# Predicting Undesired Treatment Outcome with Machine Learning in multi-site Mental Healthcare

Kasper Van Mens, Joran Lokkerbol, Ben Wijnen, Richard Janssen, Robert de Lange, Bea Tiemens

Submitted to: JMIR Medical Informatics
on: November 15, 2022

# *Table of Contents*

# Predicting Undesired Treatment Outcome with Machine Learning in multi-site Mental Healthcare

Kasper Van Mens[1] MSc; Joran Lokkerbol[2] PhD; Ben Wijnen[3] PhD; Richard Janssen[4] PhD; Robert de Lange[5] PhD; Bea Tiemens[1] PhD

[1]Radboud University Nijmegen NL
[2]Centre of Economic Evaluation & Machine Learning, Trimbos Institute (Netherlands Institute of Mental Health) Utrecht NL
[3]Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Centre Maastricht NL
[4]Erasmus University Rotterdam, Erasmus School of Health Policy & Management / Health Care Governance Rotterdam NL
[5]Alan Turing Institute Almere NL

**Corresponding Author:**
Kasper Van Mens MSc
Radboud University
Houtlaan 4
Nijmegen
NL

## *Abstract*

**Background:** It remains a challenge to predict which treatment will work for which patient in mental healthcare.

**Objective:** The aims of this multi-site study were two-fold: 1) to predict patient's response to treatment, during treatment, in Dutch basic mental healthcare using commonly available data from routine care; and 2) to compare the performance of these machine learning models across three different mental healthcare organizations in the Netherlands by using clinically interpretable models.

**Methods:** Using anonymized datasets from three different mental healthcare organizations in the Netherlands (n = 6,452), we applied three times a lasso regression to predict treatment outcome. The algorithms were internally validated with cross-validation within each site and externally validated on the data from the other sites.

**Results:** The performance of the algorithms, measured by the AUC of the internal validations as well as the corresponding external validations, were in the range of 0.77 to 0.80.

**Conclusions:** Machine learning models provide a robust and generalizable approach in automated risk signaling technology to identify cases at risk of poor treatment outcome. Results of this study hold substantial implications for clinical practice by demonstrating that model performance of a model derived from one site is similar when applied to another site (i.e. good external validation).

(JMIR Preprints 15/11/2022:44322)
DOI: https://doi.org/10.2196/preprints.44322

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
✓ **Only make the preprint title and abstract visible.**
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

# Original Manuscript

# Predicting Undesired Treatment Outcome with Machine Learning in multi-site Mental Healthcare

**Keywords: treatment outcomes, mental health, machine learning**

*Authors*

Kasper van Mens[1,2], Msc

Joran Lokkerbol[2], Phd

Ben Wijnen[2, 3], Phd

Richard T.J.M Janssen[4,5], Phd

Robert P.J. de Lange[6], Phd

Bea Tiemens[7,8,9], Phd

1. Altrecht Mental Healthcare, Utrecht, The Netherlands
2. Centre of Economic Evaluation & Machine Learning, Trimbos Institute (Netherlands Institute of Mental Health), Utrecht, The Netherlands
3. Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Centre+, Maastricht, The Netherlands
4. Erasmus University Rotterdam, Erasmus School of Health Policy & Management / Health Care Governance, Rotterdam, The Netherlands
5. Tilburg University, Tranzo, Scientific Centre for Care and Welfare, Tilburg, The Netherlands
6. Alan Turing Institute, Almere, The Netherlands
7. Indigo Service Organization, Utrecht, The Netherlands
8. Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands
9. Pro Persona Research, Renkum, The Netherlands

*Declaration of interest*

The authors declare no conflict of interest.

## Abstract

### Background

It remains a challenge to predict which treatment will work for which patient in mental healthcare.

### Objective

The aims of this multi-site study were two-fold: 1) to predict patient's response to treatment during treatment in Dutch basic mental healthcare using commonly available data from routine care; and 2) to compare the performance of these machine learning models across three different mental healthcare organizations in the Netherlands by using clinically interpretable models.

### Method

Using anonymized datasets from three different mental healthcare organizations in the Netherlands (n = 6,452), we applied three times a lasso regression to predict treatment outcome. The algorithms were internally validated with cross-validation within each site and externally validated on the data from the other sites.

### Results

The performance of the algorithms, measured by the AUC of the internal validations as well as the corresponding external validations, were in the range of 0.77 to 0.80.

### Conclusion

Machine learning models provide a robust and generalizable approach in automated risk signaling technology to identify cases at risk of poor treatment outcome. Results of this study hold substantial implications for clinical practice by demonstrating that model performance of a model derived from one site is similar when applied to another site (i.e. good external validation).

## Introduction

*Optimizing healthcare systems*

One of the main challenges in designing an efficient healthcare system is to prevent offering too much resources to some patients and too little to others. Put differently, the challenge is to maximize the opportunity for appropriate care at an individual level [1]. The recent strive for precision or personalized medicine aims to improve healthcare systems by tailoring treatments more effectively to patients. Patients are grouped in terms of their expected treatment response using diagnostic tests or techniques [2]. However, precision medicine remains a challenge in mental healthcare because treatments are effective *on average* and it is difficult to predict for exactly whom they will work [3,4]. Stepped care principles provide a framework to allocate limited healthcare resources and have been proven to be cost-effective for depression and anxiety [5,6]. In case of stepped care, treatments start with low intensity unless there is a reason to intensify. Such reasons are identified during treatment, in which at some point there is a lack of confidence in a positive outcome given the current treatment trajectory. To this extent, routine outcome monitoring (ROM) could be used to observe patterns of early treatment response and identify which patients will probably not benefit from their current treatment [7,8].

*Identification of non-responders*

The system can be improved by earlier and more accurate identification of those non-responders, such that patients do not have to endure periods of care in which they do not improve and could potentially lose interest and drop out. On top of that, scarce healthcare resources are not wasted by engaging in treatment without the desired effect. However, misclassification comes with a cost. Incorrectly classifying patients as being in need of more intensified treatment results in the unnecessary use of healthcare resources on patients that would have benefitted from a shorter, low-intensity treatment. In many clinics in Dutch basic mental healthcare, ROM measurements are part of routine care. This raises the question, whether these ROM data could be used to provide accurate prognostic feedback and support a clinician in maximizing the opportunity for appropriate care on the individual level.

*Predicting outcome with Machine Learning during treatment*

Techniques from the field of machine learning are aimed at making accurate predictions based on patterns in data. Machine learning can help to identify robust, reproducible and generalizable predictors of treatment response [3,9–11], and has already been used in healthcare research, for example in predicting healthcare costs and outcomes [12–15]. By discovering associations and understanding patterns and trends within the data, machine learning has the potential to improve care. Machine learning permits a finer detection of which patients are at elevated risk of persisting poor and costly health outcomes, and may thus give impetus to more efficient, personalized and proactive type of mental health care. Inspired by this knowledge, the aim of this study is to use machine learning on ROM data, as feedback device, to signal which patients have an elevated risk of poor response to treatment[16]. However, the use of complex data, and associated increasingly complex models, challenges researchers to ensuring that these models are clinically interpretable rather than a “black box” [17,18].

*Independent validation*

After developing a prediction model, it is recommended to evaluate model performance in other clinical data which was not used to develop the model, as mentioned in the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)-statement. For example, such a validation would require researchers to have access to a similar dataset (i.e. in terms

of predictor variables and outcome) stemming from a similar population/clinic and compare model performance on this external, independent dataset (i.e. cross-site design). Lack of independent validation is a major limitation of the extant machine learning literature in healthcare [19]. In a recent review on machine learning for suicide prediction, the majority of studies reviewed, split the data into training and testing sets, whereas none of the studies used a cross-site design in which a model was trained using data from one site and evaluated using data from another [20]. Another recent review looking at applications of machine learning algorithms to predict therapeutic outcomes in depression, concluded that most studies did not assess out-of-sample estimates of model fit, which limited their generalizability and likely overestimated predictive accuracy [21]. Therefore, the aim of this study was two-fold: 1) to predict patient's response to treatment during treatment in Dutch basic mental healthcare using limited commonly available data from routine care; and 2) to compare the performance of these machine learning models across three different mental healthcare organizations in the Netherlands by using clinically interpretable models. By using commonly available data from routine care, technical implementation of the model in clinical practice would be straightforward.

## Methods

*Study design and data collection*

Data on mental health treatment and outcomes were collected by a data collection protocol. Mental healthcare sites from six regions in the Netherlands were involved. Patients were treated with mild to severe mental health problems, and low risk of suicide or dangerous behavior. The dataset consisted of patient records with a completed treatment in the years 2014 up to 2018. A completed treatment in this setting consists of around 5 to 12 sessions[22]. The protocol consisted of a pre-defined set of variables with clear definitions and coding for each variable. Since the database was anonymized with statistical disclosure control techniques [23], there was no need for informed consent or approval by a medical ethics committee (Dutch Civil Law, Article 7:458).

In order for treatment records to be included in this study, the availability of at least the ROM data as well as certain other variables that could be used for predictions was required. As ROM questionnaires are not mandatory in routine care, ROM data were not available for all patients at all measurements. Records were included when ROM data were available at the start, during and at the end of treatment. Three from the six participating regions had sufficient treatment records (>1,000) with non-missing values to be included in the study; N(r1) = 3,020), N(r2) = 1,484, N(r3) = 1,948. In each region, patients were treated in multiple settings in both urban and rural areas. A set of 26,912 records had to be excluded from the three sites because there was a missing ROM measurement at either the start or end, such that the outcome could not be determined, or there was no measurement from during treatment, such that early treatment response patterns could not be determined. To assess the comparability of the in- and excluded treatment records in our analysis a comparison was made regarding age, sex, diagnosis, and baseline severity between both groups (see table 1).

| | **Included (n = 6,452)** | **Excluded (n =26,912)** |
|---|---|---|
| **Sex = Female** | 4,077 (63.2) | 16,872 (62.7) |
| **Age category** | | |

| | | |
|---|---|---|
| <30 | 1,978 (30.7) | 8,671 (32.2) |
| 30-40 | 1,541 (23.9) | 6,701 (24.9) |
| 40-50 | 1,238 (19.2) | 5,298 (19.7) |
| 50-60 | 1,154 (17.9) | 4,119 (15.3) |
| 60+ | 541 ( 8.4) | 2,123 ( 7.9) |
| **Diagnosis group** | | |
| Anxiety | 2,588 (40.1) | 9,955 (37.0) |
| Depression | 2,585 (40.1) | 10,831 (40.2) |
| Other | 1,279 (19.8) | 6,126 (22.8) |
| **Total OQ-45.2 Score baseline** | 80.36 (21.18) | 80.60 (23.23) |

***Table 1.*** **Comparison of patient characteristics between the included and excluded treatment records, with mean values and (SD) for numeric values and counts and (%) for categorical variables.**

*Data description*

This study utilized treatment records, as opposed to patient records. A treatment record was started whenever a patient began treatment within one of the participating centers. As a result, some patients could have multiple treatment records (5.5% of the records were not unique). ROM assessed the development in symptom severity and functioning using the standardized Dutch version of the Outcome Questionnaire (OQ-45.2) [24]. The OQ-45.2 contains three subscales: Symptom Distress (SD), Interpersonal Relations (IR) and Social Role (SR). Psychometric properties of the Dutch OQ-45.2 are adequate [25].

The idea of this study was to support a stepped care framework by predicting, during treatment, undesired treatment outcomes at the end of a treatment. These predictions can trigger a reconsideration of the chosen treatment plan, in order to improve the probability of a desired treatment outcome after finishing the treatment. Desired treatment outcomes are highly personal and depended on the type of treatment and setting. For this study we choose to define undesired outcomes as non-improvement. Based on principles of reliable change [26], we defined non-improvement as improving less than a medium effect size on the symptomatic distress scale of the OQ-45.2 [27]. Our study used data from so called ‘basic mental healthcare’ in the Netherlands. Basic mental healthcare is cost-effective short-term mental healthcare with an average effect size of Cohen’s d=.9 [22]. Despite this high effect size, the aim of this short-term treatment of 5-12 sessions is primarily to increase self-direction and get patients back on track without care as soon as possible. In this study, individual treatment goals were unknown and therefore it was decided to define non-improvement as less than a medium effect size. This is a little more than half of the average improvement in this mental healthcare setting. Our clinical outcome was derived from the observed change in the symptom distress scale on the OQ-45.2. Patients with less than half a standard

deviation improvement in symptom severity at the end of treatment were classified to have an 'undesired clinical outcome' (called non-improvement henceforth). With the standard deviation of the SD scale in a Dutch clinical population found to be 16 [25], non-improvement was defined as a patient not improving at least 8 points on the SD scale of the OQ-45.2.

Early change was defined as the difference in ROM outcome at baseline and the first ROM during treatment. For both the summed scale scores on the OQ-45.2 as well as the individual items, early change variables were created. Besides the ROM data, a set of clinical and demographic variables were included for prediction such as main diagnosis, age and living condition. The total set consisted of 163 variables, from which 144 were related to the scores on the OQ-45.2 and 19 to the context of the patient.

*Modeling and validation strategy*

The dataset was split across all included locations, such that models could be trained on a single location and externally validated on each of the other locations. Non-improvement was predicted for each location separately based on all available predictors using least absolute shrinkage and selection operator (LASSO) models. LASSO was used both to guarantee interpretability for intended model users, as well as to facilitate explicit comparison between prediction models built in different locations. Moreover, as several measures were derived from the same questionnaire, this could have led to multicollinearity between predictors in the dataset. LASSO is a technique which has been argued to be able to deal with multicollinearity and still provide stable and interpretable estimators[28]. All numeric variables were centered and scaled.

Using 10-fold cross-validation with 10 repeats, the optimal hyperparameter was determined by considering 100 possible penalty values (i.e. lambda's) between 0.001 and 1,000. For the LASSO with the optimized penalty, the probability threshold was tuned by optimizing over F1-scores over 36 possible probability values between 0.3 and 0.65. The final LASSO model selected for each site was then applied to each of the other sites for model assessment, reporting sensitivity (sen), specificity (spec), positive predictive value (ppv), negative predictive value (npv) and Area Under the Curve (AUC) using the optimized probability threshold.

Bootstrapping was used to estimate model performance in the site in which the model was built, to have an internally validated measure of model performance to compare with the two externally validated measures of model performance by estimating confidence intervals for all performance scores (e.g. sen, spec, ppv, npv). The bootstraps were performed by sampling each dataset 1,000 times with replacement, resulting in 1,000 simulated datasets for each site. The final LASSO model of each of the three site-specific models was then applied to the bootstrapped dataset, resulting in 1,000 confusion matrices per site. Next the 2,5$^{th}$ percentile and 97,5$^{th}$ percentile for each performance indicator (i.e. sen/spec/ppv/npv) were used to determine the 95% confidence interval for each estimate.

All analyses were performed in R, a statistical language and programming environment.[29] The package *caret* was used to build the models[30]. The package *glmnet* was used to perform the LASSO regression [31]. The package *pROC* was used to analyse the area under the curves [32].

## Results

### *Demographics*

The total dataset used in the analyses contained information on 6,452 treatment records and included anonymized demographic variables, care-related variables and information about the severity and types of complaints. The characteristics of the patient populations within each site are shown in Table 2. There are notable differences between baseline symptom severity, distribution of main diagnosis and percentage of patients with a paid job between sites.

| | Region 1 (n=3,020) | Region 2 (n=1,484) | Region 3 (n=1,948) | *p-value* | *Effect size* |
|---|---|---|---|---|---|
| **Care related variables** | | | | | |
| **Non-improvement** | 1,028 (34.04) | 499 (33.63) | 577 (29.62) | 0.003 | 0.042 |
| **Treatment duration in days** | 145.19 (64.87) | 208.00 (78.35) | 205.78 (77.52) | < 0.001 | 0.534 |
| **Number of treatment sessions** | 9.73 (2.92) | 13.15 (4.03) | 11.21 (4.34) | < 0.001 | 0.585 |
| **Type and severity of complaints** | | | | | |
| **Baseline Symptom Severity Score** | 51.42 (13.94) | 52.16 (13.45) | 48.72 (13.65) | < 0.001 | 0.166 |
| **Baseline Social Role Score** | 13.76 (5.06) | 14.37 (4.97) | 13.79 (5.06) | < 0.001 | 0.081 |
| **Baseline Interpersonal Relations Score** | 15.29 (6.08) | 17.01 (6.50) | 15.28 (6.11) | < 0.001 | 0.185 |
| **Baseline Total OQ-45 Score** | 80.47 (21.25) | 83.54 (20.76) | 77.79 (21.07) | < 0.001 | 0.181 |
| **Diagnosis group** | | | | < 0.001 | 0.059 |
| Anxiety | 1,300 (43.0) | 562 (37.9) | 726 ( 37.3) | | |
| Depression | 1,142 (37.8) | 568 (38.3) | 875 ( 44.9) | | |
| Other | 578 (19.1) | 354 (23.9) | 347 ( 17.8) | | |
| **Demographic variables** | | | | | |
| **Sex = F** | 1,878 (62.2) | 934 (62.9) | 1265 ( 64.9) | 0.142 | 0.025 |
| **Age category** | | | | < 0.001 | 0.051 |
| <30 | 954 (31.6) | 505 (34.0) | 519 ( 26.6) | | |
| 30-40 | 694 (23.0) | 369 (24.9) | 478 ( 24.5) | | |
| 40-50 | 577 (19.1) | 249 (16.8) | 412 ( 21.1) | | |
| 50-60 | 556 (18.4) | 241 (16.2) | 357 ( 18.3) | | |
| 60+ | 239 ( 7.9) | 120 ( 8.1) | 182 ( 9.3) | | |
| **Origin** | | | | < 0.001 | 0.612 |
| Native | 2,838 (94.0) | 1 ( 0.1) | 343 ( 17.6) | | |
| Immigrant | 68 (2.3) | 0 ( 0.0) | 97 (5.0) | | |
| Unknown | 114 ( 3.8) | 1,483 (99.9) | 1,508 ( 77.4) | | |
| **Marital Status** | | | | < 0.001 | 0.252 |
| Not married | 1,612 (53.4) | 50 ( 3.4) | 969 (49.7) | | |
| Married | 1,052 (34.8) | 24 ( 1.6) | 747 (38.3) | | |
| Divorced / Widowed | 356 (11.8) | 8 ( 0.5) | 224 (11.5) | | |
| Unknown | 0 ( 0.0) | 1,402 (94.5) | 8 ( 0.4) | | |

| | | | | | |
|---|---|---|---|---|---|
| **Living situation** | | | | < 0.001 | 0.053 |
| Alone | 981 (32.5) | 35 ( 2.4) | 571 (29.3) | | |
| With partner | 1,638 (54.2) | 43 ( 2.9) | 1,100 (56.5) | | |
| Child | 248 ( 8.2) | 9 ( 0.6) | 186 ( 9.5) | | |
| Other | 151 ( 5.0) | 6 ( 0.4) | 83 ( 4.3) | | |
| Unknown | 2 ( 0.1) | 1391 (93.8) | 8 ( 0.4) | | |
| **Paid job** | | | | < 0.001 | 0.07 |
| Employed | 1,071 (35.5) | 392 (26.4) | 536 (27.5) | | |
| Not employed | 1,949 (64.5) | 831 (56.0) | 1,412 (72.5) | | |
| Unknown | 0 ( 0.0) | 261 (17.6) | 0 ( 0.0) | | |

***Table 2.*** **Overview of research population (n = 6,452), with mean values and (SD) for numeric values and counts and (%) for categorical variables.**

The non-zero LASSO coefficients are shown in table 3. The most important coefficients, in terms of relative coefficient size, were related to early change in the SD of the OQ-45.2 and the change on the total score of the OQ-45.2. The OQ_5 at start was the only other coefficient to be non-zero at each of the three regions. The coefficient for paid employment stands out in model R1 and age had a notable coefficient in R1 and R3. Furthermore, the models contained smaller non-zero coefficients which varied between each site (e.g., some OQ-variables were non-zero in some of the models but not in all models). The results of the (hyper) parameter tuning are shown in table 4. As shown, the threshold to define a positive class was set between 0.30 (R4) - 0.34 (R3) with lambda's varying from 0.02 (R5) to 0.16 (R3).

| | R1 | R2 | R3 |
|---|---|---|---|
| (Intercept) | -0.59 | -0.76 | -1.14 |
| Age | 0.05 | | 0.04 |
| Number of days between referral and first appointment (waiting que) | | 0.05 | |
| Employment = paid job | -0.39 | -0.02 | |
| Nuisance on job = yes, very much | | -0.12 | |
| Work absence = unkown | | | 0.11 |
| *OQ start measurement* | | | |
| Self-blame | -0.08 | -0.01 | -0.07 |
| Feeling week | | -0.01 | |
| Happiness | | 0.05 | |
| Disturbing thoughts | -0.11 | | |
| Stomach | | | -0.05 |
| Relationships | -0.01 | | |
| Sadness | -0.03 | | |
| *OQ middel measurement* | | | |
| Suicidal thoughts | | | 0.03 |
| Enjoyment | | | -0.01 |
| Relationships | -0.07 | | -0.01 |
| *OQ early change* | | | |
| Stamina | | | 0.01 |
| Satisfaction in work or school | -0.01 | | -0.05 |
| Disturbing thoughts | | | 0.03 |
| Stomach | | | |
| Hearth | 0.01 | | |
| Sleeping | 0.03 | | |
| Sadness | 0.03 | | |
| Relationships | | -0.02 | |
| Headaches | | | 0.03 |
| SD OQ-45.2 score (change) | 0.97 | 0.81 | 1.09 |
| Total OQ-45.2 score (change) | 0.07 | 0.15 | |

***Table 3.*** **Non-zero LASSO coefficients of the 3 models**

| | Lambda | Probability |
|---|---|---|
| Model Region 1 | 0,16 | 0,34 |

| | | |
|---|---|---|
| Model Region 2 | 0,03 | 0,3 |
| Model Region 3 | 0,02 | 0,32 |

***Table 4.*** **The parameter settings of the three models**

The performance of the three models is shown in Table 4. Each model (row) has been evaluated internally and two times externally. Each site (columns) has been used three times, one time for internal validation and two times for the external validation of the other models. The diagonal contains the three internal validations. The confidence intervals of the AUCs overlap, which indicate that there were no significant differences in the overall performances of the models. The AUCs of the three models in the three internal validations were 0.77 (R2) and 0.80 (R1 and R2). The AUCs of the six external validations ranged from 0.77 to 0.80.

| | **metrics** | **Validation R1** | **Validation R2** | **Validation R3** |
|---|---|---|---|---|
| **Model R1** | Sensitivity | 0.784 (0.760-0.809) | 0.762 (0.725-0.800) | 0.780 (0.747-0.813) |
| | Specificity | 0.698 (0.676-0.719) | 0.647 (0.617-0.676) | 0.673 (0.650-0.697) |
| | Pos.Pred.Value | 0.572 (0.545-0.600) | 0.522 (0.486-0.560) | 0.501 (0.471-0.534) |
| | Neg.Pred.Value | 0.862 (0.846-0.880) | 0.843 (0.818-0.868) | 0.879 (0.859-0.898) |
| | AUC | 0.799 (0.783-0.816) | 0.771 (0.746-0.794) | 0.799 (0.778-0.819) |
| **Model R2** | Sensitivity | 0.841 (0.818-0.863) | 0.824 (0.789-0.856) | 0.868 (0.844-0.896) |
| | Specificity | 0.584 (0.563-0.606) | 0.586 (0.554-0.615) | 0.548 (0.520-0.574) |
| | Pos.Pred.Value | 0.511 (0.486-0.534) | 0.502 (0.466-0.533) | 0.447 (0.419-0.477) |
| | Neg.Pred.Value | 0.877 (0.860-0.893) | 0.868 (0.841-0.892) | 0.908 (0.890-0.927) |
| | AUC | 0.782 (0.765-0.799) | 0.774 (0.749-0.798) | 0.792 (0.772-0.813) |
| **Model R3** | Sensitivity | 0.696 (0.667-0.726) | 0.673 (0.633-0.716) | 0.742 (0.705-0.779) |
| | Specificity | 0.749 (0.730-0.768) | 0.726 (0.699-0.754) | 0.732 (0.708-0.754) |
| | Pos.Pred.Value | 0.589 (0.561-0.617) | 0.554 (0.517-0.596) | 0.538 (0.503-0.573) |
| | Neg.Pred.Value | 0.827 (0.809-0.846) | 0.814 (0.789-0.841) | 0.871 (0.850-0.890) |
| | AUC | 0.787 (0.771-0.803) | 0.768 (0.744-0.792) | 0.802 (0.782-0.822) |

***Table 4.*** **Comparison of internally (diagonal) and externally validated results within each site, with 1,000 bootstrapped confidence intervals for regions 1,2,3.**

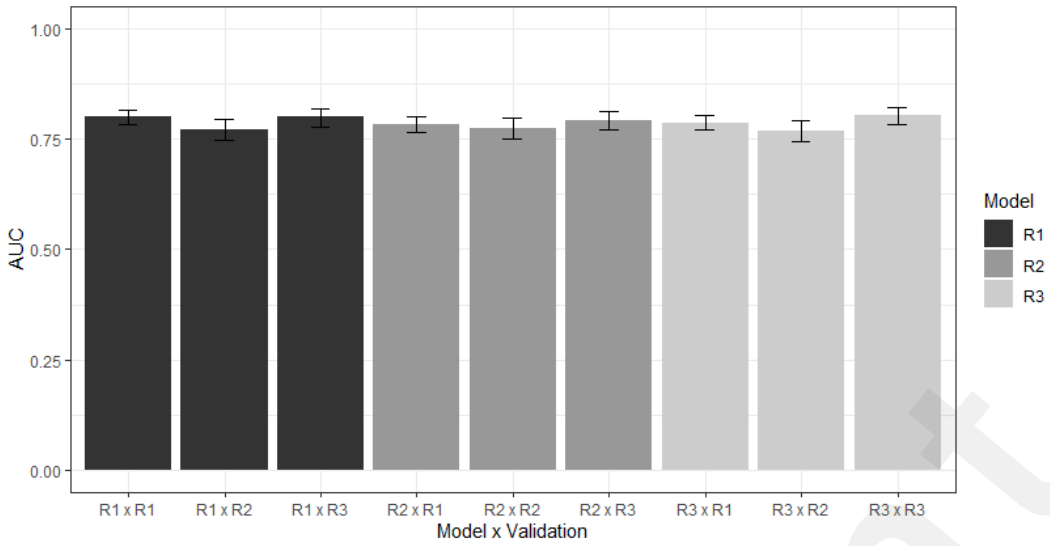


***Figure 1.*** **Comparison of the Area Under the Curve of internally and externally validated models.**

## Discussion

The aim of this study was to use machine learning to predict which patients would not substantially benefit from treatment across three different mental healthcare organizations in the Netherlands by using clinically interpretable models. This study employed a cross-site design in which model performance of a model developed in one site was compared to model performance on an external, independent dataset (i.e. 3x3 cross-site design as per the TRIPOD-statement). Data from ROM, amongst other clinical and demographic data, was used for the predictions.

*Evaluation of 3 models in 3 sites*

Both the AUC of the internal validations of the three models as well as the corresponding external validations were in the range of 0.77 to 0.80 indicating fair to good model performance [33]. In addition, confidence intervals of the AUCs overlapped in each of the nine evaluations, indicating that the performance estimates were robust and likely to be generalizable to different settings. This could be explained by the fact that LASSO regression is known to be less prone to overfitting compared to other machine learning algorithms and, when evaluated with 1,000 times bootstrapping, the internal validations give a good indication of overall performance.

All three models generalized well to the other sites. This is an interesting finding and a very promising result for the scalability of the implementation of machine learning models. Data can be gathered decentralized, within the boundaries of the general data protection regulation (GDPR). A model can be developed within the context of one site and then be exported to other sites, even if those other sites differ in certain characteristics. For example, in this research, the three sites differed in geographical location from more rural to urban. The patient populations differed, with some significant differences in the distribution of important variables such as main diagnosis, baseline symptom severity and percentage of patients with paid employment. The data sources differed in the type of EHR system used in clinical practice. Despite these substantial differences, we were able to develop three robust machine learning models with acceptable AUCs that can be applied in all three settings.

The sensitivity and specificity of the three models were consistent in each of their external validations. There were differences in these metrics between models, mainly caused by a trade-off between sensitivity and specificity when evaluating model performance with metrics from the confusion matrix. Models R1 and R2 were more shifted towards a higher sensitivity and model R3 towards a higher specificity. However, these differences were rather a shift in the balance than an ‘absolute difference’ between the models, as was indicated by the comparable AUCs.

In order to give some insight in the practical utility of the model, we translate the results can be translated to a hypothetical clinical scenario. Imagine a healthcare professional with a caseload of 30 patients working in Region 2, with a model created in Region 1. About 10 of the 30 patients will not improve according to our data (34%). The model is used by the clinician to support in identifying potential non-improving patients during treatment. With a sensitivity of 0.76 en a specificity of 0.65 (results model 1 applied to region 2), 15 patients will be classified as non-improvers and 15 will be classified as improvers. Among the improvers, 13 of them will actually improve (i.e. npv 0.84), and among the non-improvers, 8 of them would actually not improve (i.e. ppv 0.52). In half of the patients who are classified as non-improvers, therefore, the discussion would not be necessary at that time. So the question is whether these models are already good enough to actually use in practice. The idea is, however, that when the model indicates that a patient is on-track, there is little reason to change treatment. When the model indicates an elevated risk on non-improvement, the clinician and

patient should discuss the situation and adapt treatment plans if necessary. It is therefore important to see such machine learning model not as a black and white decision tool but as complementary tool in the identification and stratification of patients in need of more or less care.

*Predictive variables*

Although this research was aimed at making predictions, rather than explaining relations, we used LASSO regression in order to inform clinicians about how the algorithm works. In the healthcare setting this is important as healthcare professionals often want to understand which parameters affect and how they contribute to a prediction [34]. By looking at the coefficients of each LASSO model, it can be concluded that the algorithms rely on the variables early change in SD and the total scores of the OQ-45.2, as well as having a paid job at the start of the treatment and age. In the paper of McMahon (2014) several other studies are mentioned in which early symptom improvement, or lack of it, have been associated with psychiatric treatment outcomes [35]. In the study of Lorenzo-Lucas et al. (2017), being unemployed, amongst other factors, predicted a lower likelihood of recovery [36]. There were certain individual OQ-45.2 questionnaire items that were associated with non-zero LASSO coefficients. However these items differed between the sites and the size of the coefficients were relatively low. We are therefore reluctant to generalize findings on these individual OQ-45 items, with small non-zero coefficients, to future prediction research.

The high relative importance of the early change variable (i.e. in terms of the absolute values of the coefficients), is likely to contribute to the good external model validation as it is a straightforward defined predictor which is less likely to be subject to sampling variation. Furthermore, given the high importance of early change in the model, one could even advocate an alternative simpler predictive model (i.e. a “rule of thumb”) using early change only (or combined with less strong predictors such as age and employment status).

*Strengths and limitations*

The main strength of this study is that we used a 3x3 cross-site design to develop and evaluate the algorithms, resulting in three models with independent validation of their performance. In addition, LASSO regression was used which is a parametric approach, resulting in a prediction model that is still relatively easy to interpret. Moreover, LASSO is less prone to overfitting which increased generalizability of results.

Furthermore, with the use of a data protocol with clear data definition descriptions, we could use readily available data from routine care in the Netherlands, meaning that our approach could easily be adopted in other Dutch basic mental healthcare organizations using ROM (the R scripts to build and validate the models are available on request).

This study has a number of limitations that need to be acknowledged. First, we limited our analysis to treatment records with complete data only. In addition, we could not use every variable described in the data protocol because of missing values on these variables in one of the sites. Moreover, we had to exclude a large set of records because of missing data on the OQ-45.2. However, the excluded group of patients did not substantially differ in sex, age, diagnosis or baseline symptom severity. Nonetheless, we would like to emphasize that our models cannot be directly applied to other patient populations.

Secondly, our data did not contain information on whether the outcome of the ROM had already been used to alter the treatment strategy. This would underestimate the impact of early change as patients with only minor or no clinical improvements would have been given a possibly more intensive treatment in order for them to respond to treatment. Thirdly, although it is difficult to estimate the required sample size for developing a prognostic model, our data had a relatively small sample size [37].

A third consideration is that for this study we chose to define an undesired outcome as improving

less than a medium effect size. However, the definition of an undesired outcome is subjective and will differ between different types of treatment settings. Therefore, our definition cannot directly be generalized to other settings and each research should take an effort together with domain experts from clinical practice to define a relevant undesired outcome for that domain.
This study was performed within the context of a stepped care framework, in which treatment optimization is required during treatment. Our models heavily rely on predictors derived from early change patterns and can therefore not be applied at the start of treatment. Other research could analyze which type of predictors are more suited for a matched care framework and to what extent accurate predictions can be made in treatment response.

*Conclusion*
Machine learning models provide a robust and generalizable approach in automated risk signaling technology to identify cases at risk of poor treatment outcome. The results of this study hold substantial implications for clinical practice by demonstrating that model performance of a model derived from one site is similar when applied to another site (i.e. good external validation). This is a promising result for the scalability of machine learning models developed in single-center studies. Our findings confirm that routine monitoring provides valuable information that can be used in prognostic models to predict treatment outcomes. Such prognostic models can be used as complementary tools for practitioners in a stepped-care framework.

1. Janssen R, Van Busschbach J. Op weg naar gepaste geestelijke gezondheidszorg. *ESB*. 2012;97:81-86.
2. Fernandes BS, Williams LM, Steiner J, Leboyer M, Carvalho AF, Berk M. The new field of "precision psychiatry". *BMC Med*. 2017;15(1):80. doi:10.1186/s12916-017-0849-x
3. Gillan CM, Whelan R. What big data can do for treatment in psychiatry. *Curr Opin Behav Sci*. 2017;18:34-42. doi:10.1016/j.cobeha.2017.07.003
4. Rush AJ, Trivedi MH, Wisniewski SR, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. *Am J Psychiatry*. 2006;163(11):1905-1917. doi:10.1176/appi.ajp.163.11.1905
5. Von Korff M, Tiemens B. Individualized stepped care of chronic illness. *Cult Med*. 2000;172(February):133-137.
6. van Orden M, Hoffman T, Haffmans J, Spinhoven P, Hoencamp E. Collaborative Mental Health Care Versus Care as Usual in a Primary Care Setting: A Randomized Controlled Trial. *Psychiatr Serv*. 2015;60(1):74-79. doi:10.1176/ps.2009.60.1.74
7. Delgadillo J, Jong K De, Lucock M, et al. Feedback-informed treatment versus usual psychological treatment for depression and anxiety : a multisite , open-label , cluster randomised controlled trial. *The Lancet Psychiatry*. 2018;0366(18):1-9. doi:10.1016/S2215-0366(18)30162-7
8. Lutz W, Hofmann SG, Rubel J, et al. Patterns of early change and their relationship to outcome and early treatment termination in patients with panic disorder. *J Consult Clin Psychol*. 2014;82(2):287-297. doi:10.1037/a0035535
9. Torous J, Baker JT. Why Psychiatry Needs Data Science and Data Science Needs Psychiatry. *JAMA Psychiatry*. 2016;73(1):3. doi:10.1001/jamapsychiatry.2015.2622
10. McIntosh AM, Stewart R, John A, et al. Data science for mental health: a UK perspective on a global challenge. *The Lancet Psychiatry*. 2016;3(10):993-998. doi:10.1016/S2215-0366(16)30089-X
11. Bzdok D, Meyer-Lindenberg A. Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2018;3(3):223-230. doi:10.1016/j.bpsc.2017.11.007
12. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*. 2016;3(3):243-250. doi:10.1016/S2215-0366(15)00471-X
13. Koutsouleris N, Kahn RS, Chekroud AM, et al. Multisite prediction of 4-week and 52-week treatment outcomes in patients with fi rst-episode psychosis : a machine learning approach. *Lancet Psychiatry*. 2016;0366(16):1-12. doi:10.1016/S2215-0366(16)30171-7
14. Iniesta R, Malki K, Maier W, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *J Psychiatr Res*. 2016;78:94-102. doi:10.1016/j.jpsychires.2016.03.016
15. Lee Y, Ragguett R-M, Mansur RB, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *J Affect Disord*. 2018;241:519-532. doi:10.1016/j.jad.2018.08.073
16. Delgadillo J, de Jong K, Lucock M, et al. Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. *The Lancet Psychiatry*. 2018;5(7):564-572. doi:10.1016/S2215-0366(18)30162-7
17. Graham S, Depp C, Lee EE, et al. Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. *Curr Psychiatry Rep*. 2019;21(11):116. doi:10.1007/s11920-019-1094-0
18. Freitas AA. Comprehensible classification models. *ACM SIGKDD Explor Newsl*. 2014;15(1):1-10. doi:10.1145/2594473.2594475

19. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245. doi:10.1016/J.JCLINEPI.2015.04.005
20. Kirtley OJ, van Mens K, Hoogendoorn M, Kapur N, de Beurs D. Translating promise into practice: a review of machine learning in suicide research and prevention. *The Lancet Psychiatry*. 2022;9(3):243-252. doi:10.1016/S2215-0366(21)00254-6
21. Lee Y, Ragguett RM, Mansur RB, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *J Affect Disord*. 2018;241:519-532. doi:10.1016/j.jad.2018.08.073
22. van Mens K, Lokkerbol J, Janssen R, van Orden ML, Kloos M, Tiemens B. A Cost-Effectiveness Analysis to Evaluate a System Change in Mental Healthcare in the Netherlands for Patients with Depression or Anxiety. *Adm Policy Ment Heal Ment Heal Serv Res*. 2017;0(0):1-8. doi:10.1007/s10488-017-0842-x
23. Meindl MB, Kowarik DIA, Templ PM, Templ M, Meindl B, Kowarik A. Introduction to Statistical Disclosure Control ( SDC ). *data-analysis*. Published online 2018.
24. Lambert M, Morton J, Hatfield D, Harmon C, Hamilton S, Shimokawa K. Administration and scoring manual for the OQ-45.2 (Outcome Questionniare) (3 edn). *Wilmingt Am Prof credentialing Serv LLC*. Published online 2004.
25. De Jong K, Nugter MA, Polak MG, Wagenborg JEA, Spinhoven P, Heiser WJ. The Outcome Questionnaire (OQ-45) in a dutch population: A cross-cultural validation. *Clin Psychol Psychother*. 2007;14(4):288-301. doi:10.1002/cpp.529
26. Jacobson NS, Truax P. Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research. *J Consult Clin Psychol*. 1991;59(1):12-19. doi:10.1037/0022-006X.59.1.12
27. Cohen J. *Statistical Power Analysis for the Behavioral Sciences (2nd)*. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.
28. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Published 1996. Accessed January 29, 2023. https://www.jstor.org/stable/2346178
29. R Development Core Team. R - A language and environment for statistical computing. *Soc Sci*. 2008;2. doi:ISBN 3-900051-07-0
30. Kuhn M, Johnson K. *Applied Predictive Modeling*. Springer New York; 2013. doi:10.1007/978-1-4614-6849-3
31. Friedman J, Hastie T. Regularized paths for generalized linear models via coordinate descent (Technical Report. *Citeseer*. 2008;33(1). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.158.458%5Cnpapers://63650f89-9a27-4a9d-91f7-fca3c5048b3e/Paper/p2552
32. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77. doi:10.1186/1471-2105-12-77
33. Li F, He H. Assessing the Accuracy of Diagnostic Tests. *Shanghai Arch Psychiatry*. 2018;30(3):207-212. doi:10.11919/j.issn.1002-0829.218052
34. Hilhorst L, Stappen J van der, Lokkerbol J, Hiligsmann M, Risseeuw AH, Tiemens BG. Patients' and Psychologists' Preferences for Feedback Reports on Expected Mental Health Treatment Outcomes: A Discrete-Choice Experiment. *Adm Policy Ment Heal Ment Heal Serv Res*. 2022;49(5):707-721. doi:10.1007/S10488-022-01194-2/FIGURES/4
35. McMahon FJ. Prediction of treatment outcomes in psychiatry-where do we stand? *Dialogues Clin Neurosci*. 2014;16(4):455-464. doi:10.1164/rccm.200408-1036SO
36. Lorenzo-luaces L, Derubeis RJ, Straten A Van, Tiemens B. A prognostic index ( PI ) as a moderator of outcomes in the treatment of depression : A proof of concept combining multiple variables to inform risk- strati fi ed stepped care models. *J Affect Disord*. 2017;213(February):78-85. doi:10.1016/j.jad.2017.02.010
37. van Smeden M, Moons KGM, de Groot JAH, et al. Sample size for binary logistic prediction

models: Beyond events per variable criteria. *Stat Methods Med Res*. 2019;28(8):2455-2474. doi:10.1177/0962280218784726

## Appendix 1. Confusion matrix results

| | | *actual* | |
|---|---|---|---|
| | | non-improvement | improvement |
| *predicted* | non-improvement | 806 | 222 |
| | improvement | 602 | 1390 |

**Table 1. Confusion matrix result model 1 site 1.**

| | | *actual* | |
|---|---|---|---|
| | | non-improvement | improvement |
| *predicted* | non-improvement | 380 | 119 |
| | improvement | 348 | 637 |

**Table 2. Confusion matrix result model 1 site 2.**

| | | *actual* | |
|---|---|---|---|
| | | non-improvement | improvement |
| *predicted* | non-improvement | 450 | 127 |
| | improvement | 448 | 923 |

**Table 3. Confusion matrix result model 1 site 3.**

| | | *actual* | |
|---|---|---|---|
| | | non-improvement | improvement |
| *predicted* | non-improvement | 865 | 163 |
| | improvement | 828 | 1164 |

**Table 4. Confusion matrix result model 2 site 1.**

| | | *actual* | |
|---|---|---|---|
| | | non-improvement | improvement |
| *predicted* | non-improvement | 411 | 88 |
| | improvement | 408 | 577 |

**Table 5. Confusion matrix result model 2 site 2.**

| | | *actual* | |
|---|---|---|---|
| | | non-improvement | improvement |
| *predicted* | non-improvement | 501 | 76 |
| | improvement | 620 | 751 |

**Table 6. Confusion matrix result model 2 site 3.**

| | | *actual* | |
|---|---|---|---|
| | | non-improvement | improvement |
| *predicted* | non-improvement | 716 | 312 |
| | improvement | 499 | 1493 |

**Table 7. Confusion matrix result model 3 site 1.**

| | | *actual* | |
|---|---|---|---|
| | | non-improvement | improvement |
| *predicted* | non-improvement | 336 | 163 |
| | improvement | 270 | 715 |

**Table 8. Confusion matrix result model 3 site 2.**

| | | *actual* non-improvement | improvement |
|---|---|---|---|
| *predicted* | non-improvement | 428 | 149 |
| | improvement | 368 | 1003 |

**Table 9. Confusion matrix result model 3 site 3.**

# Supplementary Files

# Figures

Untitled.