored can become quite large due to the high lifetime prevalence of mental disorders in the population, requiring easily accessible data to monitor patients at a realistic cost.

Depression is among the most prevalent mental health disorders, with a lifetime risk estimated to exceed 10% [11, 12]. Microsimulation model-based research suggests that the actual prevalence may be even higher [13]. Previous studies have shown that individuals diagnosed with depression are often rediagnosed with the condition at a later point in time [14]; therefore, it has been suggested that preventative treatments should be incorporated into the long-term treatment plan [15]. Knowing who is at high risk of relapse can help healthcare professionals follow up on patients and intervene as soon as possible.

In the current study, we explore the potential of readily available administrative data to monitor the risk of relapsing back into depression treatment. Using a unique administrative dataset containing information about individuals registered in the Netherlands, we focus on the integration of treatment information and social factors within complex models, contrasting the results against a small, interpretable, unregularized regression model using a limited subset of variables.

This paper is structured as follows: First, we give an introduction to depression and describe a number of important predictors for depression relapse as identified from the clinical literature (Section 2). We give an introduction to the data (Section 3.1) and the different models included in our comparison (Section 3.2). In the next sections, we compare (Section 5) and discuss (Sections 6.1–6.3) the performances of the regression and ML models, and finally provide recommendations aiding the future development of large-scale depression monitoring methods (Section 6.4).

2 | Clinical Background: Depression

Depression is a mood condition characterized by persistent feelings of sadness, hopelessness, and a lack of interest in activities that the person normally enjoys. It is often accompanied by a number of psychiatric comorbidities, including anxiety disorders and substance use disorder [16, 17]. Next to the psychological impact, patients may also experience symptoms such as sleeping issues, changes in appetite, and impaired cognition [18]. This can have a significant impact on a person's daily life, including their ability to work, study, and engage in social activities. While the exact mechanism underlying risk factors for depression is not fully understood, it is believed to be caused by a combination of genetic [19], environmental [20], and psychological factors [21, 22]. In addition, research has investigated the association between depression and the probability of developing a number of other medical conditions, such as cardiovascular disease [22].

There are several factors that can lead to an elevated risk of developing depression. Previous research has examined a number of indicators including, for instance, adverse events in childhood, residual depressive symptoms, comorbid anxiety, the number of previous depressed episodes, the severity of the depressed episode, and the use of antidepressants [22–26]. Treatment depends on the severity of symptoms, but is often based on a combination of medication, such as selective serotonin reuptake inhibitors (SSRIs), and psychological approaches (e.g., cognitive behavioral therapy) [22, 27–29]. Untreated, depression can lead to serious complications, including an increased risk of suicide [30].

3 | Methodology

In this section, we describe the dataset, the models, and the evaluation metrics.

3.1 | Dataset Description

3.1.0.1 | Sample. We use an administrative dataset provided by Statistics Netherlands containing information about residents in the Netherlands collected in the period from 2013 up to (and including) 2016. From the initial dataset, we extract individuals who are diagnosed with depression (main diagnosis) and enter treatment at any point in 2013 ($N = 106{,}067$).

Due to computational restrictions, we take a random sample of 30,000 individuals from this group. The mean duration of an individual treatment trajectory registered in 2013 is 264.53 days (SD = 120.72). We exclude 981 individuals below the age of 18 from this sample. The mean age is 47.13 years (SD = 16.05), and 61.97% of the individuals are female (calculated in January 2013). We create a panel dataset where each row equals an observation for person i at time t. Each row represents a 1 month period, resulting in 48 rows per individual. We extract data for patient monitoring at six different time points t with equal spacing: In January 2014 ($t = 1$), May 2014 ($t = 2$), September 2014 ($t = 3$), January 2015 ($t = 4$), May 2015 ($t = 5$), and September 2015 ($t = 6$). We only include individuals who are not in a treatment trajectory for the main diagnosis depression at time point t. The final sample sizes can be found in Figure 1.

3.1.0.2 | Outcome variable. In the current study, we are interested in the probability of an individual experiencing a depression relapse, which we define as starting a new treatment for depression (main diagnosis) within 12 months after ending the previous depression treatment.[1] We base the categorization of being in or out of treatment on the treatment trajectory information available in our administrative data. These data contain information about a person's diagnoses and treatment. From this information, we define the outcome y_i (depression relapse) as follows: We code a treatment as a 'new treatment' when at least 1 month is registered in between two treatment trajectories. If a new treatment for depression is opened at any point within the next 12 months from time t onwards (i.e., the person is in treatment for depression again), we register this as a "relapse". If a patient commits suicide within those 12 months, this is also coded as a relapse. The percentage of cases (individuals entering a new treatment trajectory in the next 12 months) differs between time points ($t = 1$: 5.48%; $t = 2$: 5.96%; $t = 3$: 5.84%; $t = 4$: 5.11%; $t = 5$: 4.21%; $t = 6$: 3.75%).

3.1.0.3 | Predictors. Predictors used in the current study are related to the individual's health status (e.g., healthcare activity, medication prescriptions from the pharmacy, anxiety disorders and other conditions as side diagnosis), as well as social and demographic information (e.g., age, sex, income and benefits, employment situation, whether the individual has been registered with a crime). We sum up the number of hours a patient spends in a particular treatment activity per 6 and per 12 months as we are interested in predicting relapse 12 months ahead, and to strike a balance between limiting the number of

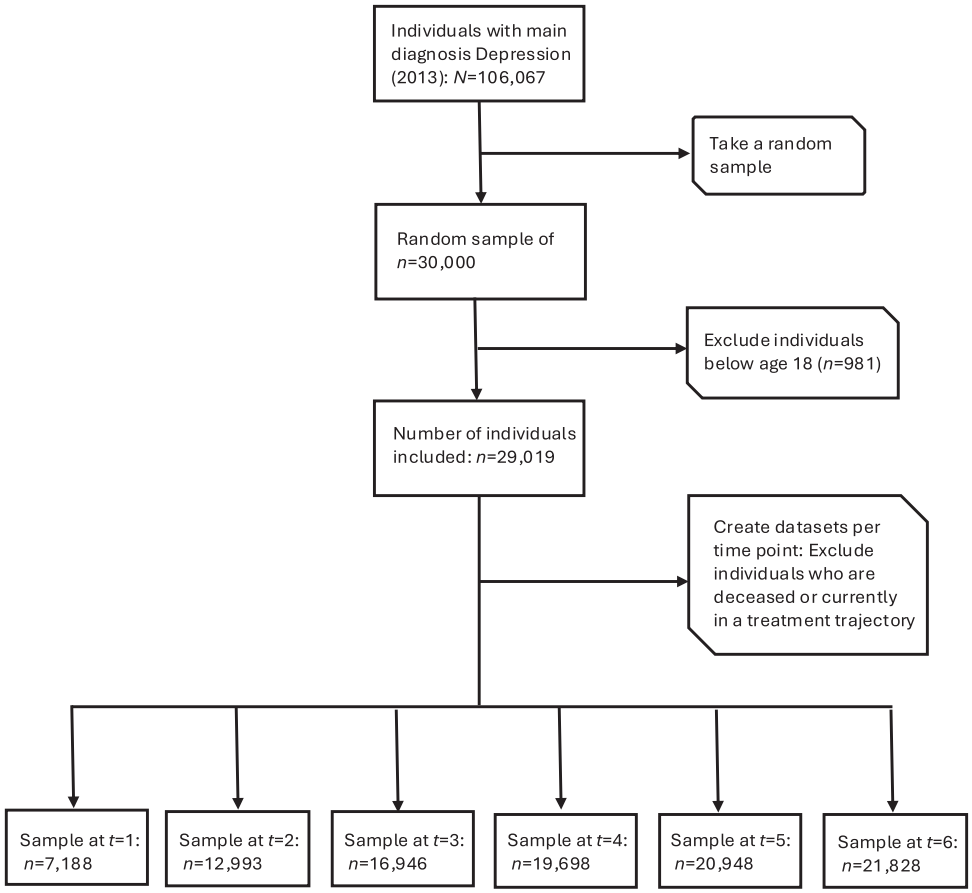

FIGURE 1 | Visual description of the repeated samples in the current study.

variables and still capturing enough past information. Note that the healthcare activities are not restricted to mental healthcare but may also include other types of healthcare; this is used as auxiliary information relating to general health indicators. If the care provided is comprised of days spent in a facility, we do not register the activities in terms of the hours spent in the treatment activity but in terms of days spent at the facility. Furthermore, some information is only available on a yearly basis (i.e., medication-related variables, income- and debt-related variables, crime-related variables, insurance information); to avoid leakage of future information, we use information from the past year for those variables. For treatment variables, we remove duplicate entries (i.e., if two treatments of the same type are registered on the same day for the same patient) (Table 1).

TABLE 1 | Summary of predictor categories and their respective data lookback periods.

Variable category	Data inclusion/Lookback period
Medication	Only previous year included
Treatment	Hours/Days: Current month, sum last 6 months, sum last 12 months
Income, benefits, crime	Only previous year included
Side diagnosis	Indicator current month, indicator last 6 months, indicator last 12 months

3.2 | Models and Evaluation Metrics

A number of different models are fitted to the data. We predict the risk of depression relapse using regression models (regularized and unregularized) as well as XGBoost. For the regression models, we include the following methods: unregularized simple logistic regression (SR)[2], Ridge logistic regression, and Lasso logistic regression. The unregularized logistic regression models serve as a baseline for comparison with the other models, which are able to handle the potential multicollinearity resulting from the large number of available predictors. As regression models require imputation for missingness, we impute the mean value for numerical variables and the mode for factors. The estimation of the mean/mode is based on the training set and subsequently used for imputation in both the training and test set. We choose XGBoost as an effective and scalable ML algorithm [32] which has also proven its use in the analysis of depression [33].

Before internal validation but after selecting individuals as described in the previous section, we exclude predictors where all values are the same. Therefore, the samples between time points have differing numbers of predictors (number of predictors *pred* per time point t after this exclusion: $pred_{t1} = 899$, $pred_{t2} = 1024$, $pred_{t3} = 1003$, $pred_{t4} = 942$, $pred_{t5} = 861$, $pred_{t6} = 822$). Following these steps, we fit the models and perform internal

validation using four-fold cross-validation (CV), stratifying on the outcome variable to ensure an approximately equal proportion of cases and controls across folds. We keep the data splits for CV across models constant. As the individuals monitored differ between time points due to not including individuals currently in treatment, we mainly focus on comparing the models per time point with each other instead of focusing on the between-time point outcomes.

3.2.1 | Models

3.2.1.1 | Unregularized logistic regression. To fit a simple unregularized logistic regression model, we include only a subset of the variables. We test models containing variables about sex, age, income, medication[3] (antidepressants yes/no, number of other nervous system medications, number of other medications as an indicator of general health), measures of estimated depression severity (i.e., (i) number of previous treatment trajectories, (ii) number of months associated with last depression treatment trajectory) and whether the patient has been registered with a side diagnosis for anxiety during one of their depression trajectories in the past 12 months. Furthermore, we include a predictor on the number of months since the last depression treatment trajectory concluded. To reduce the number of predictors further, we perform backwards selection on an independent training (IT) set (excluding 1007 individuals below the age of 18) to avoid overlap (i.e., not including any individuals who are in the datasets used for model training at a later stage). We perform the backwards selection on a subset taken at $t = 1$ ($n_{IT} = 7173$). The predictors identified from this procedure are then used as model input in the original sample. More details about these models can be found in Appendix B.

3.2.1.2 | Regularized logistic regression. To be able to include not just a subset, but all available predictors in the model we additionally fit logistic Ridge [34] and Lasso [35] regression models. Both models work by imposing a penalty on the regression coefficients. Ridge logistic regression uses the L2 penalty, while Lasso logistic regression uses the L1 penalty. The L1 penalty in Lasso logistic regression can shrink coefficients to 0, and therefore the model automatically performs variable selection. For both models, a penalty parameter λ determines the amount of shrinking and is chosen with 10-fold CV in this study, using minimal deviance as a performance criterion [36].

3.2.1.3 | XGBoost. XGBoost [32] is a gradient boosting algorithm using decision trees where new trees are built using the pseudo-residuals of the previously fitted tree to correct the prior error. We use the following tuned parameters: the regularization parameters λ (L2 regularization) and α (L1 regularization), the learning rate η, and the maximum tree depth χ. Due to computational restrictions, we only test 20 combinations of parameters during tuning with random search in `mlr` [37].

3.2.2 | Evaluation Metrics

In the current study, we use three different types of performance evaluation measures: (i) discrimination, (ii) prediction error, and (iii) calibration. To assess the models' discriminative ability, we calculate the AUC. In addition, we compute sensitivity and specificity. To compare the prediction error between models, we calculate the Brier score. We assess calibration performance by calculating the calibration slope (CS). In all cases, we compute the metrics on the test set using the model fitted on the training set. As mentioned above, internal validation is performed by means of four-fold CV.

Let TP denote true positives, FP false positives, TN true negatives, and FN false negatives. Let th denote the threshold chosen to separate classes. For sensitivity and specificity, we test the thresholds included in $th \in \{0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09\}$. For the Brier score, p_i describes the probability of depression relapse within the next 12 months for person i, and y_i describes whether individual i will experience a depression relapse within the next 12 months (binary; yes/no) at time point t. Sensitivity, specificity, and the Brier score are then defined as follows:

$$\text{Sensitivity} = \frac{TP(th)}{TP(th) + FN(th)}; \tag{1}$$

$$\text{Specificity} = \frac{TN(th)}{TN(th) + FP(th)}; \tag{2}$$

$$\text{Brier score} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2. \tag{3}$$

To assess model calibration, we examine the CS, due to it being a useful measure for internal validation [38]. As pointed out by Van Calster et al. [39], calibration-in-the-large is only relevant at external validation, which is not the focus of the current study and therefore not computed. The CS captures the spread of the risks, and shows whether the estimated risk is too extreme or not extreme enough. When the CS is estimated at <1, this means that the risk is estimated too high for samples that are at higher risk, and too low for samples that are at lower risk. When the opposite occurs (i.e., $CS > 1$), this indicates that the spread of risks is too narrow. Models with "perfect calibration" would yield a slope b_{LP} equal to 1 [39]. To obtain the CS, we fit a logistic regression model

$$logit(P(Y = 1)) = b_0 + b_{LP}\,\text{LP}, \tag{4}$$

where LP indicates the linear predictor.

4 | Software

The analyses are performed in R (version 4.4) [40]. For regression models, we use `glm` and `glmnet` [36]. As a modeling framework for the XGBoost models, we use `mlr` [37]. For data preparation and CV, as well as further analyses, we use `tidyverse` [41], `caret` [42], `runner` [43], `pROC` [44], and `haven` [45].

5 | Results

In the paragraphs below, we discuss the results.

TABLE 2 | AUC, Brier score, and calibration slopes for Simple unregularized regression (SR), Lasso regression (LAS), Ridge regression (RID), and XGBoost (XGB).

t	AUC_{SR}	AUC_{LAS}	AUC_{RID}	AUC_{XGB}	$Brier_{SR}$	$Brier_{LAS}$	$Brier_{RID}$	$Brier_{XGB}$	CS_{SR}	CS_{LAS}	CS_{RID}	CS_{XGB}
1	0.659	0.645	0.612	0.655	0.051	0.052	0.052	0.052	0.979	1.562	0.977	0.778
2	0.661	0.662	0.645	0.671	0.055	0.055	0.056	0.055	0.991	1.326	0.960	1.083
3	0.643	0.656	0.633	0.655	0.054	0.054	0.055	0.054	0.989	1.202	0.917	0.774
4	0.653	0.653	0.636	0.659	0.048	0.048	0.048	0.048	0.988	1.248	0.961	0.956
5	0.649	0.649	0.620	0.621	0.040	0.040	0.040	0.040	0.995	1.788	1.986	0.644
6	0.654	0.652	0.634	0.634	0.036	0.036	0.036	0.036	0.987	1.354	1.181	0.732

5.0.0.1 | AUC. In this study, Lasso logistic regression models outperform Ridge logistic regression in terms of the AUC (mean AUC for Lasso: $AUC_{LAS} = 0.653$; mean AUC for Ridge: $AUC_{RID} = 0.63$; Table 2). The performance of the Lasso models is very close to that of the unregularized regression models (mean AUC SR: $AUC_{SR} = 0.653$; Table 2). On average, XGBoost models demonstrate slightly lower performance than Lasso regression (mean AUC for XGBoost: $AUC_{XGB} = 0.649$; Table 2). Brier scores are nearly identical across the four model types at each time point, with a maximum difference of $Brier_{maxdiff} = 0.001$ (e.g., between XGBoost and Ridge at $t = 2$; Table 2). A separation between cases and controls is evident in most empirical cumulative distribution functions (ecdfs), with the most visible shifts in central tendency observed for SR and XGBoost models (Figure 2).

5.0.0.2 | Calibration slope. The unregularized regression models have a mean CS close to 1 (mean CS: $CS_{SR} = 0.988$; range = 0.979–0.995). Lasso logistic regression, by contrast, shows a mean CS above 1 (mean CS: $CS_{LAS} = 1.413$; range = 1.202–1.788), suggesting that the predicted risks are generally too narrow. Ridge logistic regression demonstrates a less consistent pattern, with CSs ranging from 0.917 to 1.986 and a mean CS across time points of $CS_{RID} = 1.164$, which is closer to 1 than that of Lasso. XGBoost models, on the other hand, mostly have CSs below 1 (mean CS: $CS_{XGB} = 0.828$; range = 0.644–1.083), indicating that the predicted probabilities tend to be overly extreme, with low probabilities being underestimated and high probabilities being overestimated (Table 2).

5.0.0.3 | Sensitivity and specificity. We calculate results for a range of thresholds (threshold range = 0.01–0.09). To illustrate the dependencies, consider the following example: Achieving a specificity above 0.7 across all models and time points in this study requires setting a threshold of $th = 0.08$, while achieving a sensitivity of over 0.7 would allow a maximum threshold of $th = 0.02$. As expected, sensitivity decreases within models as the threshold increases, while specificity generally shows the opposite trend, increasing with higher thresholds. To sum up a general pattern, we can see the influence of the threshold choice in terms of the calculated sensitivity and specificity: The strength in difference between the models depends on the cutoff chosen. For instance, at time points $t = 1$, $t = 3$, and $t = 5$, we see that the XGBoost models have a higher specificity than the other models, in particular for lower thresholds (e.g., difference: $spec_{diff} = 0.185$ at $t = 1$ and $th = 0.02$ between XGBoost and SR). For Ridge regression, we mostly observe a lower average sensitivity than for the other models, which is visible at higher threshold values but not at lower thresholds (Table 3).

5.0.0.4 | Threshold choices. In this section, we further elaborate on thresholds. The threshold choices made in practice can reflect a number of different points, and are highly dependent on the application. While one application may, for instance, be associated with a very high cost of missing a case (e.g., not detecting a malignant tumor), this might not be the focus in another situation. To provide some examples on potential decision paths in the current example, we split this section into two parts: First of all, we demonstrate how capacity considerations can guide threshold choices. Second, we highlight the differences in relative model performances based on the thresholds (see also Section 5 "Sensitivity and Specificity") and set this in the context of the current case study.

Threshold based on capacity. One factor that can inform the selection of a final classification threshold is capacity, such as the availability of follow-up appointments. First of all, recall that TP, FP, TN, and FN, as well as related performance metrics, such as sensitivity and specificity, are all functions of the chosen threshold th (Figure 3). In scenarios where the cost of a false positive is relatively small (e.g., if interventions following a positive signal are nonintrusive and nonexpensive), one may choose the threshold based on capacity. Specifically, this can be reduced to

$$FP(th) + TP(th) = C, \tag{5}$$

where C indicates the available capacity for interventions.

As seen in Equation (5), the focus is centered on the individuals classified as positives: For example, assuming a follow-up capacity of 2000 patients, the corresponding threshold for the Lasso model at time point 6 would be between 0.05 and 0.06 (or higher), whereas for the XGBoost model, it would be chosen as between 0.07 and 0.08 or higher (Figure 4).

Threshold based on performance metric. Alternatively, one can choose the threshold tailored to a specific performance metric. For example, von Stackelberg et al. [46] show how a threshold could be chosen to obtain a specified false positive rate. This is beneficial if the cost of intervention is high (e.g., if positive signals are followed by an intrusive or expensive treatment). Similarly, the threshold can be chosen based on any other specified metric, such as the sensitivity or specificity. In the following paragraph,

TABLE 3 | Sensitivity and specificity for simple unregularized regression (SR), Lasso regression (Las), Ridge regression (Rid), and XGBoost (XGB).

		Specificity									Sensitivity								
t	**Model**	**th = 0.01**	**th = 0.02**	**th = 0.03**	**th = 0.04**	**th = 0.05**	**th = 0.06**	**th = 0.07**	**th = 0.08**	**th = 0.09**	**th = 0.01**	**th = 0.02**	**th = 0.03**	**th = 0.04**	**th = 0.05**	**th = 0.06**	**th = 0.07**	**th = 0.08**	**th = 0.09**
1	SR	0	0.041	0.228	0.426	0.542	0.648	0.762	0.835	0.890	1	0.997	0.914	0.787	0.690	0.576	0.424	0.340	0.254
1	Las	0	0	0	0.088	0.407	0.729	0.877	0.951	0.976	1	1	1	0.972	0.787	0.482	0.264	0.112	0.046
1	Rid	0	0.001	0.009	0.059	0.340	0.834	0.935	0.964	0.978	1	1	0.992	0.964	0.791	0.264	0.081	0.043	0.018
1	XGB	0.122	0.226	0.360	0.535	0.641	0.745	0.821	0.867	0.901	0.927	0.869	0.757	0.648	0.554	0.470	0.358	0.284	0.203
2	SR	0	0.048	0.193	0.384	0.487	0.575	0.680	0.776	0.845	1	0.988	0.933	0.820	0.742	0.656	0.544	0.432	0.331
2	Las	0	0	0.038	0.228	0.424	0.597	0.745	0.848	0.908	1	1	0.996	0.926	0.792	0.637	0.455	0.335	0.204
2	Rid	0	0.001	0.006	0.064	0.346	0.658	0.843	0.922	0.954	1	0.999	0.997	0.978	0.832	0.543	0.295	0.156	0.114
2	XGB	0	0.020	0.168	0.359	0.502	0.629	0.719	0.790	0.850	1	0.996	0.944	0.853	0.740	0.618	0.521	0.404	0.328
3	SR	0	0.019	0.141	0.323	0.483	0.613	0.723	0.816	0.874	1	0.998	0.956	0.847	0.721	0.586	0.477	0.346	0.256
3	Las	0	0	0.013	0.183	0.439	0.633	0.781	0.874	0.935	1	1	0.996	0.934	0.761	0.584	0.427	0.280	0.156
3	Rid	0	0.001	0.004	0.036	0.351	0.685	0.864	0.935	0.963	1	0.999	0.996	0.974	0.810	0.495	0.258	0.139	0.091
3	XGB	0.020	0.124	0.302	0.467	0.599	0.695	0.767	0.820	0.859	0.995	0.948	0.857	0.737	0.617	0.520	0.441	0.370	0.313
4	SR	0	0.027	0.287	0.465	0.573	0.693	0.793	0.864	0.916	1	0.989	0.867	0.755	0.646	0.522	0.394	0.276	0.181
4	Las	0	0	0.050	0.400	0.546	0.737	0.866	0.932	0.968	1	1	0.980	0.802	0.683	0.472	0.275	0.149	0.075
4	Rid	0	0.002	0.010	0.199	0.600	0.844	0.928	0.960	0.976	1	1	0.992	0.918	0.607	0.298	0.152	0.078	0.050
4	XGB	0.002	0.047	0.296	0.469	0.614	0.722	0.801	0.862	0.904	0.998	0.983	0.882	0.739	0.603	0.502	0.395	0.307	0.239
5	SR	0	0.131	0.4	0.544	0.688	0.798	0.885	0.94	0.971	1	0.948	0.802	0.66	0.522	0.38	0.248	0.132	0.078
5	Las	0	0	0.132	0.482	0.746	0.959	0.998	1	1	1	1	0.952	0.72	0.438	0.091	0.002	0	0
5	Rid	0	0	0.002	0.363	0.933	0.981	0.992	0.996	0.997	1	1	0.999	0.762	0.113	0.031	0.011	0.003	0.002
5	XGB	0.052	0.246	0.469	0.621	0.742	0.817	0.871	0.911	0.936	0.968	0.861	0.705	0.557	0.41	0.312	0.242	0.176	0.122
6	SR	0	0.226	0.457	0.64	0.781	0.871	0.924	0.957	0.977	1	0.907	0.748	0.569	0.419	0.281	0.186	0.107	0.07
6	Las	0	0.014	0.377	0.654	0.853	0.935	0.983	0.995	0.998	1	0.991	0.817	0.557	0.32	0.147	0.045	0.012	0.007
6	Rid	0	0	0.006	0.816	0.951	0.983	0.992	0.995	0.997	1	1	0.991	0.341	0.101	0.034	0.015	0.006	0.006
6	XGB	0.027	0.277	0.5	0.681	0.796	0.856	0.9	0.931	0.954	0.987	0.853	0.691	0.509	0.371	0.281	0.22	0.162	0.111

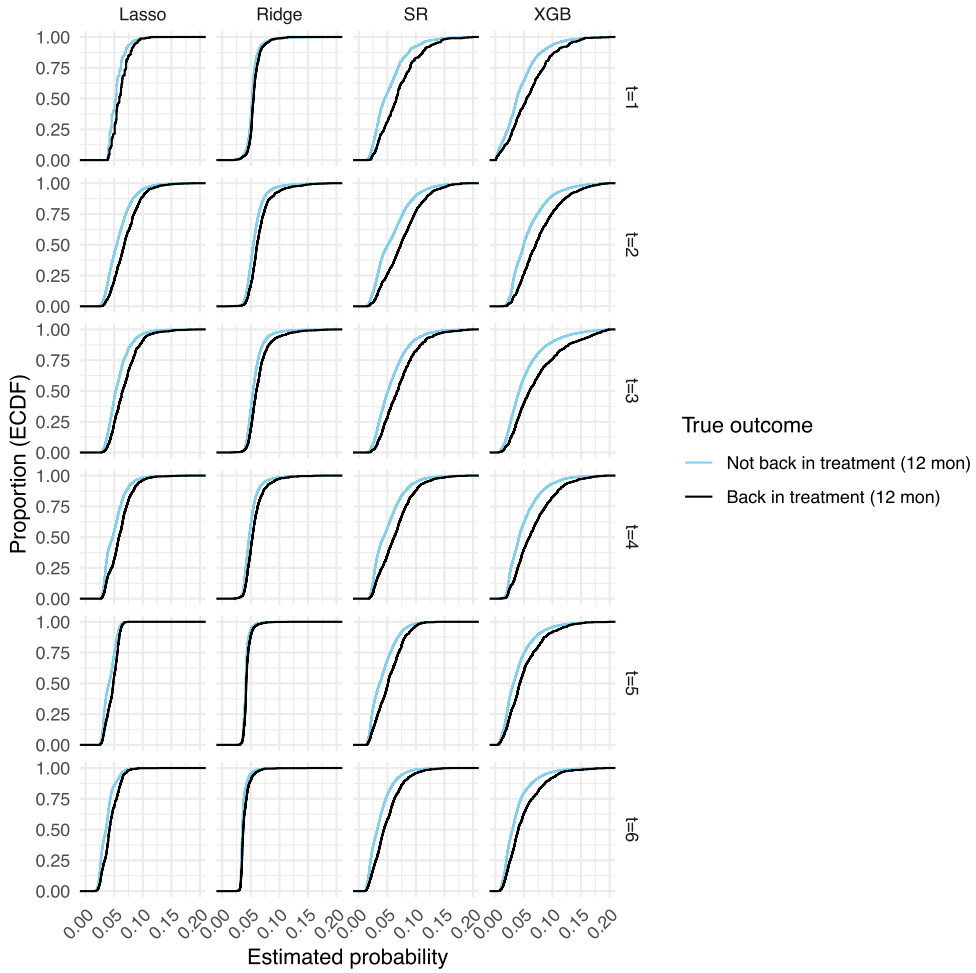

FIGURE 2 | Empirical cumulative distribution function plots for all time points and models. *Note:* The two curves show the actual outcomes—that is, the true group membership per individual.

we therefore elaborate on the consequences of threshold choices for sensitivity and specificity in our case study.

As an illustrative example to show the dependence of the relative model performance on thresholds, consider patients monitored at time point 6 with a chosen threshold of 0.03 or 0.07. Naturally, setting the threshold lower to obtain a higher sensitivity will result in a higher false positive rate, while increasing the threshold to maximize the specificity will lead to a higher false negative rate. In a sample of 1000 patients, with 3.75% being true cases, this corresponds to about 38 individuals who actually experience a relapse and 962 who do not.

At a classification threshold of 0.03, XGBoost achieves a sensitivity of 0.691, correctly identifying 26 of 38 true relapse cases among the 1000 individuals, while missing 12 (false negatives). Lasso regression shows higher sensitivity (0.817), detecting 31 true positives and missing only seven. Among the 962 true non-relapse individuals, XGBoost's specificity of 0.50 results in 481 true negatives and 481 false positives, whereas Lasso's lower specificity (0.377) results in 363 true negatives and 599 false positives—indicating more unnecessary follow-ups than for XGBoost but better detection of cases.

When the threshold increases to 0.07, the comparative model behavior shifts. XGBoost's sensitivity drops to 0.22 (eight true positives, 30 positives missed), while Lasso's decreases to 0.045 (two true positives, 36 positives missed). For non-relapse individuals, XGBoost reaches a specificity of 0.9 (866 true negatives, 96 false positives), and Lasso achieves a specificity of 0.983 (946 true negatives, 16 false positives). Overall, increasing the threshold improves overall specificity but substantially reduces sensitivity in these examples. While Lasso outperforms XGBoost at a lower threshold with respect to model sensitivity, this changes at a higher threshold. The opposite occurs for specificity, highlighting that it is not possible to arrive at a firm conclusion about the superiority of one model over another with respect to these two metrics.

5.1 | Variable Importance XGBoost

In this section, we present the 10 most important variables identified by the XGBoost algorithm, based on the mean variable importance calculated across the four CV folds. At most time points, the most influential features consistently include medication information, the duration of the previous treatment

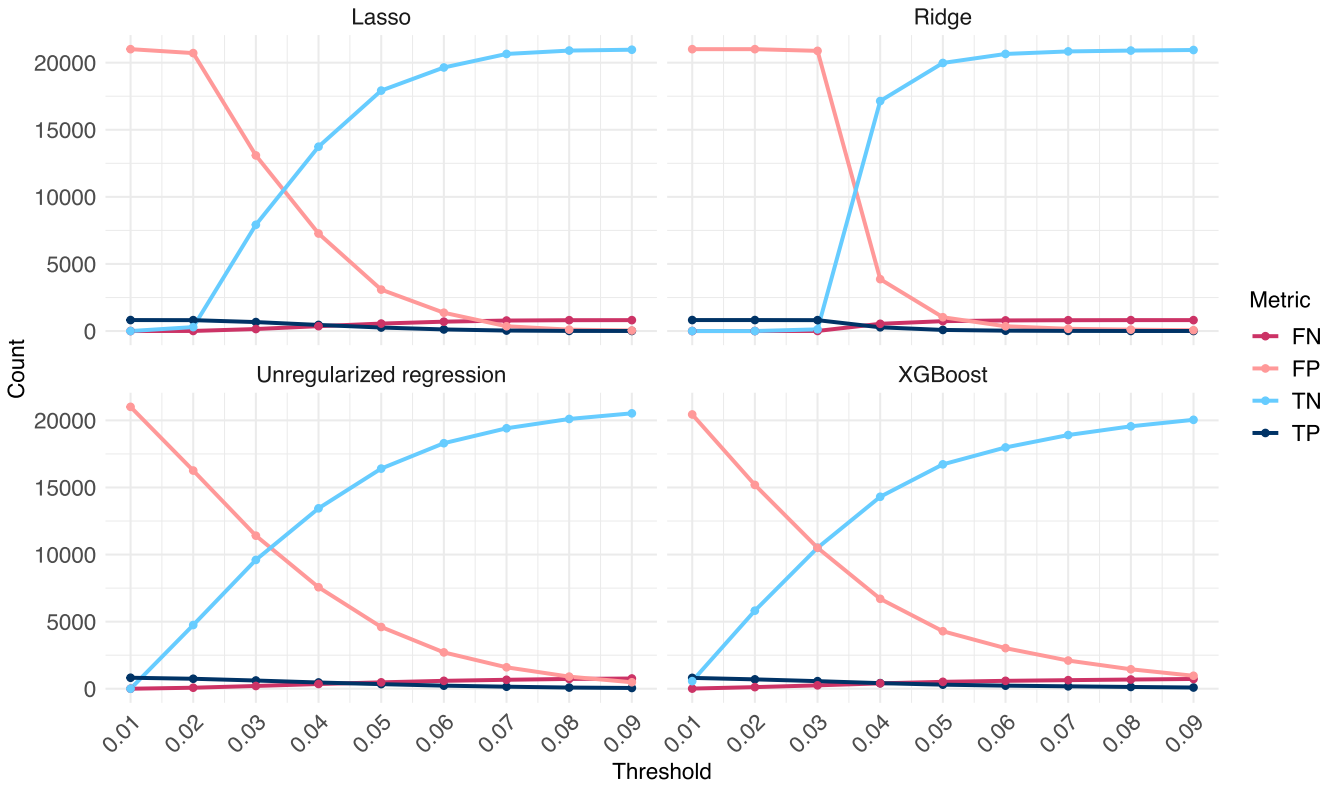

FIGURE 3 | True positives, true negatives, false positives, and false negatives by threshold at time point 6. *Note:* FN, false negative; FP, false positive; TN, true negative; TP, true positive.

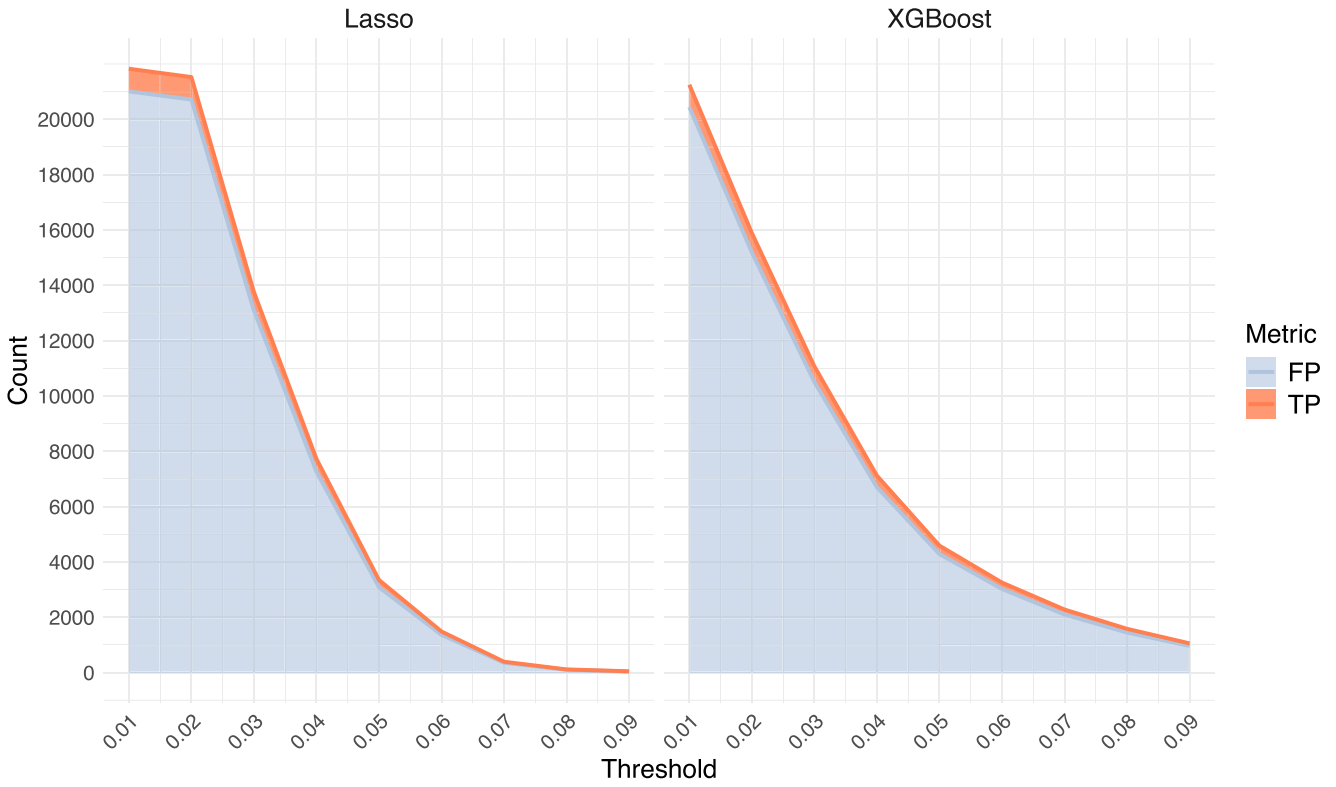

FIGURE 4 | Number of positives (added up from true positives and false positives) for Lasso and XGBoost at time point 6. *Note:* FP, false positive; TP, true positive.

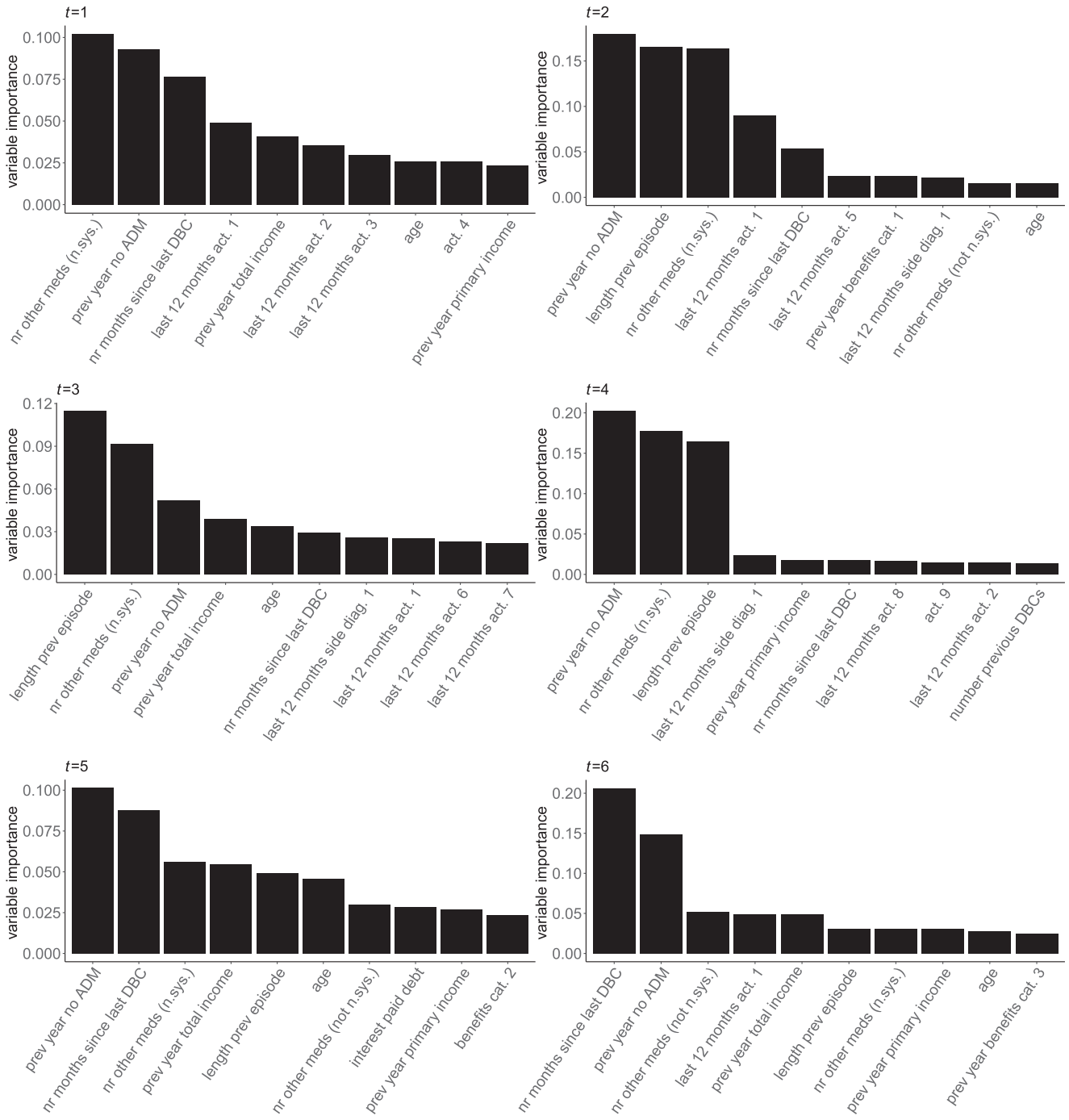

FIGURE 5 | Feature importances XGBoost. Displayed for all six time points. *Note:* act x: health activity number x.; ADM, antidepressant medication; cat. x: category number x; DBC, treatment trajectory; length prev episode, length previous treatment trajectory; nr other meds, number other medications previous year (n.sys., nervous system medication; not n.sys., other types of medication); side diag x, side diagnosis number x. Variable *total income* does not include income insurance premiums.

trajectory, and the number of months since the previous treatment trajectory concluded. Notably, several variables identified as highly important by XGBoost overlap with those selected beforehand for the unregularized regression models. Some variables not included in the SR models but ranked among the top 10 features in XGBoost models include treatment details (e.g., pharmacotherapy) and information on benefits received (Figure 5).

6 | Discussion and Conclusion

In this paper, we study the suitability of using routinely collected data from administrative databases for predictive monitoring of depression relapse. To investigate this problem, we use a large nonpublic dataset from the Netherlands. It is important to note that this study is not intended to bring forward a new prediction model for depression relapse, but rather to assess the

potential of using administrative information for this specific monitoring problem. Below, we summarize the main results and provide recommendations for future developments of monitoring methods for depression relapse using routinely collected data.

6.1 | Model Performances

We observe that in our data, an unregularized regression model (SR) yields results in the same range as those produced by more complex models. Among the regularized regression models, Lasso logistic regression has a slight advantage over Ridge logistic regression, as evidenced by differences in the AUC ranging from 0.017 to 0.033. The complexity of models increases with the number of predictors used, with regularized regression and XGBoost using over 800 predictors, while the simpler unregularized regression model is fitted using fewer than 10 predictors. Considering interpretability, while a marginal increase in AUC is observed at some time points for XGBoost (AUC difference between XGB and SR at $t = 3$: $AUC_{diff} = 0.012$; Table 2) and Lasso logistic regression (AUC difference between Lasso regression and SR at $t = 3$: $AUC_{diff} = 0.013$; Table 2), this improvement comes at the cost of a significantly increased number of parameters in the model. This should be considered in the context of the highest feature importances derived from the XGBoost models, which show that many extracted predictors align with the a priori decisions made for the simpler models in the current study.

Furthermore, the complex models show signs of requiring recalibration, with XGBoost mostly calculating probabilities that are too extreme (i.e., CSs mostly below 1; Table 2) and Lasso calculating probabilities that are not extreme enough (i.e., CSs mostly above 1; Table 2). This indicates overfitting of the XGBoost model, while pointing to underfitting of the Lasso model in this study. Our results show that the choice of threshold influences the relationship between different models in terms of the calculated sensitivity and specificity. As Wynants et al. [47] point out, a risk threshold should therefore consider the clinical context and a model may be validated for multiple thresholds to reflect such changes.

6.2 | Supplementing Smartphone and Clinical Data

Other research on depression prediction often utilizes test and self-report data, either in the form of assessments done in a clinical setting, or by implementing a study based on experience sampling methodology (ESM; see, e.g., [48, 49]). With ESM data, the patients receive a survey to assess their mental states at several points (e.g., on their smartphone), which can give clinicians an idea of, for instance, symptom fluctuations. For example, Snippe et al. [50] use ESM in combination with control charts (EWMA) to monitor depression recurrence, and Klein et al. [24] use structured clinical interviews in a sample of remitted recurrently depressed participants. It should be emphasized that we do not think that purely administrative data as used in the current study can fully replace such other information. However, based on the results obtained, we believe that there are avenues for future research into combining clinical or ESM data with routinely collected information as used in this study, due to its comparable objectivity as well as cost efficiency. Our results also underline that depression relapse is a difficult subject that is hard to predict; therefore, more research is needed to understand the dynamics of depressive disorders which can be done in a more controlled setting.

6.3 | Limitations

6.3.0.1 | Using administrative data. There is a trade-off between the general usefulness of a dataset across situations and how tailored it is to a specific condition. Since we use administrative information, the relapse definition differs from the concept in the clinical literature as we base our outcome on the treatment trajectory start and end dates instead of having direct access to psychiatric assessments. Therefore, our data are less tailored to the clinical picture of relapses as presented in the psychiatric literature, but rather describe the risk of ending up back in treatment as defined by administrative definitions. For instance, patients who do not register as back in treatment but fit the clinical picture of depression are not included in our predictions, as no self-report data are included. Particular decisions in the data preparation stage for this study, such as removing overlapping treatments, might also influence the outcome. Some relevant information (e.g., on medication) may furthermore lack granularity which could affect predictive performance because fluctuations throughout the year are not visible due to aggregation in the database. It should be emphasized that in this study, we focus on individuals who start treatment in 2013 and then monitor the predicted probability for individuals not in treatment at time point t. The mean individual treatment trajectory duration exceeds 6 months, and therefore, many patients monitored at $t = 1$ are individuals who have a shorter treatment trajectory and may be associated with other (unmeasured) characteristics. As the treatment trajectories are kept open for a long time, future studies may want to include more years than we do in the current paper.

6.3.0.2 | Model choices. In the current study, we compare regression models with a ML method in a structured, tabular dataset where most of the input is informed by previous studies on depression prediction. Therefore, the rationale of this study is in line with regression-based approaches where the main predictors are known. In a less structured setting, it is comparatively likely that a ML-based model such as XGBoost would outperform the logistic regression models. Furthermore, models such as XGBoost are dependent on hyperparameter tuning; it could be expected that more extensive tuning may influence the performance of the XGBoost models compared to the results presented in this paper.

6.4 | Recommendations for Practical Implementations and Future Research

Based on our findings and experiences with the current case study, we summarize and outline some central recommendations for projects when researchers work on practical monitoring methods for depression in future studies.

1. **Privacy and ease of data collection**: While some information may have been shown to be clinically relevant, it might not be accessible in practice due to access restrictions.

For instance, previous research points out the influence of adverse childhood events in depression relapse [51]. However, events that take place during a patient's youth and childhood may underlie stricter privacy conditions in a database as in the current case study, and may therefore not be available in similar studies. Adding to this, data on previous traumatic events may be more difficult to obtain in, for instance, interview settings as some patients might be uncomfortable sharing this information. In contrast, objective, available data (e.g., the number of days spent in a hospital in the past year) might not underlie the same strict privacy conditions. Practically informed privacy considerations (see, e.g., [52]) should therefore influence applied research on depression monitoring when choosing variables.

2. **Interpretability and model building**: Even though a complex model can result in slightly better predictions, practitioners and patients might prefer a less accurate prediction that is based on a model with a higher degree of interpretability and fewer variables. In the case studied in the current paper, the simple unregularized logistic regression model performs very similarly to a Lasso logistic regression model, even though the Lasso model allows the inclusion of much more detailed treatment information. While it should be noted that in other examples such additional information may carry more weight than in the current study, we believe that model transparency should guide applied research on mental health monitoring. Variables identified in previous clinical research should form the first step when building a large scale monitoring model that is intended to be used in practice. Our current example demonstrates that more detail in the treatment and social information does not necessarily translate to significantly better predictive performances, and that variables on, for instance, the severity of a mental health condition can be indirectly derived and aggregated without a complex model. Further research may therefore explore using ML in a more "auxiliary" fashion for depression relapse prediction; that is, to discover additional predictors beyond what is already implemented in the simple regression model and subsequently aggregate those predictors for testing and usage in an interpretable small model.

6.5 | Conclusion

Depression relapse is a well-known problem among clinicians. A meta-analysis conducted by Biesheuvel-Leliefeld et al. [15] emphasize the need for preventative treatments in the context of relapses, reinforcing the notion that depression can be viewed as a chronic recurrent disorder [53]. In this light, we argue that systematic monitoring of patients following the conclusion of their treatment could enable early identification of individuals who are at an elevated risk of relapse, and can therefore be used as a tool in depression management.

In this paper, we assess the utility of using routinely collected administrative data for this risk prediction task. Although solely using administrative information as we have it lacks granularity to predict depression treatment readmission on its own, this study is a step towards large-scale monitoring for this issue by pointing to opportunities as well as limitations in similar data. Most importantly, our findings underline the value of careful data collection and informed variable- and model selection. In conclusion, we demonstrate that simpler models and data can achieve predictions on roughly the same level as complex models including a wide range of predictors, therefore supporting the focus on small, interpretable models for practical depression relapse monitoring in routinely collected administrative data.

Funding

The study has been supported by the A Sustainable Future (ASF) initiative of the University of Amsterdam.

Data Availability Statement

The results are based on calculations by the authors using nonpublic microdata from Statistics Netherlands. Article 41 of the CBS Act states that Statistics Netherlands can give access to those data for scientific statistical research. The data are not publicly available due to privacy or ethical restrictions. For further information: microdata@cbs.nl.

Endnotes

[1] Note that in the psychiatric literature, relapse is frequently defined as a return of the depressive episode before recovery, but following remission [31]. Our definition in this study therefore differs as we base it on the administrative information available.

[2] *Note:* We do not use the term "simple logistic regression" in the sense that there is only one independent variable, but instead use this expression to describe that there is no regularization term.

[3] *Note:* The information available concerned medication that was delivered by the pharmacy.

References

1. L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko, "Predictive Monitoring of Mobile Patients by Combining Clinical Observations With Data From Wearable Sensors," *IEEE Journal of Biomedical and Health Informatics* 18, no. 3 (2013): 722–730.

2. T. Hulsen, S. S. Jamuar, A. R. Moody, et al., "From Big Data to Precision Medicine," *Frontiers in Medicine* 6, no. 34 (2019): 1–14.

3. C. Austin and F. Kusumoto, "The Application of Big Data in Medicine: Current Implications and Future Directions," *Journal of Interventional Cardiac Electrophysiology* 47, no. 1 (2016): 51–59.

4. A. M. Chekroud, J. Bondar, J. Delgadillo, et al., "The Promise of Machine Learning in Predicting Treatment Outcomes in Psychiatry," *World Psychiatry* 20, no. 2 (2021): 154–170.

5. A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, "Machine Learning and Decision Support in Critical Care," *Proceedings of the IEEE* 104, no. 2 (2016): 444–466.

6. A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, et al., "Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network," *Nature Medicine* 25, no. 1 (2019): 65–69.

7. X. Hu, "An Algorithm Strategy for Precise Patient Monitoring in a Connected Healthcare Enterprise," *NPJ Digital Medicine* 2, no. 1 (2019): 1–5.

8. E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A Systematic Review Shows no Performance Benefit of Machine Learning Over Logistic Regression for Clinical Prediction Models," *Journal of Clinical Epidemiology* 110 (2019): 12–22.

9. A. Vellido, "The Importance of Interpretability and Visualization in Machine Learning for Applications in Medicine and Health Care," *Neural Computing and Applications* 32, no. 24 (2020): 18069–18083.

10. D. Nickson, C. Meyer, L. Walasek, and C. Toro, "Prediction and Diagnosis of Depression Using Machine Learning With Electronic Health Records Data: A Systematic Review," *BMC Medical Informatics and Decision Making* 23, no. 1 (2023): 271.

11. G. Andrews, R. Poulton, and I. Skoog, "Lifetime Risk of Depression: Restricted to a Minority or Waiting for Most?," *British Journal of Psychiatry* 187, no. 6 (2005): 495–496.

12. G. Y. Lim, W. W. Tam, Y. Lu, C. S. Ho, M. W. Zhang, and R. C. Ho, "Prevalence of Depression in the Community From 30 Countries Between 1994 and 2014," *Scientific Reports* 8, no. 1 (2018): 2861.

13. M. E. Kruijshaar, J. Barendregt, T. Vos, R. De Graaf, J. Spijker, and G. Andrews, "Lifetime Prevalence Estimates of Major Depression: An Indirect Estimation Method and a Quantification of Recall Bias," *European Journal of Epidemiology* 20 (2005): 103–111.

14. C. L. Bockting, S. D. Hollon, R. B. Jarrett, W. Kuyken, and K. Dobson, "A Lifetime Approach to Major Depressive Disorder: The Contributions of Psychological Interventions in Preventing Relapse and Recurrence," *Clinical Psychology Review* 41 (2015): 16–26.

15. K. E. Biesheuvel-Leliefeld, G. D. Kok, C. L. Bockting, et al., "Effectiveness of Psychological Interventions in Preventing Recurrence of Depressive Disorder: Meta-Analysis and Meta-Regression," *Journal of Affective Disorders* 174 (2015): 400–410.

16. P. Thaipisuttikul, P. Ittasakul, P. Waleeprakhon, P. Wisajun, and S. Jullagate, "Psychiatric Comorbidities in Patients With Major Depressive Disorder," *Neuropsychiatric Disease and Treatment* 10 (2014): 2097–2103.

17. J. D. Swendsen and K. R. Merikangas, "The Comorbidity of Depression and Substance Use Disorders," *Clinical Psychology Review* 20, no. 2 (2000): 173–189.

18. C. Otte, S. M. Gold, B. W. Penninx, et al., "Major Depressive Disorder," *Nature Reviews Disease Primers* 2, no. 1 (2016): 1–20.

19. P. F. Sullivan, M. C. Neale, and K. S. Kendler, "Genetic Epidemiology of Major Depression: Review and Meta-Analysis," *American Journal of Psychiatry* 157, no. 10 (2000): 1552–1562.

20. B. W. Sahle, N. J. Reavley, W. Li, et al., "The Association Between Adverse Childhood Experiences and Common Mental Disorders and Suicidality: An Umbrella Review of Systematic Reviews and Meta-Analyses," *European Child & Adolescent Psychiatry* 31 (2022): 1–11.

21. S. Y. Struijs, P. J. de Jong, B. F. Jeronimus, W. van der Does, H. Riese, and P. Spinhoven, "Psychological Risk Factors and the Course of Depression and Anxiety Disorders: A Review of 15 Years NESDA Research," *Journal of Affective Disorders* 295 (2021): 1347–1359.

22. W. Marx, B. W. Penninx, M. Solmi, et al., "Major Depressive Disorder," *Nature Reviews Disease Primers* 9, no. 1 (2023): 44.

23. A. S. Moriarty, N. Meader, K. I. Snell, et al., "Predicting Relapse or Recurrence of Depression: Systematic Review of Prognostic Models," *British Journal of Psychiatry* 221, no. 2 (2022): 448–458.

24. N. S. Klein, G. A. Holtman, C. L. Bockting, M. W. Heymans, and H. Burger, "Development and Validation of a Clinical Prediction Tool to Estimate the Individual Risk of Depressive Relapse or Recurrence in Individuals With Recurrent Depression," *Journal of Psychiatric Research* 104 (2018): 1–7.

25. C. Wojnarowski, N. Firth, M. Finegan, and J. Delgadillo, "Predictors of Depression Relapse and Recurrence After Cognitive Behavioural Therapy: A Systematic Review and Meta-Analysis," *Behavioural and Cognitive Psychotherapy* 47, no. 5 (2019): 514–529.

26. R. Iniesta, K. Malki, W. Maier, et al., "Combining Clinical Variables to Optimize Prediction of Antidepressant Treatment Outcomes," *Journal of Psychiatric Research* 78 (2016): 94–102.

27. I. M. Anderson, D. J. Nutt, and J. Deakin, "Evidence-Based Guidelines for Treating Depressive Disorders With Antidepressants: A Revision of the 1993 British Association for Psychopharmacology Guidelines," *Journal of Psychopharmacology* 14, no. 1 (2000): 3–20.

28. J. A. López-López, S. R. Davies, D. M. Caldwell, et al., "The Process and Delivery of CBT for Depression in Adults: A Systematic Review and Network Meta-Analysis," *Psychological Medicine* 49, no. 12 (2019): 1937–1947.

29. R. W. Taylor, L. Marwood, E. Oprea, et al., "Pharmacological Augmentation in Unipolar Depression: A Guide to the Guidelines," *International Journal of Neuropsychopharmacology* 23, no. 9 (2020): 587–625.

30. Z. Rihmer, "Can Better Recognition and Treatment of Depression Reduce Suicide Rates? A Brief Review," *European Psychiatry* 16, no. 7 (2001): 406–409.

31. A. J. Rush, H. C. Kraemer, H. A. Sackeim, et al., "Report by the ACNP Task Force on Response and Remission in Major Depressive Disorder," *Neuropsychopharmacology* 31, no. 9 (2006): 1841–1853.

32. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 785–794.

33. B. Lorimer, J. Delgadillo, S. Kellett, and J. Lawrence, "Dynamic Prediction and Identification of Cases at Risk of Relapse Following Completion of Low-Intensity Cognitive Behavioural Therapy," *Psychotherapy Research* 31, no. 1 (2021): 19–32.

34. S. L. Cessie and J. V. Houwelingen, "Ridge Estimators in Logistic Regression," *Journal of the Royal Statistical Society Series C: Applied Statistics* 41, no. 1 (1992): 191–201.

35. R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58, no. 1 (1996): 267–288.

36. J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software* 33, no. 1 (2010): 1–22.

37. B. Bischl, M. Lang, L. Kotthoff, et al., "mlr: Machine Learning in R," *Journal of Machine Learning Research* 17, no. 170 (2016): 1–5, https://jmlr.org/papers/v17/15-066.html.

38. E. W. Steyerberg, A. J. Vickers, N. R. Cook, et al., "Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures," *Epidemiology* 21, no. 1 (2010): 128–138.

39. B. Van Calster, D. J. McLernon, M. Van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: The Achilles Heel of Predictive Analytics," *BMC Medicine* 17 (2019): 1–7.

40. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2024), https://www.R-project.org/.

41. H. Wickham, M. Averick, J. Bryan, et al., "Welcome to the Tidyverse," *Journal of Open Source Software* 4, no. 43 (2019): 1686.

42. Kuhn and Max, "Building Predictive Models in R Using the Caret Package," *Journal of Statistical Software* 28, no. 5 (2008): 1–26, https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

43. D. Kaledkowski, *runner: Running Operations for Vectors*, R Package Version 0.4.4 (2024), https://CRAN.R-project.org/package=runner.

44. X. Robin, N. Turck, A. Hainard, et al., "pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves," *BMC Bioinformatics* 12 (2011): 77.

45. H. Wickham, E. Miller, and D. Smith, *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*, R Package Version 2.5.4 (2023), https://CRAN.R-project.org/package=haven.

46. P. von Stackelberg, R. Goedhart, Ş. İ. Birbil, and R. J. M. M. Does, "Comparison of Threshold Tuning Methods for Predictive Monitoring," *Quality and Reliability Engineering International* 40, no. 1 (2024): 499–512.

47. L. Wynants, M. Van Smeden, D. J. McLernon, D. Timmerman, E. W. Steyerberg, and B. Van Calster, "Three Myths About Risk Thresholds for Prediction Models," *BMC Medicine* 17 (2019): 1–7.

48. A. C. Smit, E. Snippe, L. F. Bringmann, H. R. Hoenders, and M. Wichers, "Transitions in Depression: If, How, and When Depressive Symptoms Return During and After Discontinuing Antidepressants," *Quality of Life Research* 32, no. 5 (2023): 1295–1306.

49. M. Wichers, A. C. Smit, and E. Snippe, "Early Warning Signals Based on Momentary Affect Dynamics Can Expose Nearby Transitions in Depression: A Confirmatory Single-Subject Time-Series Study," *Journal for Person-Oriented Research* 6, no. 1 (2020): 1–15.

50. E. Snippe, A. C. Smit, P. Kuppens, H. Burger, and E. Ceulemans, "Recurrence of Depression Can Be Foreseen by Monitoring Mental States With Statistical Process Control," *Journal of Psychopathology and Clinical Science* 132, no. 2 (2023): 145–155.

51. J. E. Buckman, A. Underwood, K. Clarke, et al., "Risk Factors for Relapse and Recurrence of Depression in Adults and How They Operate: A Four-Phase Systematic Review and Meta-Synthesis," *Clinical Psychology Review* 64 (2018): 13–38.

52. A. A. de Hond, A. M. Leeuwenberg, L. Hooft, et al., "Guidelines and Quality Criteria for Artificial Intelligence-Based Prediction Models in Healthcare: A Scoping Review," *NPJ Digital Medicine* 5, no. 1 (2022): 2.

53. M. B. Keller, "The Long-Term Treatment of Depression," *Journal of Clinical Psychiatry* 60 (1999): 41–45.

Appendix A: Dataset

The following table (Table A1) shows some (fictional) example rows. *Patient ID* describes the individual identifier per person. *Month* shows the month in the dataset (note that in the original data, every person has 48 months of information). *Depression treatment* describes whether a patient is currently registered in a treatment trajectory for depression as a main diagnosis. *Relapse* describes whether the patient starts a new depression treatment trajectory within the next 2 months. Do note that in the actual dataset, a relapse is registered looking at the next 12 months. We adjusted this here for readability. *Act 1* describes the number of minutes registered for healthcare activity number one for that month. *Act 2* also calculates the number of minutes spent in a particular treatment, but then for healthcare activity two. *Income* describes the income of the previous year, and *Age* shows the patient's age in that month. As described in the main text, we use this dataset to subset individuals based on the month of interest and other criteria to fit the models.

TABLE A1 | Fictional example rows from the balanced panel data.

Patient ID	Month	Depression treatment trajectory	Relapse	Act 1	Act 2	Income	Age
1	1	0	1	0	0	40,000	38
1	2	0	1	0	0	40,000	38
1	3	1	—	40	0	40,000	38
1	4	1	—	120	10	40,000	38
1	5	1	—	0	0	40,000	38
2	1	0	0	0	0	29,000	23
2	2	0	0	0	0	29,000	23
2	3	0	1	0	0	29,000	23
2	4	0	1	0	0	29,000	23
2	5	1	—	20	0	29,000	24

Appendix B: Results of all Unregularized Regression Models

Below, we report the predictors contained in the regression models that are evaluated during the stepwise variable selection procedure, as well as corresponding results (Table B1). As can be seen below, the results roughly fall in the same range, so we focus on the results from the final model (SR 6) in the main text.

1. SR 1: Age + Sex + Previous year primary income + Previous year ADM + Previous year ONM + Previous year OM + Anxiety side diagnosis in treatment trajectory of last 12 months + Number of months since last treatment trajectory + Number of previous treatment trajectories + Length previous treatment trajectory
2. SR 2: Age + Sex + Previous year ADM + Previous year ONM + Previous year OM + Anxiety side diagnosis in treatment trajectory of last 12 months + Number of months since last treatment trajectory + Number of previous treatment trajectories + Length previous treatment trajectory
3. SR 3: Age + Sex + Previous year ADM + Previous year ONM + Previous year OM + Number of months since last treatment trajectory + Number of previous treatment trajectories + Length previous treatment trajectory
4. SR 4: Age + Previous year ADM + Previous year ONM + Previous year OM + Number of months since last treatment trajectory + Number of previous treatment trajectories + Length previous treatment trajectory
5. SR 5: Age + Previous year ADM + Previous year ONM + Number of months since last treatment trajectory + Number of previous treatment trajectories + Length previous treatment trajectory
6. SR 6 (final predictor set as reported in the results in the main text): Age + Previous year ADM + Previous year ONM + Number of months since last treatment trajectory + Number of previous treatment trajectories

TABLE B1 | AUC, Brier score, and calibration slopes for all six fitted regression models per time point.

t	AUC_{SR1}	AUC_{SR2}	AUC_{SR3}	AUC_{SR4}	AUC_{SR5}	AUC_{SR6}	$Brier_{SR1}$	$Brier_{SR2}$	$Brier_{SR3}$	$Brier_{SR4}$	$Brier_{SR5}$	$Brier_{SR6}$	CS_{SR1}	CS_{SR2}	CS_{SR3}	CS_{SR4}	CS_{SR5}	CS_{SR6}
1	0.653	0.656	0.657	0.658	0.658	0.659	0.051	0.051	0.051	0.051	0.051	0.051	0.936	0.953	0.963	0.975	0.975	0.979
2	0.662	0.663	0.664	0.664	0.664	0.661	0.055	0.055	0.055	0.055	0.055	0.055	0.951	0.959	0.962	0.967	0.969	0.991
3	0.649	0.649	0.649	0.651	0.652	0.643	0.054	0.054	0.054	0.054	0.054	0.054	0.959	0.964	0.965	0.978	0.989	0.989
4	0.652	0.652	0.652	0.654	0.656	0.653	0.048	0.048	0.048	0.048	0.048	0.048	0.950	0.951	0.955	0.964	0.982	0.988
5	0.647	0.648	0.648	0.648	0.648	0.649	0.040	0.040	0.040	0.040	0.040	0.040	0.956	0.966	0.978	0.983	0.984	0.995
6	0.653	0.653	0.654	0.654	0.654	0.654	0.036	0.036	0.036	0.036	0.036	0.036	0.958	0.961	0.966	0.976	0.978	0.987