





Predictive monitoring using machine learning algorithms and a real-life example on schizophrenia

Leo C. E. Huberts¹  | Ronald J. M. M. Does¹  | Bastian Ravesteijn²  |
Joran Lokkerbol³ 

¹ Department of Operations Management, University of Amsterdam, Plantage Muidergracht 12, Amsterdam 1018TV, Netherlands

² Erasmus School of Economics, Erasmus University Rotterdam, Burgemeester Oudlaan 50, Rotterdam 3062PA, Netherlands

³ Epidemiology, Trimbos Institute, Da Costakade 45, Utrecht 3521VS, Netherlands

Correspondence

Leo C. E. Huberts, University of Amsterdam, Department of Operations Management, Netherlands.

Email: L.c.e.huberts@uva.nl

Abstract

Predictive process monitoring aims to produce early warnings of unwanted events. We consider the use of the machine learning method extreme gradient boosting as the forecasting model in predictive monitoring. A tuning algorithm is proposed as the signaling method to produce a required false alarm rate. We demonstrate the procedure using a unique data set on mental health in the Netherlands. The goal of this application is to support healthcare workers in identifying the risk of a mental health crisis in people diagnosed with schizophrenia. The procedure we outline offers promising results and a novel approach to predictive monitoring.

KEYWORDS

extreme gradient boosting, false alarm rate, machine learning, mental health, predictive process monitoring, schizophrenia, tuning algorithm

1 | INTRODUCTION

Predictive monitoring is a promising research area that focuses on forecasting potential problems during process execution before they occur.¹ Applications have been developed in a wide range of domains, such as manufacturing,^{2–4} healthcare,^{5–7} networking,⁸ and business processes.⁹ Increasingly comprehensive data collection provides more process visibility. Advances in machine learning make use of this increase in detail and frequency of data. Data-driven techniques in this area can be used to improve process quality control by forecasting and monitoring potential process problems.

Gradient boosting is an important recent development within machine learning for regression and classification. This technique produces an ensemble of decision trees that minimize an appropriate loss function. Such an ensemble often produces better predictions than a single, more comprehensive model. When used for classification, as in most process monitoring applications, the predictions are in the form of probabilities, similar to the output of regression models. In this paper, these predicted probabilities will be used in a process monitoring procedure.

An area that has a real interest in predictive monitoring for quality control is mental health.¹⁰ Nineteen percent of adults in the United States have a mental, behavioral, or emotional disorder.¹¹ These disorders pose a heavy burden on the patient and affected families, the healthcare system, and healthcare expenditure. Early intervention is important for many of the severe mental disorders and could prevent the escalation of the disease. However, it is very hard to predict the progress of a disease and thus determine when to intervene.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2021 The Authors. *Quality and Reliability Engineering International* published by John Wiley & Sons Ltd

One of the most debilitating mental disorders is schizophrenia. According to the World Health Organization,¹² schizophrenia is characterized by distortions in thinking, perception, emotions, language, sense of self, and behavior. Common experiences include hallucinations and delusions. Around 0.75% of people suffer from the disease worldwide.¹³ Schizophrenia is one of the top 15 leading causes of disability worldwide.¹⁴ The all-cause standardized mortality rate is around 3.7 times higher for people diagnosed with schizophrenia compared to the general adult population.¹⁵ Suicide is much more frequent among people suffering from schizophrenia. An estimated 4.9% of people with schizophrenia die by suicide compared to 0.013% for the general population.¹⁶

The overwhelming majority of people suffering from schizophrenia will relapse into crisis care over time, even with access to good care.¹⁷ Relapse averages are reported between 20% and 40% per year, depending on many factors.¹⁸ In the United States, the estimated total cost of schizophrenia was \$155.7 billion in 2013.¹⁹ The healthcare costs for people diagnosed with schizophrenia are significantly higher than the national average, where relapse events (i.e., hospital admissions) are the most expensive.²⁰ This motivates the need for early identification of patients at high risk of having a mental health crisis, to facilitate preventive measures, and mitigate the high costs associated with these crises.

In a systematic review, Sullivan et al.²¹ investigated models to predict crises and found a lack of high-quality evidence on prediction methods. Paxton et al.²² and Amarasingham et al.²³ highlighted the challenges that predictive modeling based on Electronic Medical Records faces. According to Sullivan et al.²¹ the number of studies with promising results is very limited. One of the exceptions is the study by Vigod et al.²⁴ which used Canadian data from 2008 to 2011 to predict 30-day readmission rates for acute psychiatric units using logistic regression. The study shows moderate discriminative capacity with an area under the receiver operating characteristic curve of 0.630.

In this study, we have access to all mental health treatment records covered by the Dutch Health Insurance Act, which covers all specialist mental health treatment of schizophrenia for all 17 million residents of the Netherlands between 2010 and 2014. We use this data to predict readmission into crisis care for the 75,000 people diagnosed with schizophrenia in the Netherlands. We compare the predictive power of logistic regression as used by Vigod et al.²⁴ to a hierarchical regression model.

Subsequently, a gradient boosting algorithm named extreme gradient boosting (abbreviated by XGBoost) is applied to the data. Teinema et al.²⁵ reviewed available predictive process monitoring techniques and concluded that XGBoost is reasonably fast and often the most accurate technique. It can deal with class imbalance and incomplete observations, which are often found in medical data.²² Furthermore, Zhang et al.²⁶ used XGBoost for predictive monitoring of faults in wind turbines.

The XGBoost predictions are then used to monitor the set of people diagnosed with schizophrenia during a monitoring phase. We present an algorithm to determine the threshold to signal and consider the monitoring performance. The goal of the procedure is to support healthcare workers in identifying individuals at risk in a crisis.

The article is structured as follows. In the next section, we present the mental health problem context. We describe the predictive models we consider in this study and their performance in Section 3. In Section 4 we propose an algorithm to determine the monitoring threshold and present the monitoring results when using the XGBoost technique. In the last section, we provide concluding remarks and limitations.

2 | PROBLEM DESCRIPTION

This section describes the setting in which this case study attempts to monitor the risk of mental health crises. First, we give a summary of the mental healthcare system in the Netherlands. We then describe the available data, followed by the definition of a crisis event.

2.1 | The mental healthcare system

The total cost of mental healthcare in the Netherlands was estimated at 6.5 billion Euros in 2017,²⁷ of which 416 million Euros was directly related to schizophrenia.²⁸ The details of the healthcare system are out of scope, but this section describes the basics.

Similar to the US system, mental healthcare in the Netherlands is organized through managed competition. In contrast to the United States, health insurance is mandatory for all Dutch citizens and covers 99.9% of the population.²⁹ Insurers are required to accept all applicants and offer community rating. The deductibles are relatively low in the Netherlands and there is risk adjustment among insurers. Coverage includes a broad set of essential health benefits, including

out- and in-patient treatment of nearly all disorders in the Diagnostic and Statistical Manual of Mental Disorders 5,³⁰ except select diagnoses such as adjustment disorder (since 2012). Curative healthcare expenditure is relatively high with per capita spending of 3791 Euros in 2017 and the long-term care spending is the highest of all EU countries. The system is comparatively effective and the Netherlands reports the lowest rate of unmet medical needs among EU countries.²⁹

2.2 | Data description

The non-public microdata used in this paper is provided by Statistics Netherlands. It consists of a wide range of de-identified administrative data sets on all 17 million Dutch citizens. As the data is very sensitive, it is stored on secure servers at the Statistics Netherlands and can solely be accessed on their local terminals.

A patient requires a diagnosis in order for the health insurer to reimburse the mental healthcare treatments. The diagnosis, together with the amount of treatment provided, constitutes the so-called diagnosis–treatment–combination (DBC in Dutch) that determines the amount of reimbursement. As reimbursement is dependent on consistent registration, the system results in a clear timeline of the diagnoses and activities for a patient.

Mental health data were available for the years 2010 through 2013. The data includes all registered healthcare information, such as detailed mental health diagnoses (1.4 million) and psychological treatments (25 million). The categories of treatments were as defined by the Dutch Healthcare Authority.³¹ Furthermore, individual data on employment, housing, and personal information were available. The data included in the following concerns the subset of 75,000 people diagnosed with schizophrenia. The resulting selection consists of an unbalanced panel data set, with a large variation in the number of registered treatment activities per individual.

The gathered data cannot be used directly for statistical modeling, as the sequences of diagnoses and treatments vary widely in length, frequency, and type. Time series models thus require padding and aggregation to balance the data which will produce sparse sequences for many of the included individuals. The decision of the level of aggregation was motivated by domain experts and set at the week level. The week-level aggregation had enough details to expect predictive value in the data while resulting in an actionable time frame for intervention.

Diagnoses and treatments were described in detailed labels. These were condensed into broader categories to avoid high dimensionality and sparsity. The 36 resulting categories of diagnoses are tracked cumulatively. The weekly aggregations of treatments are further aggregated into a short-, medium-, and long-term history of 4, 12, and 64 weeks, respectively. The 4 weeks represent the past month of data, the 12 weeks represent the last quarter, and the past 64 weeks consist of the last year plus one quarter to make sure there is overlap between separate (administrative) years. These three levels result in 264 predictors (74 for each of the three treatment aggregation levels, 36 cumulative diagnoses, and 6 fixed variables) for more than 15 million person-weeks. The final structure of the set of predictors, largely informed by expert input, is illustrated in Table 1.

2.3 | The definition of a crisis

A variable indicating a crisis was not readily available and had to be constructed from the raw data. This section describes how the variable to signal a crisis event was constructed. As described previously, individuals are assigned DBCs. Some of these DBCs are directly defined as “crisis care,” that is, treating a patient in a mental health institution due to a crisis. In other cases, crisis care is registered to an existing, non-crisis DBC. Therefore, we define the start of a crisis event as the first moment a crisis DBC was opened or any crisis care was given.

Furthermore, as the goal of the procedure is to support healthcare workers in identifying individuals at risk of a crisis, the model should not focus on individuals that are already known to be in a mental health crisis. The crisis care service in the Netherlands aims to provide a maximum of 12 weeks of crisis care. In the data, following the start of a crisis an individual will be excluded for the following 20 weeks. This 20-week period covers the 12 weeks that a crisis can cover, plus some additional weeks after that, where it could be argued that a patient will be on the radar of the healthcare professional and a signaling mechanism is not needed to achieve this.

This results in a binary dependent variable on a weekly basis, where a TRUE value equals the start of the crisis in that week and a FALSE indicating the subject is not in a crisis. This binary dependent variable is sparse with only 0.285% TRUE values for all person-weeks.

As the long-term aggregation level (64-week aggregations) includes all diagnoses and healthcare activities from the past 64 weeks, the first week we can model is week 65 (i.e., March 2011). Figure 1 plots the aggregated number of crises per

TABLE 1 Illustration of the set of predictors after feature engineering, where ID identifies the person, Week represents the current week, Age is the age in years, Gender contains the gender of ID (M = male, F = female), Diag. 1 indicates the number of diagnoses of type 1 that ID received, GAF 4w represents the Global Assessment of Functioning score in the past 4 weeks (0–100 scale), Inc. 4w is the total income over the past 4 weeks in Euros, Med. 4w signals medication prescribed over the past 4 weeks, Act. 1 4w contains the number of hours of activity 1 in the past 4 weeks and Act. 74 64w represents the number of hours of activity 74 in the past 64 weeks. There are 36 categories of diagnoses, 74 categories of activities all of which were aggregated into short-term (4 weeks), medium-term (12 weeks) and long-term (64 weeks) history variables

ID	Week	Age	Gender	Diag. 1	Diag. 2	...	Diag. 36	GAF 4w	Inc. 4w	Med. 4w	Act. 1 4w	Act. 2 4w	...	Act. 73 64w	Act. 74 64w
1	100	37	M	0	1	...	0	50	1200	4	32	0	...	0	120
1	101	37	M	0	1	...	0	45	900	2	42	0	...	0	120
1	102	37	M	0	2	...	0	45	600	1	37	10	...	0	105
:	:	:	:	:	:	...	:	:	:	:	:	:	...	:	:
75,000	209	62	F	1	0	...	0	90	2200	0	0	0	...	0	0
75,000	210	62	F	1	0	...	0	90	2200	0	0	0	...	0	0

FIGURE 1 Number of crises per week from March 2011 (week 65) to December 2013 (week 209)

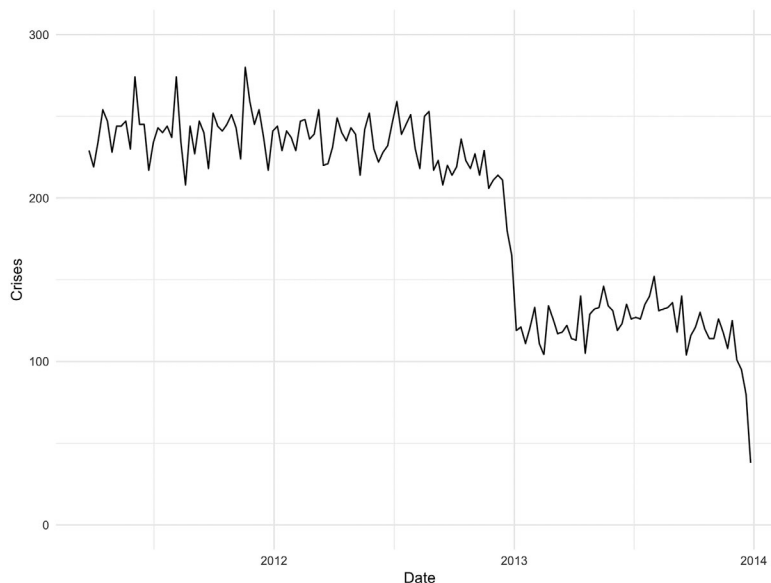
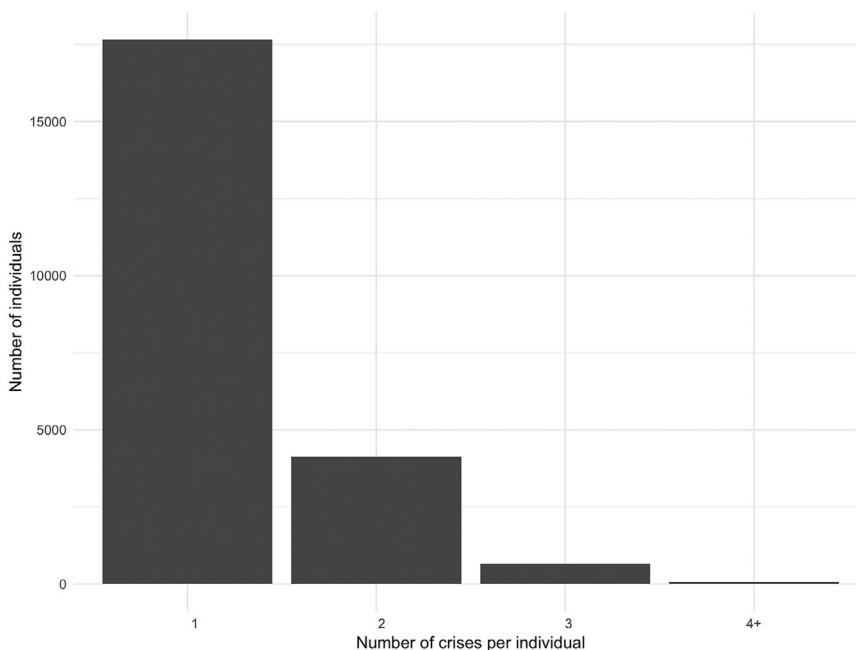


FIGURE 2 Number of crises per individual from March 2011 (week 65) to December 2013 (week 209)



week. There are a lot fewer crises that start in 2013 than the other 2 years due to administrative changes and incomplete data for that year. This is important to consider when evaluating the results.

In total there are just over 28,300 crises that start between March 2011 and December 2013. These crises are divided over 22,600 people with an average of around 0.4 crisis per person in the data. The mean number of weeks between crises (for people with multiple crises in the data) equals 51 weeks. Figure 2 gives an overview of the number of crises per individual that had a crisis at least once. A large majority has one crisis during the 3 years we consider. Table 2 gives an overview of the 10 mental healthcare activities with the largest weekly mean values in the data.

2.4 | The status quo

This section describes how healthcare practitioners in the Netherlands currently monitor people diagnosed with schizophrenia. The majority of institutions work using the Flexible Assertive Community Treatment method (FACT) described by Nugter et al.³² FACT is based on the widely used Assertive Community Treatment (ACT) method developed

TABLE 2 Descriptive statistics for the 10 most frequent mental healthcare activities from March 2011 (week 65) to December 2013 (week 209) per person-week

Activity	Mean	Std.Dev
Individual contact	4.81	16.92
General no-show/other	1.13	8.04
Individual activating guidance	0.66	7.23
Diagnostics	0.65	8.23
Pharmacotherapy	0.60	4.86
Individual other communication	0.31	3.82
Individual supportive guidance	0.29	4.72
Group contact	0.21	2.91
Patient system	0.17	3.34
Crisis treatment	0.16	4.23

in the United States. The method uses multidisciplinary teams, a low number of clients per team member, assertive outreach, shared caseloads, and indefinite intensive care. FACT applies these tools in a flexible way spanning the entire caseload.

The FACT teams meet on a weekly basis to discuss the caseload. The team will focus on clients that are on the FACT board. The board displays the clients with the highest priority, that is, clients that are experiencing severe mental health issues. Every week, the team members are allowed to add or remove clients from the board. If a client is added to the board, a crisis plan is triggered and the client and their surrounding network are informed that care is being intensified. Depending on the urgency, the client will be visited by the head practitioner within a week. A client can be added to the board if the FACT team members detect an increase in psychiatric symptoms, behavioral problems, substance abuse, or criminal behavior. Furthermore, care-avoiding behavior, imminent crises, (forced) admission, life events, changes in medication, and the introduction of new clients can persuade a team member to propose to add a client to the board. Note that this process relies entirely on the observations of the team members.

3 | PREDICTIVE MODEL

This section describes the logistic regression model (similar to Vigod et al.²⁴), the hierarchical regression model, and the XGBoost algorithm that are used to predict weekly probabilities of getting a crisis for people diagnosed with schizophrenia.

Define $y_{i,t}$ as a binary variable for individual $i = 1, \dots, N$ and week $t = 1, \dots, M$. If person i has a crisis in week t , then $y_{i,t} = 1$, if there is no sign of crisis $y_{i,t} = 0$. Let $p_{i,t}$ be the predicted probability of crisis in week t for person i . Furthermore, $x_{i,t}$ contains the constructed features based on the 4-, 12-, and 64-week history of individual i , as well as the individual characteristics.

3.1 | Regression

Using regression models for prediction has the advantage of high explainability. In contrast to many of the available machine learning techniques, the parameters of the logistic and hierarchical regression models we discuss can offer insight into the process dynamics.

3.1.1 | Logistic regression

Logistic regression is used to model the probabilities of a categorical outcome variable. In this case, the categorical outcome is binary. The model for the probability of crisis $p_{i,t} = P(y_{i,t} = 1|x_{i,t})$ for individual i in week t with vector of predictors $x_{i,t}$ has the form³³



FIGURE 3 Two-level structure of the case study data with individuals ($i = 1, \dots, N$) as the top level. Weeks in the data ($t = 1, \dots, M$) belonging to individual i are the bottom level

$$p_{i,t} = \frac{\exp(\beta_0 + \beta' x_{i,t})}{1 + \exp(\beta_0 + \beta' x_{i,t})}, \quad (1)$$

where β_0 is a vector of constants and β' is the transposed vector of parameters for the predictors $x_{i,t}$. The model is fitted using maximum likelihood. The estimated parameters $\hat{\beta}_0, \hat{\beta}$ can be used for inference and prediction of $p_{i,t+1}$.

3.1.2 | Hierarchical regression

Hierarchical modeling is almost always an improvement over single-level regression models.³⁴ In this case, each observation equates to one person-week. The mental healthcare activities, diagnoses, income changes, and crises leading up to a specific week all relate to a single person. We can model this as a simple two-level hierarchy, with the weeks as the lower level and the individual as the upper level (see Figure 3). For more details on predictive monitoring using hierarchical regression see Huberts et al.³⁵

Suppose we have p_0 predictors on the person-week level and p_1 predictors on the person level. The hierarchical model for the log-odds ratio of the probability of a crisis for individual i in week t is then defined as

$$\log\left(\frac{p_{it}}{1 - p_{it}}\right) \sim N(X_{it}\alpha_i, \sigma^2), \text{ for } t = 1, \dots, M \text{ (Week level)}, \quad (2)$$

where the individual level is modeled as

$$\alpha_i \sim N(\gamma W_i', \Sigma), \text{ for } i = 1, \dots, N \text{ (Individual level)}, \quad (3)$$

where X_{it} is a $1 \times (p_0 + 1)$ row vector of person-week specific variables such as mental healthcare activities and diagnoses (see Table 2); α_i is a $(p_0 + 1) \times 1$ vector of parameters for individual i ; σ^2 is the variance for the person-week level; γ is a $(p_0 + 1) \times (p_1 + 1)$ parameter matrix determined by the person i that person-week t is a part of; W_i is a $1 \times (p_1 + 1)$ row vector of person specific variables such as age and Σ is the covariance matrix for parameters α_i . We estimate the parameters using restricted maximum likelihood (REML) in the *lme4* package in R.³⁶ The values of the estimated parameters $\hat{\alpha}_i, \hat{\sigma}, \hat{\gamma}, \hat{\Sigma}$ offer insight into the effects of variables X_{it} and W_i and the size of the variance at the person-week level and person level.

3.2 | Machine learning

A lot of progress has been made in the machine learning domain in recent years. Increased availability of data and computing power extend the range of models that can be estimated. In this section, we discuss a gradient boosting framework called XGBoost and a few alternative techniques.

3.2.1 | Extreme gradient boosting

Gradient boosting is one of the most important recent developments in machine learning.³³ XGBoost is an open-source framework to apply gradient boosting in various programming languages.³⁷ The gradient boosting decision tree algorithm within XGBoost creates an ensemble of weak learners that minimize an appropriate loss function. Each weak learner consists of a regression tree grown on the residuals. Each tree has a number of terminal nodes $j = 1, \dots, J_k$ that each represents

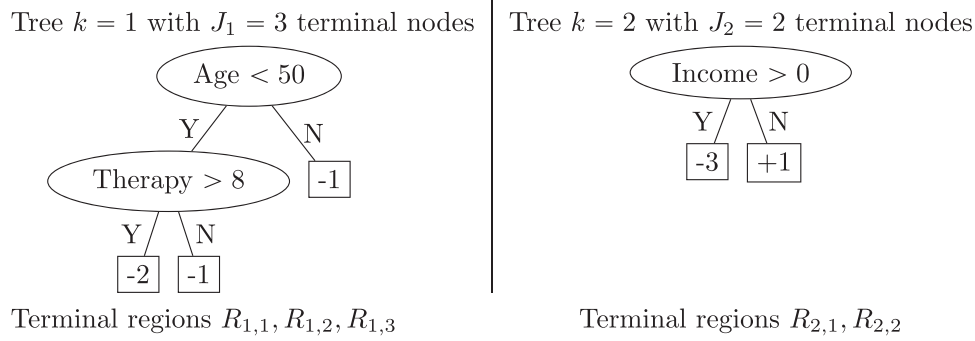


FIGURE 4 An example of a tree ensemble model for the risk of a mental health crisis using two trees ($K = 2$) and three variables. The Therapy variable contains hours of therapy in the past 4 weeks, the Income variable the amount in Euros earned over the past 4 weeks. The learning rate is $\eta = 0.3$. The initial prediction equals $p_0 = 0.5$

a terminal region $R_{k,j}$, containing the predictions for that specific tree. The output of a regression tree k is multiplied by learning rate η and then added to the predictions of tree $k - 1$.

Figure 4 gives a fictional example of two trees grown to predict the risk of having a mental health crisis. The example uses the variables age, hours of therapy, and income in Euros over the past 4 weeks. The values in the terminal nodes are the log-odds $\log\left(\frac{p}{1-p}\right)$. The final prediction equals the initial log-odds prediction plus the predictions of the regression trees multiplied by the learning rate η . Given the example in Figure 4, the final prediction of the log-odds for a 35-year-old person with 16 h of therapy and an income of 1500 in the past 4 weeks equals $0 - 0.3 \times 2 - 0.3 \times 3 = -1.5$ ($p \approx 0.18$). The log-odds prediction for 25-year-old person with no therapy and no income in the past 4 weeks equals $0 - 0.3 \times 1 + 0.3 \times 1 = 0$ ($p = 0.5$).

In this study, the outcome is binary thus the logistic loss function is used, given by

$$L(y_{i,t}, p_{i,t}) = -y_{i,t} \log(p_{i,t}) + (y_{i,t} - 1) \log(1 - p_{i,t}), \quad (4)$$

for individual i at time t . The algorithm is initialized by setting $p_{i,t,0} = 0.5$. Then a manually specified number of trees $k = 1, 2, \dots, K$ is grown on the pseudo-residuals. The pseudo-residuals are $r_{i,t,k} = y_{i,t} - p_{i,t,k-1}$, where $p_{i,t,k-1}$ are the predicted probabilities of the previous iteration $k - 1$.

For each tree k the objective function $obj(t, k)$ for log-odds output values $f_k(x_{i,t})$ consists of the loss function, the pruning term and a regularization term

$$obj(t, k) = \left[\sum_{i=1}^n L(y_{i,t}, f_k(x_{i,t}) + f_{k-1}(x_{i,t})) \right] + \chi J_k + \frac{1}{2} \lambda f_k^2(x_{i,t}), \quad (5)$$

where χ is a user specified pruning parameter and λ is a regularization parameter.

The output value at time t for the individuals i in terminal node $j = 1, \dots, J_k$ in tree k that minimizes $obj(t, k)$ is approximated using the second-order Taylor expansion. The gradient for $L(y_{i,t}, p_{i,t})$ equals $g_{i,t} = y_{i,t} - p_{i,t}$ and the hessian equals $h_{i,t} = p_{i,t}(1 - p_{i,t})$. The output value for tree k is then given by

$$f(x_{i \in j, t, k}) = \frac{\sum_{i \in j} (y_{i,t} - p_{i,t,k-1})}{\sum_{i \in j} [p_{i,t,k-1}(1 - p_{i,t,k-1})] + \lambda}. \quad (6)$$

The similarity score used to determine the splits when constructing the decision trees is found by plugging the output value back into the second-order Taylor approximation and is given by $\frac{\sum_i (y_i - p_i)^2}{\sum_i p_i(1 - p_i) + \lambda}$.

There are four parameters $\Theta = \{\eta, \lambda, \chi, K\}$ that need to be tuned using cross-validation. Sparsity-aware split finding in the XGBoost framework ensures the algorithm works efficiently for the sparse data in this study. We use the implementation of *xgboost* in R.³⁸

TABLE 3 Table of the mean estimated probabilities for the three methods, aggregated by the binary crisis outcome variable

Model	$\bar{p}_{t y=0}$	$\bar{p}_{t y=1}$	r_t	AUC
Logistic regression*	0.0021	0.0026	1.2260	0.6130
Hierarchical regression*	0.0021	0.0030	1.4370	0.5856
XGBoost	0.0017	0.0062	3.5440	0.6533

*Using 50% of the persons in the data due to limited memory size

3.2.2 | Other machine learning methods

Several other machine learning techniques can be used in this setting. Examples include (one-class) support vector machines (SVM), decision trees, random forest, and elastic nets. For an overview of machine learning methods in process monitoring see Weese et al.³⁹ We applied one-class SVM, random forest, and the elastic net methods to the same data in this paper. Compared to the XGBoost predictions, these methods produced inferior results. This is the reason that we have excluded these methods from the analysis. There are some methods, such as recurrent neural networks (see, e.g., Choi et al.⁴⁰), that require more computing power than was available at the Statistics Netherlands terminals.

3.3 | Estimation

To train the models, weeks 1–175 are used. The first 64 weeks are incorporated into the features as described in Section 2.2, thus the outcomes of weeks 65–175 are used in training. Weeks 176–209 are used to test the predictions. Due to limits in available computing power, the regression models were trained on a random 50% of the people in the training set. The XGBoost model was trained using both the random 50% of the people as well as the full set. The results are very similar, but as the XGBoost model requires extensive cross-validation we use the full data set for that method in the following. To cross-validate the parameters of the XGBoost model $\Theta = \{\eta, \lambda, \chi, K\}$ we use random splits (75%/25%) of the person-weeks in the training data. A large grid of values for Θ was implemented. The model used in the following is based on $\eta = 0.3, \lambda = 1, \chi = 0.5, K = 123$. Note that the performance of the XGBoost model varies with the values of these parameters. Extensive cross-validation is thus an important step before implementation.

3.4 | Results

This section describes the results of predicting weekly crises using the logistic regression model, hierarchical model, and the XGBoost algorithm described previously. These results are based on the predictions for the test set of weeks 176–209. Because of the highly imbalanced nature of the outcome, we consider the mean assigned probabilities for weeks with crises, $\bar{p}_{t|y=1} = \frac{1}{N} \sum_{i=1}^N p_{i,t} I(y_{i,t} = 1)$, and without crises $\bar{p}_{t|y=0} = \frac{1}{N} \sum_{i=1}^N p_{i,t} I(y_{i,t} = 0)$. The ratio of these probabilities, $r_t = \frac{\bar{p}_{t|y=1}}{\bar{p}_{t|y=0}}$, is a measure of the predictive power of a model. Values of $r_t \leq 1$ indicate that, on average, the procedure assigns the same or a lower probability to person-weeks with positive outcome values. Values $r_t > 1$ show that, on average, the method estimates a higher probability to person-weeks with positive outcome values. We also report the widely used area under the receiver operating characteristic curve (AUC) values (see Bradley⁴¹). AUC values close to 0.5 indicate a total lack of predictive power, values close to 1 represent perfect prediction.

Table 3 shows the measures of performance for the three methods using the test set. The logistic regression model does predict a slightly higher probability for person-weeks with crises with a ratio of 1.226. The limited predictive power is also shown by the AUC value of 0.613. The hierarchical model incorporates some more of the structure in the data. The performance in terms of the ratio of predicted probabilities seems slightly better than the logistic model with $\bar{r} = 1.437$. Conversely, the AUC value is lower than for the logistic regression, which indicates a limited predictive power. Lastly, Table 3 shows the results for the XGBoost algorithm predictions on the test set. The mean ratio \bar{r} is around 3.5 and it has the highest AUC score of the three, which shows there is more predictive power than for the logistic and hierarchical regressions. On average, for person-weeks with a crisis, the XGBoost algorithm predicts a probability 3.5 times higher than for person-weeks with no crisis. This suggests the predicted probabilities of the algorithm might be used as a risk score to guide mental health workers toward unstable individuals.

4 | MONITORING

The three models in the previous section predicted if a crisis is likely to occur for an individual in a given week. This section will discuss the use of these predictions for monitoring.

Using logistic regression, a hierarchical model or the XGBoost algorithm to model the risk of a crisis will result in weekly estimated probabilities. Monitoring these probabilities requires a probability control limit C . Once a probability $p_{i,t}$ passes this limit, the procedure will signal and practitioners can intervene. The choice of C will determine the number of false/true signals. A higher value of C will decrease the share of false signals, but also decrease the absolute number of true signals.

In process monitoring, the performance of a monitoring procedure is often quantified using the false alarm rate (FAR) or the average run length (ARL) (see, e.g., Shu et al.⁴² and Does et al.⁴³). The ARL equals the average time it takes for the procedure to signal. A monitoring procedure is configured to satisfy a required ARL or FAR as determined by the practitioner. This generally involves adjusting parameters based on distributional assumptions or simulation. Distributional assumptions are not realistic with a machine learning method such as XGBoost. Thus, in the following section, we propose a simple tuning procedure that achieves a desired FAR for a predictive monitoring approach. This involves cross-validating the predictions and results in a non-parametric monitoring threshold. Note that the procedure is data driven, thus no distributional assumptions are needed.

4.1 | Tuning procedure

In this section, we propose a tuning procedure to achieve a desired FAR . Assume a practitioner determines a FAR based on the monitoring context. In the case of predictive monitoring, a signal is produced when the predicted probability $p_{i,t}$ exceeds the threshold C .

The FAR set by the practitioner in this procedure translates to all observations during monitoring. For example, setting $FAR = 0.01$ will result in 1% of observations being false alarms. The other 99% of observations consist of true/false negatives and true positives. A higher FAR will result in a lower value of C .

The following steps determine the value of C that produces the desired FAR

1. Set FAR , the size of the training set N_0 and the test set N_1 , the number of cross-validation splits S , the cross-validation proportion q and initialize an empty vector \hat{C} .
2. Split data $\{X, y\}$ of size $N \times k$ into $i = 1, 2, \dots, N_0$ training set $\{X_0, y_0\}$ of size $N_0 \times k$ and $i = N_0 + 1, N_0 + 2, \dots, N$ test set $\{X_1, y_1\}$ of size $N_1 \times k$ with $N_1 = N - N_0$.
3. Draw a random sample R of integers $1, 2, \dots, N_0$ of size qN_0 and define the vector V of size $(1 - q)N_0$ as integers $1, 2, \dots, N_0 \notin R$.
4. Split the training set $\{X_0, y_0\}$ into $\{X_R, y_R\}$ using rows R and $\{X_V, y_V\}$ using rows V .
5. Use $\{X_R, y_R\}$ to estimate model $F(X)$ and calculate $p_V = F(X_V)$.
6. Use the grid-search algorithm below to find value \hat{c} for which $\frac{1}{(1-q)N_0} \sum_{i \in V} I(p_{i,V} \geq \hat{c})I(y_{i,V} = 0) \approx FAR$, with $I()$ the indicator function.
 - (a) Set resolution $r > 2$, $w = 1/r$, search limit w_{lim} to a small value (i.e., 10^{-5}) and initiate grid $G = \{0, 1/r, 2/r, \dots, 1 - 2/r, 1 - 1/r, 1\}$ of length $r + 1$.
 - (b) Set $c = \min(g \in G : \frac{1}{(1-q)N_0} \sum_{i \in V} I(p_{i,V} \geq g)I(y_{i,V} = 0) < FAR)$.
 - (c) Calculate $FAR_s = \frac{1}{(1-q)N_0} \sum_{i \in V} I(p_{i,V} \geq c)I(y_{i,V} = 0)$.
 - (d) Update w as $w = 2w/r$ and redefine grid $G = \{c - rw/2, c - rw/2 + w, c - rw/2 + 2w, \dots, c - rw/2 + (r - 1)w, c - rw/2 + rw\}$ of length $r + 1$.
 - (e) If $|FAR - FAR_s| > 0.01FAR$ and $w > w_{lim}$ go back to step (b). If $|FAR - FAR_s| \leq 0.01FAR$ set $\hat{c} = c$. If $|FAR - FAR_s| > 0.01FAR$ and $w \leq w_{lim}$ set $\hat{c} = NA$.
7. Save value \hat{c} in vector \hat{C} , if the length of \hat{C} is smaller than S return to step 3.
8. Set threshold value C_{tuned} as $\max(\hat{C})$ for $\hat{C} \neq NA$. If all values in \hat{C} are NA no threshold was found. Use $\{X_0, y_0\}$ to estimate model $F(X)$ and calculate $p_1 = F(X_1)$. The expected false alarm rate \widehat{FAR} then equals $\widehat{FAR} = \frac{1}{N_1} \sum_{i=N_0+1}^N I(p_{1,i} \geq C_{tuned})I(y_{1,i} = 0)$.

FIGURE 5 Average probability per week grouped for people that have a crisis in the monitoring time frame (orange) and that do not have a crisis (blue)

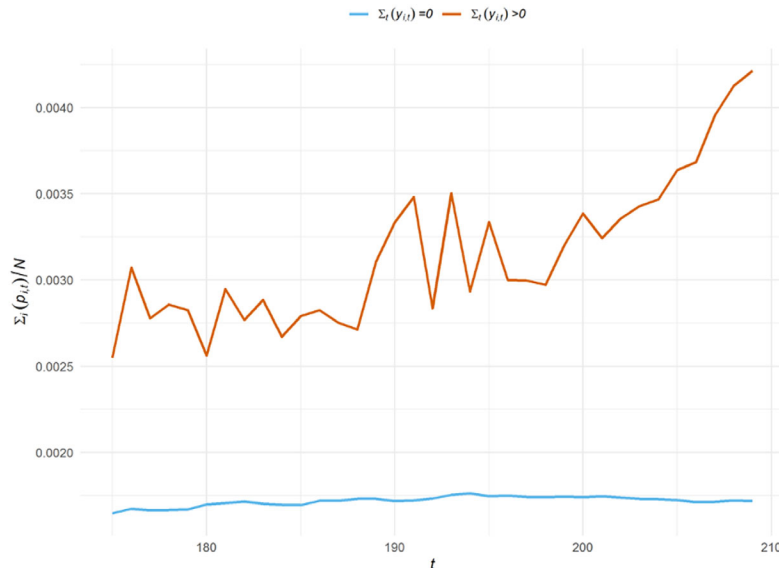


TABLE 4 Precision and recall values using the XGBoost estimated probabilities and various values for threshold C

C	Precision	Recall
0.0001	0.0017	1.0000
0.0010	0.0023	0.7955
0.0100	0.0081	0.0517
0.1000	0.1224	0.0087
0.5000	0.5000	0.0018
0.7500	1.0000	0.0003

Note that the maximum estimated value in \hat{C} is set as C_{tuned} . Assuming the model generalizes well to new data, this will result in an expected false alarm rate \widehat{FAR} that is smaller than the desired FAR . We will demonstrate the procedure in the following section.

4.2 | Results

In this section, we monitor the probability of a crisis as predicted by the XGBoost algorithm. Threshold C determines when the procedure signals. As measures of performance we consider the precision and recall values. The precision is given by

$$\text{Precision}(C) = \frac{tp(C)}{tp(C) + fp(C)}, \quad (7)$$

with $tp(C)$ equal to the number of true positives for threshold C and $fp(C)$ the number of false positives for threshold C . The recall is defined as

$$\text{Recall}(C) = \frac{tp(C)}{tp(C) + fn(C)}, \quad (8)$$

where $fn(C)$ equals the number of false negatives for threshold C .⁴⁴

Figure 5 shows the average estimated probability per week, grouped by the observed value of $y_{i,t}$. On average, the procedure estimates a visibly higher probability for the individuals that have a crisis for all weeks. We cannot show individual probabilities to ensure the privacy of the persons.

Table 4 shows the precision and recall values for a wide range of values for C . The table shows perfect recall for $C = 0.0001$, but the precision is very low. Perfect precision is achieved for $C = 0.75$, but the recall is very low. This shows that, although the model does have some predictive power, a high false alarm rate is needed to detect a large portion of the crises.

TABLE 5 Average estimated difference in probabilities per crisis/no crisis if we consider a 10-week period

Crisis within 10 weeks	Average estimated difference in probability
No	-0.000002
Yes	0.000296
Ratio	-135.846300

TABLE 6 Precision and recall values using the XGBoost estimated probabilities of a crisis within 10 weeks and various values for threshold C

	C	Precision	Recall
1	0.0001	0.0148	1.0000000
2	0.0010	0.0183	0.8244052
3	0.0100	0.0438	0.0265457
4	0.1000	0.1603	0.0010674
5	0.5000	0.5000	0.0001685
6	0.7500	1.0000	0.0000281

4.2.1 | Weeks before a crisis

The individual probability plots often show high volatility in the weeks leading up to a crisis. In this subsection, we thus consider the average difference in estimated probability in the weeks before a crisis occurs.

Table 5 gives the average difference in estimated probabilities over 10 weeks ($\frac{1}{9} \sum_{l=t-9}^t (p_{i,l} - p_{i,l-1})$ for week t and individual i), grouped by whether a crisis was observed at the end of those 10 weeks. This shows that, on average, the estimated average difference in probability of a crisis in the 10 weeks leading up to crisis is 135 times higher than if no crisis is observed after these 10 weeks.

Table 6 gives the precision and recall values for a range of C values when predicting if a crisis will occur within 10 weeks. This shows a higher precision for low values of $C \leq 0.1$ compared to the weekly predictions (cf. Table 4).

4.2.2 | Tuning the procedure

In this section, we run the procedure including the tuning algorithm for various required FARs. The desired FAR produces a value for C through the tuning algorithm outlined in Section 4.1. This value C_{tuned} is then used to monitor the test set. We use $S = 10$ splits in the procedure and use cross-validation proportion $q = 0.75$ for all values of FAR .

Table 7 gives the tuned values C_{tuned} for a set of predetermined false alarm rates $FAR \in \{0.5, 0.1, 0.05, 0.01, 0.001, 0.0001, 0.00001\}$, as well as the precision/recall values in the test set and the actual observed $FAR_{observed}$. The table shows that all the $FAR_{observed}$ values are smaller than their respective FAR values. In step 8 of the procedure in Section 4.1 the maximum estimated value in \hat{C} is set as C_{tuned} . This results in $FAR_{observed} \leq FAR$ for all FAR values of Table 7.

TABLE 7 C_{tuned} -values for various predetermined values of FAR , as well as the precision/recall and the observed $FAR_{observed}$

FAR	C_{tuned}	$FAR_{observed}$	Precision	Recall
0.50000	0.001961	0.233681	0.003285	0.451351
0.10000	0.006116	0.029413	0.006130	0.106306
0.05000	0.008648	0.014699	0.007784	0.067568
0.01000	0.017988	0.003014	0.014247	0.025526
0.00100	0.051656	0.000327	0.064516	0.013213
0.00010	0.139054	0.000060	0.180556	0.007808
0.00001	0.333604	0.000009	0.400000	0.003604

4.3 | False positives/negatives

The predictive monitoring procedure of the previous sections will result in a number of false positives (i.e., a crisis is falsely predicted) and false negatives (i.e., a crisis is not detected). Both are unwanted outcomes, with different consequences. A false positive will result in unnecessary intervention. The responsible healthcare worker will be triggered by the procedure to investigate the mental health state of the individual and these hours might be better spent otherwise. On the contrary, a false negative will result in a crisis. This crisis might have been mitigated or even prevented if correctly predicted in previous weeks. The practitioner can utilize the tuning procedure of Section 4.1 to produce an acceptable number of false positives. Furthermore, the monitoring procedure might enable automated checkups where individuals at risk are automatically asked to provide some input for the healthcare worker. The results of such an inquiry might warrant further investigation or indicate a high likelihood of a false positive. In the next section we further discuss possible applications.

4.4 | Recommendations

In this section, we summarize the findings of the previous sections into some recommendations for researchers and practitioners in the mental healthcare domain.

In terms of feature engineering, we propose to aggregate healthcare activities, diagnoses, and personal data such as income and job status into short-, medium-, and long-term features. A crisis should be defined based on input from the healthcare practitioners. This will ensure that signals produced by the monitoring system are of value to the team. Section 2.3 describes the definition of a crisis as used in this study.

As discussed in Section 3, gradient boosting method XGBoost offers a flexible approach to modeling the probability of mental health crises. As input, the team of healthcare workers employing the monitoring system should decide on an acceptable number of false alarms per week. This number of alarms can be achieved using the tuning procedure of Section 4.1. The desired FAR is calculated by dividing the number of acceptable false alarms by the number of clients in a given week. This FAR is the input for the algorithm of Section 4.1 which gives the appropriate threshold C_{tuned} . The number of false alarms and associated FAR and C_{tuned} should regularly be updated.

Section 4.2 shows that both the predictions for the coming week as well as the variation in probabilities over a 10 week period can be of value toward identifying crises. The predictive monitoring can include both indicators, where the variation in the past 10 weeks can be investigated to support the single week signals based on C_{tuned} . Interventions triggered by the monitoring system should be recorded in the data.

For example, suppose a FACT team has 20 h per week to invest in client checkups. Each checkup takes around 12 min to complete and the team is responsible for 1000 clients. In this case, 100 false alarms would not harm their core activities. This suggests a desired $FAR = 10\%$ and according to Table 7 a C_{tuned} of 0.006116. The observed FAR equals 2.94%, which means the team will have around 30 checkups every week spending around 6 h. In expectation, around 11% of crises will be detected by the system using these checkups. Predicting crises within the next 10 weeks can increase the percentage of crises detected by the system. Furthermore, registering the result of a checkup will decrease the number of false alarms and can increase the precision and recall of the monitoring system.

5 | CONCLUSION AND DISCUSSION

In this study, predictive monitoring using XGBoost is investigated. We develop a procedure that can produce early warnings of problematic events and can be tuned to deliver a desired false alarm rate.

Advances in data collection and machine learning techniques can improve process quality control by forecasting and monitoring potential process problems. XGboost is a recently developed powerful machine learning framework that efficiently combines weak learners to minimize an appropriate loss function. The predictive monitoring procedure is demonstrated using a real-life example on mental health in the Netherlands.

Predictive monitoring is an area of tremendous interest in mental health.¹⁰ A unique non-public data set on mental health in the Netherlands was provided by Statistics Netherlands. We focused on predictive monitoring of mental health crises in people diagnosed with schizophrenia. These crises are harmful, frequent, and expensive, which motivated the need for early warnings of these events. All 75,000 people diagnosed with schizophrenia in the Netherlands were included

in the study. The individual healthcare treatments, diagnoses, admissions, and incomes were aggregated on a weekly interval. Subsequently, we built explanatory variables on three levels of aggregation, short (4 weeks), medium (12 weeks), and long term (64 weeks). The final data set consisted of more than 15 million person-weeks and 264 predictors.

The data was then used to predict the probability of a crisis in a future week. We compared the performance of logistic regression, hierarchical regression, and the XGBoost algorithm. All three methods showed predictive power, assigning a higher probability of a crisis to individuals that end up in crisis care in the coming week. The XGBoost framework achieved the highest discriminative capacity and was subsequently used in the monitoring procedure.

Predicted probabilities were monitored using a threshold value C . The procedure signals when the predicted probability exceeds this value. Each value for C will result in a number of true/false signals. A higher threshold will result in fewer false positives, but it will also miss more cases of crisis in the monitoring phase. We propose a search algorithm to find a value for C that results in a desired false alarm rate. This algorithm uses cross-validation on the training data and delivers good results in this case study.

We also considered the monitoring performance looking up to 10 weeks ahead. More specifically, the average estimated difference in probabilities in the weeks leading up to a crisis was 135 times higher than for weeks that do not lead to a crisis. This can be used by practitioners to produce early warnings of crises in people diagnosed with schizophrenia.

Mental health crises on a weekly basis are rare events, occurring in less than 0.3% of the recorded cases. The predictive monitoring procedure shows promising results, although a high degree of uncertainty remains. The administrative changes in the final year of the data made predicting crises more challenging. Further tuning of the parameters could produce more accurate results but are out of the scope of this study. The proposed tuning algorithm provides a tool for practitioners to configure the monitoring procedure based on the available capacity.

Some challenges in the application of the predictive monitoring procedure in mental health remain. The data wrangling operation is extensive. The administration of treatments and diagnoses is complicated causing inconsistencies among healthcare providers. The facilities to process and clean the numerous protected microdata sources are currently limited in the Netherlands. Improving the consistency and improving the computational facilities will boost the predictive performance.

A logical avenue for further research is the application of recurrent neural networks to a similar data set. This requires more computational capacity than was available in this study. An interesting follow-up would be to compare the accuracy of expert opinion to the results of a predictive monitoring system as presented in this paper. Especially, studying the changes in the accuracy of experts using the predictive monitoring system is of importance toward successful implementation. Furthermore, applying the predictive monitoring procedure to a process with a less sparse outcome is of interest.

In summary, predictive monitoring using XGBoost can produce good results in the mental health domain, as well as other areas. It can be tuned to achieve a desired false alarm rate and is capable of handling large amounts of (sparse) data. The tuning procedure and unique data set in this study represent a new direction in process monitoring.

ACKNOWLEDGMENTS

The authors wish to thank the Statistics Netherlands Institute for giving access to the unique data set used in this study. Furthermore, the authors wish to thank the Trimbos Institute for providing mental healthcare practitioners' input on the current operation and the design of the monitoring system proposed in this article. The authors would also like to thank the editors and reviewers at Quality and Reliability Engineering International for their valuable comments on this paper.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from Statistics Netherlands. Restrictions apply to the availability of these data, which were used under license for this study.

ORCID

Leo C. E. Huberts  <https://orcid.org/0000-0001-5139-5522>

Ronald J. M. M. Does  <https://orcid.org/0000-0003-3452-6441>

Bastian Ravesteijn  <https://orcid.org/0000-0003-1914-2431>

Joran Lokkerbol  <https://orcid.org/0000-0001-9949-5442>

REFERENCES

1. Metzger A, Leitner P, Ivanović D, et al. Comparing and combining predictive business process monitoring techniques. *IEEE Trans Syst Man Cybern Syst*. 2015;45(2):276–290.
2. Spiewak S, Duggirala R, Barnett K. Predictive monitoring and control of the cold extrusion process. *CIRP Ann*. 2000;49(1):383–386.
3. Zhou J, Li X, Andernrooer A, et al. Intelligent prediction monitoring system for predictive maintenance in manufacturing. *31st Ann Conf IEEE Ind Electron Soc*. 2005;1:2314–2319.
4. Smith WS, Coleman S, Bacardit J, Coxon S. Insight from data analytics with an automotive aftermarket SME. *Qual Reliab Eng Int*. 2019;35(5):1396–1407.
5. Reifman J, Rajaraman S, Gribok A, Ward W. Predictive monitoring for improved management of glucose levels. *J Diabetes Sci Technol*. 2007;1(4):478–486.
6. Clifton L, Clifton DA, Pimentel MA, Watkinson PJ, Tarassenko L. Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors. *IEEE J Biomed Health Inf*. 2013;18(3):722–730.
7. Luo J. Predictive monitoring of COVID-19. Singapore University of Technology and Design Data-Driven Innovation Lab, 2020.
8. Ali A, Khelil A, Shaikh FK, Suri N. Efficient predictive monitoring of wireless sensor networks. *Int J Auton Adapt Commun Syst*. 2012;5(3):233–254.
9. Tax N, Verenich I, La Rosa M, Dumas M. Predictive business process monitoring with LSTM neural networks. *Int Conf Adv Inf Sys Eng*. 2017;1:477–492.
10. Hahn T, Nierenberg AA, Whitfield-Gabrieli S. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Mol Psychiatry*. 2017;22(1):37–43.
11. Substance Abuse and Mental Health Services Administration. Key substance use and mental health indicators in the United States: Results from the 2017 national survey on drug use and health, 2018, September 1. <http://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHFFR2017/NSDUHFFR2017.pdf>
12. World Health Organization. Schizophrenia fact sheet, 2019, October 4. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/schizophrenia>
13. Moreno-Küstner B, Martin C, Pastor L. Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses. *PLoS One*. 2018;13(4):1–25.
14. Vos T, Abajobir AA, Abate KH, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *Lancet*. 2017;390(10100):1211–1259.
15. Olsson M, Gerhard T, Huang C, Crystal S, Stroup TS. Premature mortality among adults with schizophrenia in the United States. *JAMA Psychiatry*. 2015;72(12):1172–1181.
16. Palmer BA, Pankratz VS, Bostwick JM. The lifetime risk of suicide in schizophrenia: a reexamination. *Arch Gen Psychiatry*. 2005;62(3):247–253.
17. Emsley R, Chiliza B, Asmal L, Harvey BH. The nature of relapse in schizophrenia. *BMC Psychiatry*. 2013;13(50):1–8.
18. Ruetsch C, Un H, Waters HC. Claims-based proxies of patient instability among commercially insured adults with schizophrenia. *ClinicoEconomics and Outcomes Res*. 2018;10:259–267.
19. Cloutier M, Aigbogun MS, Guerin A, et al. The economic burden of schizophrenia in the United States in 2013. *J Clin Psychiatry*. 2016;77(6):764–771.
20. Karve SJ, Panish JM, Dirani RG, Candrilli SD. Health care utilization and costs among medicaid-enrolled patients with schizophrenia experiencing multiple psychiatric relapses. *Health Outcomes Res Med*. 2012;3(4):183–194.
21. Sullivan S, Northstone K, Gadd C, et al. Models to predict relapse in psychosis: A systematic review. *PLoS One*. 2017;12(9):1–12.
22. Paxton C, Niculescu-Mizil A, Saria S. Developing predictive models using electronic medical records: challenges and pitfalls. In *AMIA Annual Symposium Proceedings*, pp. 1109–1115. American Medical Informatics Association; 2013.
23. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. *Health Affairs* 2014;33(7):1148–1154.
24. Vigod SN, Kurdyak PA, Seitz D, et al. Readmit: a clinical risk index to predict 30-day readmission after discharge from acute psychiatric units. *J Psychiatric Res*. 2015;61:205–213.
25. Teinmaa I, Dumas M, Rosa ML, Maggi FM. Outcome-oriented predictive process monitoring: review and benchmark. *ACM Trans Knowl Discovery Data (TKDD)*. 2019;13(2):1–57.
26. Zhang D, Qian L, Mao B, Huang C, Huang B, Si Y. A data-driven design for fault detection of wind turbines using random forests and XGboost. *IEEE Access*. 2018;6:21020–21031.
27. Statistics Netherlands. Health, lifestyle, medical care use and offer, death causes; key figures [data file]. Retrieved May 1, 2019, from <https://opendata.cbs.nl/#/CBS/nl/dataset/81628NED/table?ts=1597217730991>
28. National Institute for Public Health and the Environment. Costs of diseases 2017 [data file]. Retrieved May 1, 2019, from <https://statline.rivm.nl/#/RIVM/nl/dataset/50050NED/table?ts=1597217173008>
29. OECD & European Observatory on Health Systems and Policies. *Netherlands: Country Health Profile 2019*. Retrieved May 1, 2019, from <https://www.oecd-ilibrary.org/content/publication/9ac45ee0-en>
30. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. Arlington, VA: American Psychiatric Association; 2013.

31. Dutch Healthcare Authority. Spelregels GGZ 2015; bijlage 1 activiteiten- en verrichtingenlijst (rules mental care 2015; appendix 1 activity and treatment list). Retrieved May 1, 2019, from <https://dbcregels.nza.nl/2015/documents/20150101Spelregelsggzv20141028Activiteiten-enverrichtingenlijst.pdf>
32. Nugter MA, Engelsbel F, Bähler M, Keet R, van Veldhuizen R. Outcomes of flexible assertive community treatment (fact) implementation: a prospective real life study. *Community Mental Health J.* 2016;52(8):898–907.
33. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer; 2009.
34. Gelman A. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics.* 2006;48(3):432–435.
35. Huberts LCE, Schoonhoven M, Does RJMM. Multilevel process monitoring: a case study to predict student success or failure. *J Qual Technol.* 2020. <https://doi.org/10.1080/00224065.2020.1828008>
36. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Software.* 2015;67(1):1–48.
37. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *Proc 22nd ACM SIGKDD Int Conf Knowl discovery Data Min.* 2016;1:785–794.
38. Chen T, He T, Benesty M, et al. *XGBoost: Extreme Gradient Boosting*. R package version 0.90.0.2. Retrieved May 1, 2019, from <https://CRAN.R-project.org/package=xgboost>
39. Weese M, Martinez W, Megahed FM, Jones-Farmer LA. Statistical learning methods applied to process monitoring: an overview and perspective. *J Qual Technol.* 2016;48(1):4–24.
40. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inf Assoc.* 2016;24(2):361–370.
41. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 1997;30(7):1145–1159.
42. Shu L, Tsung F, Tsui K-L. Run-length performance of regression control charts with estimated parameters. *J Qual Technol.* 2004;36(3):280–292.
43. Does RJMM, Goedhart R, Woodall WH. On the design of control charts with guaranteed conditional performance under estimated parameters. *Qual Reliab Eng Int.* 2020;36(8):2610–2620.
44. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Technol.* 2011;2(1):37–63.

AUTHOR BIOGRAPHIES

Dr. Leo C. E. Huberts is an assistant professor in the Department of Business Analytics at the University of Amsterdam, the Netherlands. His current research focuses on statistical and predictive process monitoring with an emphasis on analytics for a better world.

Prof. Dr. Ronald J. M. M. Does is professor of industrial statistics at the University of Amsterdam and independent statistical engineer. He is member of the editorial boards of the Journal of Quality Technology (since 2015), Quality Technology and Quantitative Management (since 2003), and Quality Engineering (since 2001). He is the recipient of the Hunter Award (2008), Shewhart Medal (2019), Box Medal (2019), and Lancaster Medal (2021). He is a Fellow of the American Society for Quality and American Statistical Association, an elected member of the International Statistical Institute, an Academician of the International Academy for Quality, and Secretary of the International Statistical Engineering Association.

Dr. Bastian Ravesteijn is an assistant professor at the Erasmus School of Economics. His research focuses on health economics, public economics, and applied econometrics.

Dr. Joran Lokkerbol is director of the Centre of Economic Evaluation & Machine Learning at the Netherlands Institute of Mental Health and Addiction. He is the recipient of a Harkness Fellowship (2017–2018) and a Mental Healthcare Fellowship (2017–2022) funded by the Netherlands Organisation for Health Research and Development.

How to cite this article: Huberts LCE, Does RJMM, Ravesteijn B, Lokkerbol J. Predictive monitoring using machine learning algorithms and a real-life example on schizophrenia. *Qual Reliab Eng Int.* 2021;1–16. <https://doi.org/10.1002/qre.2957>