Predicting therapy success and costs using baseline characteristics – An Approach for personalized treatment recommendations

Vincent Bremer, M.A.¹, Dennis Becker, M.Sc.¹, Spyros Kolovos, M.Sc.², Burkhardt Funk, Ph.D.¹, Ward van Breda, Ph.D.³, Mark Hoogendoorn, Ph.D.³, and Heleen Riper, Ph.D.²,⁴

¹Institute of Information Systems, Leuphana University, Lüneburg, Germany.
²Department of Clinical, Neuro- & Developmental Psychology, Vrije University, Amsterdam, Netherlands.
³Department of Computer Science, Vrije University, Amsterdam, Netherlands.
⁴Amsterdam Department of Health Sciences, Vrije University, Amsterdam, Netherlands.

Correspondence should be addressed to:
Vincent Bremer
Universitätsallee 1, C4.320
21335 Lüneburg, Germany
Phone: +49.4131.677-1157
Fax: +49.4131.677-1149
Mail: vincent.bremer@leuphana.de
Abstract

**Background:** Different treatment alternatives exist for psychological disorders. Clinical and cost-effectiveness of treatment is a crucial aspect for policy makers, therapists, as well as patients and thus plays a major role for healthcare decision-making. At the start of an intervention, however, it is often not clear which specific individuals benefit from a particular intervention alternative most and how costs will be distributed on an individual patient level.

**Objective:** This paper aimed at predicting the individual outcome and costs for patients before the start of an Internet-based intervention. Based on these predictions, treatment recommendations can be provided on an individual patient level. Thus, we expand the discussion of personalized treatment recommendation.

**Methods:** Outcomes and costs were predicted based on baseline data of 350 patients from a two-arm randomized controlled trial focusing on the comparison of treatment as usual and blended therapy for depressive disorders. For this purpose, we evaluated various machine learning techniques, compared predictive accuracy, and revealed features that contributed most to the prediction performance. We then combined these predictions and utilized the incremental cost-effectiveness ratio in order to derive individual treatment recommendations before the start of treatment.

**Results:** Predicting clinical outcomes and costs was a challenging task that came with high uncertainty when only utilizing baseline information. However, we were able to predict more accurate than a predefined reference measure in the shape of mean values of the outcome and costs. Questionnaires that include anxiety/depression items and questions regarding the mobility of individuals and their energy level contributed to the prediction performance. It was then illustrated how patients can be allocated to the individually most appropriate treatment type. For an incremental cost-effectiveness threshold of 25,000 €/quality adjusted life years, we demonstrated that our recommendations would have led to slightly worse outcomes (1.8%) but simultaneously a decrease in costs (5.9%).

**Conclusions:** Our results indicate that it was feasible to provide personalized treatment recommendations at baseline and thus allocate patients to the most beneficial treatment type. This approach can potentially lead to an improved decision-making, better outcomes for individuals, and simultaneously reduced health care costs.

**Keywords:** Treatment recommendation, Selection Criteria, Cost Effectiveness, e-Mental-Health, Machine Learning

Introduction

In a clinical context, different forms of behavioral interventions such as face-to-face or Internet-based treatments exist for a depressive disorder. Clinical and cost-effectiveness studies provide important knowledge regarding these treatment alternatives [1]. However, the question remains as to which particular individuals do not only prefer particular treatment types but potentially receive an increased benefit from a specific treatment option over the other; especially before the treatment begins. Therapists or other clinicians often decide based on personal understanding and experience - leading to high uncertainty or non-optimal decisions [1]. This uncertainty can potentially result in worse
treatment outcomes for individuals and increased health care costs. Simultaneously, policy makers and stakeholders increasingly demand cost-effectiveness evidence in order to support their conclusions and decisions [2].

For supporting these admittedly difficult and complex decisions, approaches exist based on cost analysis or decision analysis [1,3]. The incremental cost-effectiveness ratio (ICER) is a widespread indicator for cost-effectiveness [4]. The goal is to support the mentioned decisions by identifying actions that, on average, maximize a specific result [1] such as the quality adjusted life years (QALYs). The ICER is applied on a population level, which means that average values of costs and outcomes are considered for population level decisions [1,5]. This procedure does not consider any heterogeneity amongst individuals regarding outcomes and costs. Individual patients, for example, respond differently to treatment and have varying mindsets regarding risks [6,7]. Thus, the average outcomes and costs often do not necessarily represent the best decision for an individual [6]. Even though these aspects are well known, cost-effectiveness analyses is, nonetheless, still widely used on a basis of average values [6].

Predictive analyses can provide crucial insight into aspects that influence outcome and costs of interventions and can be beneficial for patients as well as the society [8]. Research that seeks to forecast outcomes in depressive patients already exists. One study, for example, predicted treatment success in the domain of depression and showed that baseline data has predictive power in this context [9]. Another study predicted treatment outcome of treatment-resistant depressive patients and thereby revealed important predictors such as severity and suicidal risk amongst others [10]. These types of statistical procedures can ultimately result in the development of decision support systems in the context of health interventions. In the field of depression treatment, these systems often lead to positive effects and even a reduction of symptoms in various situations [11].

This paper aimed at making personalized treatment recommendations. For this purpose, we predicted the outcomes and costs for different treatment types at baseline on an individual patient level. We applied various machine learning techniques, evaluated them based on their predictive performance, and revealed important features that contributed to the prediction. In order to derive personalized treatment recommendations, we applied an individualized cost-effectiveness analysis based on the ICER. Unlike traditionally utilized based on the ratio of average values, we used the individual predictions for each treatment type and its alternative. The predictions and their generated information can provide additional knowledge and enable practitioners as well as researchers to assign patients at baseline to their individually most appropriate treatment type in terms of outcomes and costs. This approach is applied to data from an Internet-based two-arm randomized controlled trial in the domain of depression.

The forecast of outcome and costs on an individual level is one of the most important aims in clinical research [12] and personalized analyses and illustration of cost-effectiveness in this context are of increased interest and need [6,13]. Thus, we contribute to existing research by attempting to predict these factors at the start of treatment for each
individual and by further proposing a conceptual approach for treatment recommendation which is applied to empirical data.

Methods

Data & Pre-Processing
The data we utilized originate from the EU funded project E-Compared in which the clinical and cost-effectiveness of blended treatment (BT) for depression, where Internet and face-to-face treatments are combined in one integrated treatment protocol, is evaluated and compared to treatment as usual (TAU) in nine different countries [14]. Participants were 18 years or older, met criteria for a major depressive disorder, were not of high suicidal risk, were not being treated for depression, and had access to an Internet connection. Table 1 illustrates the different questionnaires used in the study.

Table 1. Data utilized in this paper.

<table>
<thead>
<tr>
<th>Data</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - Demographic data</td>
<td>—</td>
</tr>
<tr>
<td>2 - Current treatment</td>
<td>Current treatment type, medication, provider</td>
</tr>
<tr>
<td>3 - M.I.N.I.</td>
<td>Structured clinical interview for making diagnoses</td>
</tr>
<tr>
<td>4 - QIDS-SR16</td>
<td>Quick Inventory of Depressive Symptomatology</td>
</tr>
<tr>
<td>5 - PHQ-9</td>
<td>Questions regarding depressive symptoms</td>
</tr>
<tr>
<td>6 - EQ-5D-5L</td>
<td>EuroQol questionnaire; measuring generic health status; for calculation of QALYs</td>
</tr>
<tr>
<td>7 - TIC-P</td>
<td>Measurement of medical costs and productivity losses</td>
</tr>
<tr>
<td>8 - Treatment Preferences</td>
<td>Individual preferences for BT or TAU</td>
</tr>
</tbody>
</table>

The data consist of information regarding depressive symptoms, medical costs, and other aspects on an individual level. Because these questionnaires are widely utilized and known, we refer for more information to [14–18]. The data in the E-Compared project were collected at multiple times during the trial; at baseline, three months, six months, and twelve months. Questionnaire three, four, six, and seven (according to Table 1) are also available not only at baseline but after the other times of data acquisition. Because we were interested in recommendations before the start of the actual treatment, we solely used the baseline information as features in this study.

As outcome we used the QALY measured by the EuroQol questionnaire (EQ-5D-5L). Utility weights were calculated by using the Dutch tariffs [19]. These weights are a preference-based measure of quality of life anchored at 0 (worst perceivable health) and 1 (perfect health). QALYs were calculated by multiplying the utility weights with the amount of time a participant spent in a particular health state. Transitions between the health states were linearly interpolated. The costs that we aimed to forecast were measured from the societal perspective (including healthcare utilization and productivity losses) based on the adapted version of the Trimbos and iMTA questionnaires on Costs Associated with Psychiatric Illness (TiC-P) [18]. Dutch unit costs were used to value healthcare utilization and productivity losses [20]. Costs for the online part of the blended
treatment included maintenance and hosting of the treatment and the costs that occurred for a therapist to provide feedback to participants. We decided to use costs from a societal perspective because they represent the society’s interest and all stakeholder groups [1]. For more information on the calculation of the costs see [21]. As dependent variables, we utilized the QALY and costs that appear after a six-month period. This allowed for more observations compared to the data at twelve months (350 patients vs 212 patients) because not all patients had already finished the treatment process. Because we focused on the outcome data up to six months, the QALY can have a maximum value of .5 in our analysis.

During the data pre-processing phase, we merged all mentioned data from Table 1. This process then led to 309 features that could be utilized for the prediction. We then calculated the costs and QALY for each individual. We only included patients for which both dependent variables were not missing. By splitting the dataset into groups for the different treatment types (TAU and BT), some factor levels of an item/feature can go missing. We removed 97 features that had just one level or were missing. Multimedia Appendix 1 lists the omitted items from the questionnaires. The resulting dataset still contained 29,568 missing values. Disregarding these values and thus deleting them would lead to a substantial decrease in observations. We therefore utilized two different methods for handling them in order to evaluate which method would perform better regarding the predictive performance. We first imputed the numeric values by sampling from a normal distribution based on the mean value and the standard deviation of the corresponding feature. We imputed the categorical predictors by sampling from the categorical distribution of those features. As a second approach, we imputed the missing values by the median (numeric variable) and mode (categorical variable). Finally, we ended up with a dataset of 350 observations/patients and 212 features. In the following, we only report the results for the latter imputation procedure. In Multimedia Appendix 2, we also demonstrate the final performances for the first imputation method. However, we decided to utilize the latter method because it leads to the best performance in terms of prediction.

**Approach & Statistical Analysis**

In order to derive individual treatment recommendations, we utilized the baseline features as input for predicting individual level outcome and costs based on the treatment type (Figure 1). We applied various machine learning techniques to evaluate which yields the highest prediction performance. As mentioned by several studies, it is beneficial to compare different statistical procedures in order to eventually find the most precise model, especially when predicting costs due to its challenging nature [8,22,23]. Because the data consist of numerous features, we applied a feature selection method to reveal variables that contribute to the prediction performance. To demonstrate how the forecasts can be beneficial in recommending treatment types on an individual patient level, we applied the incremental cost-effectiveness ratio to the predictions.
Specifically, we estimated the conditional probability \( p(o, c|b, tt) \) for each treatment type, where \( o \) is the outcome, \( c \) is the costs, \( b \) reflects the baseline features, and \( tt \) is one of the two treatment types. Given the limited amount of data, we assumed that the conditional probability can be factorized:

\[
p(o, c|b, tt) = p(o|b, tt) p(c|b, tt).
\]

For the prediction of outcome and costs, we used linear regression and support vector regression (SVR). The latter method has shown good predictive capabilities in various fields [24]. We further utilized regression trees and ridge regression. For finding the optimal parameters, we applied a grid-based search and cross-validation. Additionally, we defined the mean of all outcomes/costs as a reference measure. If unable to achieve a better prediction performance compared to the reference measure, it is questionable if the application of more advanced statistical methods is appropriate in this context. For finding the model that achieves the highest prediction performance, we used leave-one-out cross validation, that is, one observation is utilized as the test set and the remaining observations are used for training the model. This procedure is repeated for every single observation in the dataset. The error measures we used are the root mean square error (RMSE) and mean absolute error (MAE). We present both error measures because it is debated which measure is more appropriate for the demonstration of predictive performance [25,26].

When utilizing a vast number of features, overfitting presumably occurs. Thus, we used Lasso regression to select features that contribute to the predictive performance. Lasso is a linear regression that introduces a penalty term, the so-called regularizer [27]. The error function of the regression, which is to be optimized, consists of the mean square error of the misclassified samples and a term that penalizes the absolute value of the sum of regression coefficients. This linear penalty enforces useless coefficients to shrink towards zero in order to produce a sparse solution. The corresponding optimization problem is:

\[
\min_{\beta} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\},
\]
where $X$ is the baseline features, $Y$ is the outcome/costs, and $\beta$ the coefficients. The parameter $\lambda$ influences the strength of the penalty. Specifically, the higher the value of $\lambda$, the higher the penalty. A higher penalty leads to sparser solutions (more coefficients are shrunk to zero). The optimal $\lambda$'s are found by utilizing cross validation. After obtaining the specific features that appear to add to the predictive accuracy, we again predicted the outcome values and costs based on the aforementioned machine learning techniques. This time, however, we only utilized the features that were identified by the Lasso regression. Finally, we selected the algorithm that produces the smallest error and therefore performs best for the outcome and cost predictions. Based on these individual predictions, we calculated the ICER as follows:

$$ICER = \frac{\left( Cost_{BT} - Cost_{TAU} \right)}{\left( Outcome_{BT} - Outcome_{TAU} \right)}.$$ 

The ICER is then visualized in the cost-effectiveness plane [28]. By predicting the costs and outcomes at baseline and utilizing the ICER, we could then make recommendations about individual patient allocation. We implemented the mentioned models and processes in R [29].

**Results**

Before we focus on the outcome and cost predictions, we illustrate the general improvements of the patients for TAU and BT. The E-Compared project hypothesized non-inferiority between both treatment types (meaning that BT is not less effective) [14]. Improvement was defined as the difference of the start and end value of the cumulated PHQ9 values. The PHQ9 questionnaire is a reliable measure for depression severity [16]. Because we only investigated the improvements for a six-month period, these results are not final, however, they can indicate a trend. Table 2 shows that the mean baseline score for PHQ9 is 15.35 for BT and 15.42 for TAU. At the six-month measurement, the scores are 7.85 and 9.49 respectively. Furthermore, 154 patients improved in the BT group and 140 patients in the TAU group. Therefore, we can see that the PHQ9 value decreased stronger for BT and that the number of improvements for BT excels the outcome of TAU. Applying a t-test for the comparison of the Mean End values resulted in the rejection of the hypothesis that both samples have the same mean ($P = .006$).

Table 2. Mean of PHQ9 scores at baseline and end for TAU and BT as well as the numbers of patients in each condition that improved.

<table>
<thead>
<tr>
<th></th>
<th>TAU</th>
<th>BT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Start PHQ9</td>
<td>15.42</td>
<td>15.35</td>
</tr>
<tr>
<td>Mean End PHQ9</td>
<td>9.49</td>
<td>7.85</td>
</tr>
<tr>
<td>#Improvements</td>
<td>140</td>
<td>154</td>
</tr>
<tr>
<td>#No Improvements</td>
<td>38</td>
<td>18</td>
</tr>
</tbody>
</table>

**Outcome & Cost Prediction**

Table 3 illustrates the prediction performance for all utilized machine learning techniques and all baseline features. Overall, the SVR and the regression tree had the smallest errors for the performance measures. The ridge regression also performed better than the
reference measure. Based on a Wilcoxon-test, the mean absolute errors differed significantly (SVR: \( P_O = .030, \ P_C < .001 \); Tree: \( P_O = .001, \ P_C < .001 \); Ridge: \( P_O = .049, \ P_C < .023 \)). Since we had more features than observations, we did not apply OLS regression when utilizing all baseline features.

Table 3. Results for prediction performance based on all baseline features for varying machine learning approaches (MAE: mean absolute error, RMSE: root mean square error).

<table>
<thead>
<tr>
<th>Model</th>
<th>Outcome</th>
<th>Costs in €</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( MAE_O )</td>
<td>( RMSE_O )</td>
</tr>
<tr>
<td>SVR</td>
<td>0.0714</td>
<td>0.0997</td>
</tr>
<tr>
<td>Regression Tree</td>
<td>0.0698</td>
<td>0.0992</td>
</tr>
<tr>
<td>Ridge Regression</td>
<td>0.0711</td>
<td>0.1000</td>
</tr>
<tr>
<td>Reference Measure</td>
<td>0.0770</td>
<td>0.1017</td>
</tr>
</tbody>
</table>

We then performed Lasso regression in order to select the important features that contributed to the prediction performance. The tables in Multimedia Appendix 3 show the important features that were utilized and their corresponding coefficient. By applying cross-validation, we chose specific \( \lambda \) values which minimize the mean cross-validated error. For TAU and BT, we used all features up to a \( \lambda \) value of 0.01485 and 0.01479 respectively (433.83 and 651.14 for the cost prediction).

Multiple features were appearing repeatedly. Various questions regarding the medication use and the amount of consultations of some kind of therapist, practitioner, or treatment program occurred most often (24 and 16 times respectively). Furthermore, the anxiety/depression items (6 times), mobility (5 times), origin of the patient (7 times), and energy level questions (4 times) appeared to also have an influence on the prediction performance. Using the selected features, we then repeatedly applied the above specified statistical methods in order to achieve a better accuracy.

Table 4. Results for prediction performance based on selected baseline features for varying machine learning approaches (MAE: mean absolute error, RMSE: root mean square error).

<table>
<thead>
<tr>
<th>Model</th>
<th>Outcome</th>
<th>Costs in €</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( MAE_O )</td>
<td>( RMSE_O )</td>
</tr>
<tr>
<td>SVR</td>
<td>0.0575</td>
<td>0.0812</td>
</tr>
<tr>
<td>Regression Tree</td>
<td>0.0590</td>
<td>0.0793</td>
</tr>
<tr>
<td>Ridge Regression</td>
<td>0.0684</td>
<td>0.0952</td>
</tr>
<tr>
<td>Reference Measure</td>
<td>0.0770</td>
<td>0.1017</td>
</tr>
</tbody>
</table>

We observe a general increase in performance (Table 4). All statistical methods performed better than the reference measure (except for the RMSE for the linear regression and cost prediction) which is again confirmed by a significant Wilcoxon-Test.
for the mean absolute errors (SVR: $P_O < .001$, $P_C < .001$; Regression: $P_O < .001$, $P_C < .001$; Tree: $P_O = .002$, $P_C < .001$; Ridge: $P_O < .001$, $P_C < .001$). This suggests that feature selection results in more accurate predictions in this context. The overall results demonstrate that some machine learning approaches are beneficial when predicting the outcomes and costs. Since ridge regression predicts the outcome and costs best, we utilized this model in the following analysis.

Figure 2 illustrates the predicted and observed values for each treatment type and dependent variable (QALY/costs). The predictions are sorted in ascending order. The blue markers/line are the predictions and the black markers are the observed values where the y-Axis demonstrates the value of the QALY/costs and the x-Axis represents the corresponding patient. We see that the predicted outcome and costs come with high uncertainty. The broader range of the actual observations around the blue markers for the cost predictions indicates that these were more difficult to achieve compared to outcome predictions in this context. Visually, however, the trend of the predictions appears to be as expected - and as illustrated by the increased performance compared to the reference measure, this result indicates a step in the right direction.

Figure 2. Predicted and observed values for QALY and costs and both treatment types.

Treatment Recommendation
In order to derive individual treatment recommendations, we represent the differential outcomes and costs in the cost-effectiveness plane where the y-Axis is the difference
between the costs of each treatment type and the x-Axis is the difference between the clinical effect (Figure 3) [28]. Each quadrant has a different meaning. In our context, the NE quadrant represents higher costs and positive effects for BT; the SE quadrant indicates that BT is less expensive and more effective (BT dominates); the SW quadrant demonstrates the case where BT is less expensive but less effective; and the NW quadrant displays the situation where BT is more expensive and less effective (TAU dominates) [30]. As a first step, a threshold has to be defined which specifies up to which point an additional improvement is worth the costs. In the context of this paper, the monetary amount or willingness to pay for gaining one QALY differs by country [30]; the commonly used UK WTP thresholds for QALYs are between 25,000-35,000 €/QALY [31]. For this study, we used the conservative estimation of 25,000 €/QALY. A value above this threshold indicates that the treatment type is too expensive. Each patient represented by a green cross received the treatment type we would have recommended based on the prediction. On the contrary, each patient that has a red circle should have received the other treatment type based on the forecasts. Following this process, it is possible to recommend the probably most beneficial treatment type on an individual level at baseline.

Figure 3. Expected improvement for all patients in relation to costs.

Table 5 demonstrates a contingency table consisting of the patients for which we recommended a specific treatment type. Only 47.43% of all patients have been treated by the treatment type we would recommend based on our models and the particular ICER threshold.
We then calculated potential outcomes and costs on a population level assuming the patients would have been allocated according to the predictions. For patients that already received the recommended treatment type, we utilized the observed outcomes and costs. For those patients for which the actual treatment type is not recommended, we utilized the predictions of the model. Then, the QALYs would have decreased by 1.81% while at the same time a reduction in costs of 5.92% could have been achieved.

Table 5. Treatment recommendation for all patients.

<table>
<thead>
<tr>
<th></th>
<th>Recommended BT</th>
<th>Recommended TAU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Received BT</td>
<td>22.6%</td>
<td>26.6%</td>
</tr>
<tr>
<td>Received TAU</td>
<td>26.0%</td>
<td>24.8%</td>
</tr>
</tbody>
</table>

**Discussion**

**Principal Results**

Given the growth in demand of personalized treatments and the need for a reduction in costs, predictions of outcome and costs in the context of mental health are increasingly important [3]. In this paper, we proposed an approach for personalized treatment recommendations at baseline. Here, individuals are assigned to most beneficial treatment types before the commencement of the treatment which can, if desired, even be automated. We derived these recommendations by predicting patient individual QALYs and costs based on data from an EU funded project. We then used the ICER and the cost-effectiveness plane as an individualized treatment recommendation tool. Nowadays, decisions are often made based on the incremental cost-effectiveness ratio - we proposed a feasible path that allows the individualization and tailoring of this process.

We illustrated that the utilization of all baseline features is not necessarily appropriate in this context. Taking advantage of feature selection techniques can increase prediction performance. As a result, we found that consultations with some kind of therapist, medication usage, anxiety/depression information (severity), mobility items (i.e. "I have no problems in walking about"), and origin of the patient play an important role when predicting outcomes and costs in the context of digital health interventions. Therefore, including questionnaires that contain these factors and subsequently utilizing these features in statistical analyses when predicting outcomes and costs can be beneficial. We further illustrated that experimentation with different statistical methods benefits the final results since considerable varying performances occurred amongst the methods.

However, we demonstrated that the prediction is a challenging task – even though the results suggest that predictive power exists in the baseline features, our analyses indicated that the predictions and thus the recommendations come with uncertainty when only given the baseline information. In general, the predictive uncertainty is due to two sources. The first source is the uncertainty in the estimated parameters. With an increased amount of data, the uncertainty in parameter estimation reduces. This does not mean that we would achieve perfect predictions because the second source is related to the variance...
of the treatments that cannot be explained by the model. More specifically, the models do not fully represent the reality and all its complexity. Hence, although the estimation of the model parameters improves with more data, the uncertainty that results from the model specifications and inability of the baseline information to precisely predict results remains. Nevertheless, we showed that we were able to predict the outcomes and costs better compared to using the mean of the dependent variables as prediction (reference measure). Therefore, we are convinced that the baseline features do include some information regarding the forecast of outcomes and costs and can support practitioners in their decision-making process. Thus, combining these results with the ICER enabled us to provide treatment recommendation on an individual level.

As mentioned earlier, if the patients would have been allocated according to our predictions, the QALYs would have decreased by 1.81% and simultaneously a reduction in costs of 5.92% could have been achieved. These results are based on a specific ICER threshold. When applying this procedure in a real-world setting, this threshold can be adjusted to values set by experts/policy makers or available budgets. These experts must decide how many monetary resources they would want to spend on a specific QALY gain. Thus, the outcome and costs can be controlled by setting this threshold. As suggested by [32], it might also be required to model the cost-effectiveness decision rule in a non-linear form. The value of improvements may vary amongst the outcome levels, for example. Particularly, a difference between .1 and .2 on the scale might be more important than a difference between .8 and .9; even though the absolute difference is the same. The absolute severity of the symptoms can also play an additional role in this context. Possibly, it might not be justifiable to spend additional monetary effort if a specific patient already does not suffer from severe symptoms. Therefore, experts in the field need to choose appropriate values for the ICER threshold based on their experiences and knowledge and might even consider a non-linear specification.

**Limitations**

This study also comes with limitations. One limitation is the fact that we utilized data after a six-month period. Usually, the preferred outcome for cost-effectiveness analysis is based on twelve months. Another limitation, which is closely associated with the previous aspect, is the size of the used dataset. Given the complexity of the problem, it is inevitable that variations in performance occur when predicting other datasets. Thus, for achieving higher accuracy in predictions, obtaining more data is crucial. Even though our results are promising, more data and evaluations are needed in order to investigate the generalizability of these outcomes and improve predictive accuracy of statistical techniques. Besides the size of the dataset, the data are heterogeneous in different ways. For example, the data were collected from nine different European countries which each have their own country specific conditions [14]. This can result in country specific patterns in the data. Given the limited amount of observations on a national level, we have not explored this multi-level structure. Additionally, the dataset consists of a large amount of missing values that needed imputation. Making all baseline questions mandatory for the patients can lead to an increased performance of the statistical procedures and can therefore lower uncertainty.
Conclusions
This study investigated how patients can be allocated to different treatment types in order to increase clinical and cost-effectiveness. We demonstrated how to predict outcomes and costs in this context and proposed an approach for individualized treatment recommendation by utilizing the incremental cost-effectiveness ratio. Simultaneously, we evaluated a variety of machine learning techniques and demonstrated specific features that contribute to the prediction performance. The results are indicative of progress. We hope that policy makers increasingly understand the benefit of predictive modeling in this context and apply these types of models to make better and simultaneously more personalized treatment choices. We further hope we can contribute to the decision-making process in this field by providing a path that allows the prediction of eventual outcomes and costs on an individual basis before the onset of treatment.

Acknowledgements
The current study has been conducted in the context of the EU FP7 project E-COMPARED (project number 603098). We therefore thank the EU for funding and the E-COMPARED consortium for the fantastic cooperation.

Conflicts of Interest
none declared

Multimedia Appendix
Multimedia Appendix 1: Omitted items from analysis.

Multimedia Appendix 2: Results for prediction performance based on sampling from normal and categorical distribution for varying machine learning approaches.

Multimedia Appendix 3: Important baseline features based on Lasso regression for QALY and cost prediction for TAU and BT.

References
7. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of