Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Madoc Smith 04/17/2020

Inference and Prediction of Atypical Response to Major Depressive Disorder Treatment With C-Reactive Protein and Interleukin-6 Inflammatory Markers

By

Madoc Smith

MSPH

Department of Biostatistics and Bioinformatics

Mary Kelley, PhD

Thesis Advisor

Rebecca Zhang, MS

Thesis Reader

Inference and Prediction of Atypical Response to Major Depressive Disorder Treatment With C-Reactive Protein and Interleukin-6 Inflammatory Markers

By

Madoc Smith

B.S., University of Pittsburgh, 2018

B.A., University of Pittsburgh, 2018

Thesis Advisor: Mary Kelley, PhD

An abstract of

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health
in Biostatistics

2020

Abstract

Inference and Prediction of Atypical Response to Major Depressive Disorder Treatment With C-Reactive Protein and Interleukin-6 Inflammatory Markers

By Madoc Smith

Background: Major depressive disorder is a common mood disorder with complex patient response patterns over time. A patient's improvement may vary across the treatment period. We aim to predict patient response patterns in a clinically meaningful manner using inflammatory markers historically associated with depression.

Methods: In the PReDICT study (Dunlop, 2012), 316 patients were randomly assigned to three equally effective major depressive disorder treatments over a 12-week treatment period. Clinical and biological measurements were taken to assess possible predictors of patient response to treatment. We classified patients into four groups based on patient responses in early and late treatment intervals. After verifying response group patterns, we conducted pairwise comparisons of potential confounding clinical measures to control in logistic models that predict disagreement between early and late response using C-Reactive Protein (CRP) and Interleukin-6 (IL6). To better predict MDD symptom recurrence and late sustained response, we constructed minimum deviance cross-validation classification trees with inflammatory marker concentrations and potential confounders.

Results: For MDD symptom recurrence, previous antidepressant trials (P=0.050), anxiety diagnosis (P=0.104), baseline Hamilton anxiety score (P=0.077) and employment status (P=0.023) were identified as potential confounders. For late sustained response, age (P=0.064) was identified. When modeling recurrence, there is a relationship between CRP and response group (OR=1.42; P=0.053). The large effect size of IL6 (OR=0.62 P=0.162) suggests a relationship that we are underpowered to detect.

Conclusions: Unemployment and concurrent anxiety may be associated with increased likelihood of patient experience MDD symptom recurrence. Young age may be associated with increased likelihood of a patient experiencing late sustained response. CRP and IL6 appear to have a relationship with a patient experiencing atypical response patterns to MDD treatment. Prediction efforts poorly classified recurrence and late sustained response correctly likely due to small sample size in the recurrence (N=28) and late response (N=20) groups. Tree classification prediction rate was typically better than logistic models at the expense of inferential power.

Inference and Prediction of Atypical Response to Major Depressive Disorder Treatment With C-Reactive Protein and Interleukin-6 Inflammatory Markers

By

Madoc Smith

B.S., University of Pittsburgh, 2018

B.A., Univeristy of Pittsburgh, 2018

Thesis Advisor: Mary Kelley, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health
in Biostatistics

2020

I. Introduction

Major Depressive Disorder (MDD) is a common mood disorder consisting of a sustained negative emotional state that departs from a person's habitual functioning. According to Kaplan & Sadock's *Comprehensive Textbook of Psychiatry*, symptoms can manifest as daily depressed mood, diminished interest or pleasure in activities, significant weight change without diet, insomnia or hypersomnia, psychomotor agitation or retardation, fatigue, feelings of worthlessness and guilt, diminished ability to concentrate, and recurrent thoughts of death. One in five women and one in ten men will experience a depressive disorder at some point in their lifetimes (Sadock, 2017). MDD is commonly measured by the Hamilton Depression Rating Scale (HDRS), which measures melancholic and physical symptoms of depression. Higher scores on the HDRS are associated with worse MDD symptoms (Williams, 1988).

There are both therapeutic and pharmaceutical treatments for MDD. The three treatments utilized in the trial for this analysis were Cognitive-behavioral therapy (CBT), Duloxetine, and Escitalopram. CBT is a therapeutic intervention focusing on patient beliefs and challenging automatic thoughts that encourage a depressive state (Sadock, 2017). CBT is considered a valid treatment for MDD even though it cannot be performed under double-blind conditions (Berger, 2015; Zhang, 2018). Duloxetine is a serotonin norepinephrine reuptake inhibitor (SNRI), and Escitalopram is a selective serotonin reuptake inhibitor (SSRI). Duloxetine and Escitalopram show superiority to placebo in previous studies with no significantly differential patient response or time to effect relative to one another (Nierenberg, 2007; Khan, 2007; Zdanowicz, 2017). Both are considered second generation antidepressant medications with documented superiority to

earlier drugs such as tricyclic antidepressants (TCAs) (Nemeroff, 2002). There is little-to-no evidence of differential efficacy between CBT, Escitalopram, and Duloxetine (Dunlop, 2017; Kennedy, 2018; Kuyken, 2015) unless therapist quality is in question (DeRubeis, 2005). Studies suggest additional improvement when a patient is treated with medication and therapeutic interventions concurrently (DeRubeis, 2005; Huijbers, 2019; Ma, 2014; Mullen, 2018; Sung, 2015; Hollon, 2014).

Szegedi et al. (2009) found that early improvement in the first two weeks of treatment is a sensitive predictor of MDD patient outcome at the end of treatment in weeks 8-12. They recommend clinicians should change treatment strategies if treatment does not work in immediate weeks (Szegedi et al., 2009). Similar studies support early response as a predictor of patient response later in the treatment period for both CBT (Beard, 2019; Schlagert, 2017) and medication treatments (Lin, 2019) with varied definitions of early and late responses. Additionally, some studies suggest diminishing likelihood of responding the longer the patient remains unimproved (Posternak, 2011). However, using early medication change in practice for persons who do not improve immediately did not show improved outcome in a randomized trial setting (Tadic, 2016). This goes against the recommended practices outlined in Szegedi et al. (2009) but is not surprising given that early response in the first two weeks is a sensitive predictor with low specificity. For early improvement, the sensitivity – or probability of improving early given the patient responds later – is 88% [82%, 93%]. The specificity – or probability of not improving early given the patient does not respond later – is 60% [56%, 93%] (Szegedi et al. 2009).

This suggests that using this test in practice would yield many "false negative" patients who do not respond early but would have improved later. Similarly, using early improvement as a predictor for later improvement is not 100% sensitive. Thus, there are bound to be some patients who improve initially then experience a recurrence and fail to sustain that improvement later.

If we are able to identify which patients may experience MDD treatment in such a way and better understand the specific trajectories of their responses over time, we will be able to build upon Szegedi et al.'s work to guide treatment recommendations in a way that is inclusive for the many MDD patients whose early response to treatment is not indicative of their later response. Using data from PReDICT study patients, we will use longitudinal data analysis methods to understand response trajectories for MDD patients whose early treatment response within six weeks does not match their response in later treatment weeks 8-12.

Two biological inflammatory measures for depression, C-reactive protein (CRP) and interleukin-6 (IL6), were tested as possible biomarkers for atypical response. Studies have shown that high CRP levels are associated with worse depression scores (Panagiotakos et al., 2004), indicating a possible role of inflammation in depression. In support of this hypothesis, Miller et al. (2002) found significantly higher concentrations of both CRP and IL6 inflammatory markers in the depressed patients relative to the controls. If immune markers are associated with depression, they might also be associated with response to treatment, i.e. lessening depressive symptoms.

By classifying, verifying, inferring from, and predicting patient response group using these inflammatory markers, we could possibly provide clinicians with information

to better anticipate an MDD patient's response trajectory and be able to take a more informed course of action when planning a personalized treatment routine.

II. Methods

Study Details

The Predictors of Remission in Depression to Individual and Combined Treatments (PReDICT) study involves 344 treatment naïve MDD patients. Each patient is assigned to CBT, Duloxetine, or Escitalopram treatments and followed for a maximum of twelve weeks. The goal of the PReDICT study was to identify biological markers to predict patient outcome to treatment (Dunlop, 2012). The primary outcome of interest for the current analysis is response – defined as at least a 50% reduction from the baseline HDRS-17 item total score. Potential predictors from the study include inflammatory markers, neuroimaging (Dunlop et al., 2017), clinical and demographic characteristics, and genetic information (Dunlop, 2012).

Definition of Response Patterns

To assess patient response over time, it was necessary to exclude 28 patients who never returned for any follow-up appointments. For the remaining 316 patients, we defined two treatment intervals to assess "early" and "late" response trends using a benchmark of six weeks. Response outcome at each week determines whether a patient met the criterion for sustained response in the "early" (baseline to week 6) or "late"

(week 6 to week 12) intervals. We defined an "early response" as a patient responding within the first six weeks and maintaining that response in subsequent early weeks. In the event of a missing week 6 response, we accepted week 5 response as the final indicator of early response. We then defined a "late response" as a patient experiencing sustained response for each recorded (non-missing) week in weeks 8, 10, and 12. If all three weeks were missing, the late response indicator for that patient was also considered missing.

This resulted in 4 distinct patient sustained response groups based on the sustained "early response" and "late response" status of each patient. "Non-responders" had neither an early response nor a late response. "Recurrence" patients had an early response with no late response. "Late sustained responders" had no early response by week 6 but did experience a response later. "Early sustained responders" experienced an early response and kept response to treatment through the late interval. Under this coding scheme, 51 patients were missing response group assignment due to a missing response indicator in at least one interval – leaving 265 patients with group assignments for verification of longitudinal patterns and calculation of possible clinical confounders to the inflammatory marker models.

Verification of Longitudinal Patterns

Using the lme4 package in R, we constructed piecewise, bent-line regression models for each classification method to examine patient response trajectory over time by group and verify the classification scheme. We implemented one knot at week 6 to represent the cutoff of when we considered a response to be early or late, so the model

can estimate the change in slope at that time. Each model was a mixed-effect model with fixed effects for response group, week, "week after week 6", as well as an interaction of response group and time with a random intercept and slope for each subject. The mixed-effect model of depression scores over time (HDRS) is represented by:

$$HDRS_{ij} = \beta_0 + \beta_1 RespGrp_i + \beta_2 Week_j + \beta_3 (RespGrp_i \times Week_j) +$$
$$\beta_4 (Week_j - 6)_+ + \beta_5 \left(RespGrp_i \times (Week_j - 6)_+\right) + \beta_{0i} +$$
$$\beta_{2i} Week_{ij} + \beta_{3i} (RespGrp_i \times Week_{ij})$$

Where the $(Week_j - 6)_+$ term is equal to 0 when week is less than 6, i.e. the "positive part" of that term. Index "j" represents week of treatment with values 1, 2, 3, 4, 5, 6, 8, 10, or 12. Index "i" represents a subject.

Potential Confounders of Response Pattern

To evaluate if any patient characteristics may confound the inferential and predictive models of inflammatory markers on response group, we assessed means and proportions of clinical measures across response groups, performed in a pairwise fashion. More specifically, the question of interest is whether early response predicts later response. For this reason, our recommendations operate under the assumption that a clinician would already know the early response of the patient they are working with. We compare patients who respond early – the early sustained response and recurrence groups – separately from the patients who do not respond early – the non-responders and late sustained response groups.

Prediction of Endpoint Response given Early Response

We will perform prediction using two methods: traditional logistic regression and classification trees. The questions of interest are as follows:

1) Will a patient with a sustained response by week 6 become a recurrence patient or an early sustained response patient by the end of treatment?
2) Will a patient with no sustained response by week 6 become a late response patient or a non-responder by the end of treatment?

Traditional logistic regression assumes a linear relationship between the predictors (CRP, IL6) and the outcome (probability of recurrence/late sustained response), which may not exist. In contrast, classification trees classify patients with nonparametric split values to achieve the best possible prediction rate. The classification trees may discover a critical value of CRP and IL6 that predicts recurrence/late sustained response and does not appear in the linear relationship assumed by traditional logistic regression. In addition, classification trees are built with the primary goal of prediction. By using a second methodology that emphasizes prediction, we may correctly classify more patients into the response groups of interest in ways not captured by logistic regression. For these reasons, we will perform both methods to achieve our goals of prediction and clinical utility.

Logistic Regressions

Roughly 5% of inflammatory marker measures were contaminated due to a protocol deviation and removed from analysis. Our primary analysis was to test baseline CRP and IL6 measurements as possible predictors of response pattern. We performed logistic regressions that model late response in weeks 8, 10, and 12 to conduct statistical inference on the effect of inflammatory markers on response group. These models include logarithmic concentrations of both CRP and IL6 – to mitigate the influence of outlier values – as well as any marginally significant clinical measures we may find to reduce confounding.

Classification Trees

The process of classification with a tree begins at the top of the tree, which contains all subjects to be classified. Predictors are then used to form recursive binary splits that maximize the separation of the two groups. Once the process is complete, traversing down the tree based on the characteristics of a patient, we will eventually arrive at a decision as to which response group that patient belongs to – with a certain amount of misclassification error.

To build classification trees to answer these questions, we must consider what constitutes a node's goodness of split. This "Goodness of Split" term is composed of measures of the "impurity" of its two daughter nodes. In each daughter node, we would ideally have all members of one class in the left node and all members of the other class in the right node. A split would be extremely poor at classifying the outcome if each

daughter node contained equal proportions of class 1 and class 2 patients, i.e. $p = 0.5$ is the highest degree of "impurity". We measure node impurity through the Gini index:

$$Gini\ Impurity = p(1-p)$$

Where p is the proportion of patients in the class of interest at the node – in this case, the proportion of patients who have sustained response in weeks 8, 10, and 12 of treatment. The Gini impurity measure reaches a minimum (0) when p is 1 or 0, and a maximum (0.25) when p is 0.5. Impurity can also represent homogeneity, i.e. high homogeneity is measured with low Gini impurity, and low homogeneity is measured with high Gini Impurity. The goal is to have high homogeneity, which suggests good classification. Using the Gini impurity, we can calculate the "Goodness of Split" term. This term represents the reduction in impurity by splitting from parent node to two daughter nodes and is given by the equation:

$$\Delta I = i(T_P) - P(T_L)i(T_L) - P(T_R)i(T_R)$$

Where $i(T_K)$ represents Gini impurity of the parent node (T_P), the left daughter node (T_L), and the right daughter node (T_R). $P(T_L)$ and $P(T_R)$ represent the proportion of patients assigned to each daughter node by the split. Ideally, we would like to maximize ΔI to make the best split possible.

Now that we have a measure of goodness of split, we must decide which parameters we use to create each split. Additionally, we must decide which value of each parameter produces the best split. A continuous predictor with "k" unique observations would have k-1 potential splits. A nominal predictor with "k" levels would have $2^{k-1} - 1$ potential splits. At each potential split, we calculate goodness of split for every potential

value of a predictor. Once we have the best value to form a potential split for each predictor, we compare the best values of all predictors of interest and create a split at the best value of the best predictor. All predictors are considered at each new split, and the algorithm continues until no more splits are possible (saturation).

Once the tree has reached saturation, it is common practice to "prune" the tree to avoid overfitting the data and forming unnecessarily small splits at the terminal nodes of the tree. We measure tree quality with:

$$R(T) = \sum_{T \in S} P(T) r(T)$$

Where *T* represents a node in the set of terminal nodes, *S*. *r(T)* is a measure of misclassification (usually impurity), and *P(T)* is the probability of a patient being in class 1 given the patient is classified into node *T*. Therefore, *R(T)* measures the "cost" of misclassification. For the "tree" command in R, the measure of impurity used for pruning is entropy impurity which takes the form:

$$Entropy\ Impurity = -plog(p) - (1-p)log(1-p)$$

For entropy impurity, R(T) is proportional to traditional deviance from a binomial model. R(T) is then used to calculate the "cost complexity" of a tree by computing the following term for each **terminal** node:

$$R_{\alpha}(T) = R(T) + \alpha|S|$$

Where $|S|$ is the number of terminal nodes. α represents a complexity parameter that is specific to each terminal node. The α term gives us an idea of what the loss incurred by

pruning the tree at node *T* would be. Each α is calculated with the data used to build the tree via the following formula:

$$\alpha = \frac{R^S(T) - R^S(T_\tau)}{|T_\tau| - 1}$$

Where $R^S(T)$ represents the deviance of node *T*, $R^S(T_\tau)$ is the sum of deviances of *T*'s offspring nodes, and $|T_\tau|$ is the number of offspring nodes. T_τ can be thought of as an optimal subtree extending from node *T* (Zhang & Singer 2010).

Once a tree is calculated to saturation, we can then use that tree to calculate a sequence of α parameters with corresponding optimal subtrees (Zhang & Singer 2010). The best tree under this cost complexity model would be selected based on either minimizing deviance *R(T)* or minimizing misclassification rate. However, given that the α parameter is calculated with the same data used to build the tree, the optimal tree would be better judged on an independent sample using new data with α values computed from the original data. In lieu of another independent test sample, we perform cross-validation of the existing sample and use the average deviance across samples to decide which tree is optimal based upon minimizing deviance.

Cross-validation is the process of breaking the original dataset into "k" equal subsets, using "k-1" of those subsets to build a model (in this case, a sequence of α parameters with corresponding optimal subtrees), and using the remaining subset of the original data to test the sequence of α parameters. This concept is much like bootstrapping, where the data is resampled numerous times to get an estimate of the underlying distribution of data. Through using cross-validation, a clearer optimal number of nodes to minimize deviance will appear.

We then select the minimum deviance tree indicated by our cross-validation efforts and report it as the optimal tree for classifying later patient response. This will provide future studies with nonparametric insight as to which predictors might have the most influence on prediction and, additionally, which **values** of various predictors might be considered critical values for analyses regarding patient response group classification.

Because trees tend to make many and, in some cases, non-meaningful splits on continuous variables, we will construct two classification trees for each question: one will use continuous CRP and IL6 values in order to see if there are any possible thresholds that properly classify patient late sustained response, and the other will use more clinically meaningful cutoffs for CRP and IL6. According to CDC, the critical values of interest of CRP are: <1mg/L, 1-3mg/L, 3-10mg/L, and >10mg/L – which each indicate increasing risk of cardiovascular disease. IL6 does not have similar clinically defined levels, so we use quantiles from the total sample to create 4 distinct levels of IL6.

III. Results

Verification of Longitudinal Patterns

We observed 108 (40.5%) “Non-Responders”, 38 (14.3%) “Recurrence” patients, 29 (10.9%) “Late sustained response” patients, and 90 (34.0%) “Early sustained response” patients. The fitted values of the piecewise linear regression model predicting HDRS verify the expected response group trends over time.

From the interaction term of response group and week after week 6, we ascertain that the “Recurrence” group experienced a statistically significant change in slope –

initially starting with a decrease in symptoms, followed by a subsequent increase in symptoms after week 6 ($\beta = 1.82$ (0.21), $P < 0.001$) [Figure 1]. The "Non-Responders" and "Early Sustained Response" groups experienced a statistically significant decrease in rate of improvement after week 6 ($\beta = 0.25$ (0.11), $P = 0.024$; $\beta = 1.61$ (0.16), $P < 0.001$ respectively), but continued to improve in depressive symptoms after week 6 nonetheless. The "Late Sustained Response" group is the only group where we did not observe a statistically significant change in slope after week 6 ($\beta = 0.10$ (0.24), $P=0.68$), which indicates no difference in rate of improvement before and after the 6^{th} week.

This model behavior verifies the expected group HDRS trends over time. Estimates of each response group's rate of change in symptoms before and after week 6 is provided in Table 1. The recurrence patient group improves initially followed by a worsening of depressive symptoms later in treatment. Non-responders consistently improve slightly across the treatment period – which is understandable given some non-responders experience response yet do not meet the criteria for a sustained response in either the early or late treatment intervals. The early sustained response group's reduction in improvement after week 6 is likely due to many of these patients achieving response by the late interval, and thus, they cannot reduce depressive symptoms much more. Late sustained response patients did not experience a significant change in slope – which may represent the opposing forces of this group's improvement later in treatment against the mitigation of improvement rate due to prior sporadic individual responses in early treatment.

Potential Confounders of Response Pattern

Table 2 contains comparisons of clinical measures by valid response group (265 of the original 316 patients), and mean comparisons of inflammatory marker log concentration by response group. Sample size was reduced to 74 early sustained response patients, 28 recurrence patients, 90 non-responders, and 20 late sustained response patients on subsequent analyses (modeling) using the inflammatory markers due to missing data.

When comparing the early sustained response group to the recurrence group, there are significantly more people in the recurrence group who are unemployed ($p=0.023$) and had antidepressant trials in the past ($p=0.05$). Although this patient population was supposed to be treatment naïve, this suggests there is an association with past treatment history and later response to treatment. Additionally, losing response later in treatment for unemployed patients suggests social determinants influencing longitudinal response to treatment. Baseline Hamilton anxiety score was marginally significant ($p=0.077$) between these two groups, which suggests a possible link between higher Hamilton anxiety (HAM-A) scores at baseline. Concurrent anxiety diagnosis was near marginal significance ($p=0.104$) as well. Due to the correlated nature of baseline HAM-A score and anxiety diagnosis, two adjusted models were constructed with the idea that including concurrent anxiety diagnosis (Model 2) and baseline Hamilton anxiety score (Model 3) would be repetitive to include in the same model. Neither inflammatory marker had significantly different mean log concentrations between these two groups.

When comparing the non-responders to the late response group, there are no significantly different demographic variables at the 5% significance level. However, we

see some marginal significance with a lower mean age (p=0.064) in the late response group, indicating there might be a lower likelihood of late response across patients who are older. No inflammatory markers had significantly different mean log concentrations.

Prediction of Endpoint Response given Early Response

Logistic Regressions

Logistic regressions that model recurrence for patients with early response (Models 1, 2, and 3) are given in Table 3. There were no significant predictors of recurrence at the 5% significance level in any of the models, however the effect sizes for both CRP and IL6 remain substantial and near significance. CRP (OR=1.33, p=0.081) IL6 (OR=0.62, p=0.162). In terms of prediction, Model 1 correctly classifies 72.5% of patients. However, this model does not correctly classify any patient from the recurrence group, which is our goal.

After adjustment for covariates, the significance of CRP improves (p=0.065 in Model 2, p=0.053 in Model 3) while the significance of IL6 lessens (p=0.239 in Model 2, p=0.268 in Model 3). This suggests that we are underpowered to detect a linear relationship of low IL6 on recurrence, but there might be a threshold value that correctly classifies these patients, which will be explored later in tree construction. Although we detected a linear effect of CRP, logistic regression provided minimal classification of the recurrence group; only 5 of 28 (17.9%) recurrence patients were correctly classified in the best predictive model (Model 2).

Table 4 contains the models for late sustained response in patients who did not have an early sustained response in the first six weeks. We again see relatively strong effect sizes for both IL6 (OR=0.51, p=0.116) and CRP (OR=1.23, p=0.27). It is worth noting that although the model has prediction rate of 81.8%, it fails for our purposes as all patients are classified as non-responders regardless of CRP or IL6 values. Adjusting for age improves the CRP signal (OR=1.29, p=0.172), but once again fails by not correctly predicting any late sustained response patients.

Classification and Regression Trees

Recurrence:

Figures 2 and 3 provide two separate minimum deviance decision trees that aim to predict recurrence in individuals who had an early response. Figure 2's tree was constructed using only CRP and IL6 raw baseline concentration values to determine if there is a threshold that may correctly classify patients, with all relevant confounders as possible predictors (employment status, number of AD trials, and concurrent anxiety diagnosis). Figure 2's tree uses only immune measures and classifies 84.2% of patients correctly. However, of the three paths that classify someone as a recurrence patient rather than an early sustained response patient, only one of these paths has any clinical utility. Patients with a CRP value greater than 4.45 mg/L (high CRP) and an IL6 value less than 2.31 mg/L (low or normal IL6) were classified as recurrence. The other two paths to recurrence do not have much clinical utility because they rely on having CRP concentrations within a certain normal range. Since the primary goal of classification

trees is prediction rather than inference, this type of model loses some clinical interpretability by design.

In an attempt to force clinical interpretability into the classification tree, we generated the results in Figure 3 with similar variables as input, but now using meaningful thresholds for immune measure splits. Since R's tree function treats ordinal variables as factors, every clinically meaningful threshold value was coded as an indicator. When we prune the tree and use cross-validation to minimize deviance, we see that the clinically meaningful CRP and IL6 cutoffs are not included in Figure 3's tree's path to recurrence classification. The only path to recurrence in this model is classifying patients who are both unemployed and have a concurrent anxiety diagnosis as recurrence patients. This provides a correct prediction rate of 73.3%.

Late sustained response:

For prediction of late sustained response, we first constructed a tree using continuous values of CRP and IL6 along with age, the only relevant confounder. However, in this case, the cross-validated tree failed because it classified all patients as non-responders, and therefore, the tree was not considered further. We then fit our clinically relevant cutoffs, along with age, in a second tree [Figure 3]. This tree correctly classifies 84.5% of the patients, which is a slight improvement to classifying all patients as non-responders (81.8% correct). However, the path to late response has no clinical utility. As it classifies a patient as a late sustained responder if they are 29 or 30 years old with an IL6 measure less than the 75^{th} percentile of our data (IL6 < 1.93 mg/L).

IV. Discussion

An important clinical recommendation from this study comes from the observation that 25% of our patients are “atypical” responders whose early response to treatment does not match their later response. Early response alone should not be used to plan treatment because there are many patients for whom that might not be enough in practice. More importantly, it is the patients who do not have typical patterns that we wish to target for management, not those who are responding well to treatment.

We were able to verify longitudinal response patterns to show there is a group of patients that experiences initial improving depression symptoms followed by a subsequent worsening of symptoms in later weeks of treatment. Clinical measures associated with a greater probability of recurrence are unemployment, previous antidepressant trials, and – possibly – concurrent anxiety diagnosis. Patients with these characteristics might have a higher risk of losing response to treatment after their initial reduction in symptoms. Also, there might be a greater chance of responding late for younger patients even if they do not respond early in treatment.

In terms of traditional modeling of odds of recurrence with CRP and IL6, the significance of CRP at $p=0.053$ in Model 3 and the reasonably large effect sizes of both markers throughout suggest a relationship of inflammatory markers and response group. We may be underpowered to detect the influence of IL6 due to a sample size of less than 30 in both response groups of interest: recurrence and late sustained response. Furthermore, that signal of IL6 may be in fact nonlinearly related to response group. All attempts at modeling recurrence using traditional regression had poor prediction with a 24.8% misclassification rate in the best predictive model.

When we attempted to sacrifice some interpretability for more predictive power by using classification and regression trees (CART), we did find a cutoff for CRP and IL6 that might have clinical utility. More specifically, CRP levels greater than 4.45 mg/L and low-to-normal IL6 levels below 2.31 mg/L correctly predicted recurrence in 21.4% of recurrence patients – better than the best predictive logistic regression model for recurrence. The gains in predictive accuracy through the other paths to recurrence classification in the tree come at the expense of interpretability, as the ranges captured as predictive are all within the normal range. This illustrates one of the primary drawbacks of tree (CART, random forest) methodology for use in scientific research, where interpretation is the goal. However, the fact that the clinically relevant tree shows that unemployed persons with anxiety can be classified as recurrence or early sustained response with 73.3% accuracy is likely useful in clinical practice. In contrast, prediction of late response did not provide clinically useful results in either types of analysis.

A potential limitation of this study is that there is inherent difficulty in finding biomarkers to predict rare responses, especially with a limited sample size. It would likely be easier to identify predictors of patients who respond to treatment quite well, like the early sustained response group, and patients who respond to treatment quite poorly, like the non-responders. However, the recurrence group and late response group are of great interest clinically because physicians wish to know whether to change or not to change treatment strategy early on.

Additionally, we are limited by classifying patients into arbitrary groups that reflect what is a continuous phenomenon. Most notably, the late sustained response group with only 29 patients had a response trajectory that resembled the sustained early

response group. The small sample size of this group may have led the trajectory to be misrepresentative of the underlying trend of late improvement. Losing even more sample size with missing inflammatory marker data made it quite difficult to predict recurrence (N=28) and late sustained response (N=20). Future studies may classify response groups in different ways, which may identify new predictors of atypical response. With larger recurrence and late sustained response groups, another study might be able to draw out different associations between non-responders and late sustained response patients that we were unable to detect.

This analysis clearly exhibits a tradeoff of inferential power and predictive power. The regression models almost always had a lower correct prediction rate than the tree models, but the regressions excel in interpretability. With the logistic regressions, we could see the significance of predictors as well as the effect sizes and interpret the influence on recurrence and late response with odds ratios. The tree always had a better prediction rate relative to the logistic regressions when it was free to classify patients using CRP and IL6 as it pleased with little clinical interpretability. The choice of whether to value interpretability or prediction accuracy depends entirely on the classification problem at hand. We believe that in this clinical setting, results lose meaning if they cannot be applied to help clinicians determine if a new MDD patient outside of this existing dataset will experience a recurrence or a late response. For this reason, we value inferential power of the logistic regression models in this study over the predictive power of the classification trees.

Works Cited

Beard, J. I. L., & Delgadillo, J. (2019). Early response to psychological therapy as a predictor of depression and anxiety treatment outcomes: A systematic review and meta-analysis. Depress Anxiety, 36(9), 866-878. doi:10.1002/da.22931

Berger, D. (2015). Double blinding requirement for validity claims in cognitive-behavioral therapy intervention trials for major depressive disorder. Analysis of Hollon S, et al., Effect of cognitive therapy with antidepressant medications vs antidepressants alone on the rate of recovery in major depressive disorder: a randomized clinical trial. *F1000Res, 4*, 639. doi:10.12688/f1000research.6954.1

DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., . . . Gallop, R. (2005). Cognitive therapy vs medications in the treatment of moderate to severe depression. *Arch Gen Psychiatry, 62*(4), 409-416. doi:10.1001/archpsyc.62.4.409

Dunlop, B. W., Binder, E. B., Cubells, J. F., Goodman, M. M., Kelley, M. E., Kinkead, B., . . . Mayberg, H. S. (2012). Predictors of remission in depression to individual and combined treatments (PReDICT): study protocol for a randomized controlled trial. *Trials, 13*, 106. doi:10.1186/1745-6215-13-106

Dunlop, B. W., Kelley, M. E., Aponte-Rivera, V., Mletzko-Crowe, T., Kinkead, B., Ritchie, J. C., . . . Team, P. R. (2017). Effects of Patient Preferences on Outcomes in the Predictors of Remission in Depression to Individual and Combined Treatments (PReDICT) Study. *Am J Psychiatry, 174*(6), 546-556. doi:10.1176/appi.ajp.2016.16050517

Hollon, S. D., DeRubeis, R. J., Fawcett, J., Amsterdam, J. D., Shelton, R. C., Zajecka, J., . . . Gallop, R. (2014). Effect of cognitive therapy with antidepressant medications vs antidepressants alone on the rate of recovery in major depressive disorder: a randomized clinical trial. *JAMA Psychiatry, 71*(10), 1157-1164. doi:10.1001/jamapsychiatry.2014.1054

Huijbers, M. J., Wentink, C., & Speckens, A. E. M. (2019). Preventive cognitive therapy could be a viable and effective addition to antidepressant medication in preventing relapse or recurrence in major depressive disorder. *Evid Based Ment Health, 22*(1), e7. doi:10.1136/ebmental-2018-300054

Khan, A., Bose, A., Alexopoulos, G. S., Gommoll, C., Li, D., & Gandhi, C. (2007). Double-blind comparison of escitalopram and duloxetine in the acute treatment of major depressive disorder. *Clin Drug Investig, 27*(7), 481-492. doi:10.2165/00044011-200727070-00005

Kelley, M. E., Dunlop, B. W., Nemeroff, C. B., Lori, A., Carrillo-Roa, T., Binder, E. B., . . . Mayberg, H. S. (2018). Response rate profiles for major depressive disorder: Characterizing early response and longitudinal nonresponse. Depress Anxiety, 35(10), 992-1000. doi:10.1002/da.22832

Kennedy, J. C., Dunlop, B. W., Craighead, L. W., Nemeroff, C. B., Mayberg, H. S., & Craighead, W. E. (2018). Follow-up of monotherapy remitters in the PReDICT study: Maintenance treatment outcomes and clinical predictors of recurrence. *J Consult Clin Psychol, 86*(2), 189-199. doi:10.1037/ccp0000279

Kuyken, W., Hayes, R., Barrett, B., Byng, R., Dalgleish, T., Kessler, D., . . . Byford, S. (2015). Effectiveness and cost-effectiveness of mindfulness-based cognitive

therapy compared with maintenance antidepressant treatment in the prevention of depressive relapse or recurrence (PREVENT): a randomised controlled trial. *Lancet, 386*(9988), 63-73. doi:10.1016/S0140-6736(14)62222-4

Lin, C. H., Park, C., & McIntyre, R. S. (2019). Early improvement in HAMD-17 and HAMD-7 scores predict response and remission in depressed patients treated with fluoxetine or electroconvulsive therapy. J Affect Disord, 253, 154-161. doi:10.1016/j.jad.2019.04.082

Ma, D., Zhang, Z., Zhang, X., & Li, L. (2014). Comparative efficacy, acceptability, and safety of medicinal, cognitive-behavioral therapy, and placebo treatments for acute major depressive disorder in children and adolescents: a multiple-treatments meta-analysis. *Curr Med Res Opin, 30*(6), 971-995. doi:10.1185/03007995.2013.860020

Miller, G. E., Stetler, C. A., Carney, R. M., Freedland, K. E., & Banks, W. A. (2002). Clinical depression and inflammatory risk markers for coronary heart disease. Am J Cardiol, 90(12), 1279-1283. doi:10.1016/s0002-9149(02)02863-1

Mullen, S. (2018). Major depressive disorder in children and adolescents. *Ment Health Clin, 8*(6), 275-283. doi:10.9740/mhc.2018.11.275

Nemeroff, C. B., & Owens, M. J. (2002). Treatment of mood disorders. *Nat Neurosci, 5 Suppl*, 1068-1070. doi:10.1038/nn943

Nierenberg, A. A., Greist, J. H., Mallinckrodt, C. H., Prakash, A., Sambunaris, A., Tollefson, G. D., & Wohlreich, M. M. (2007). Duloxetine versus escitalopram and placebo in the treatment of patients with major depressive disorder: onset of antidepressant action, a non-inferiority study. *Curr Med Res Opin, 23*(2), 401-416. doi:10.1185/030079906X167453

Panagiotakos, D. B., Pitsavos, C., Chrysohoou, C., Tsetsekou, E., Papageorgiou, C., Christodoulou, G., . . . study, A. (2004). Inflammation, coagulation, and depressive symptomatology in cardiovascular disease-free people; the ATTICA study. Eur Heart J, 25(6), 492-499. doi:10.1016/j.ehj.2004.01.018

Posternak, M. A., Baer, L., Nierenberg, A. A., & Fava, M. (2011). Response rates to fluoxetine in subjects who initially show no improvement. J Clin Psychiatry, 72(7), 949-954. doi:10.4088/JCP.10m06098

Sadock, B. J., Sadock, V. A., & Ruiz, P. (2017). Kaplan & Sadocks comprehensive textbook of psychiatry. Philadelphia: Wolters Kluwer.

Schlagert, H. S., & Hiller, W. (2017). The predictive value of early response in patients with depressive disorders. Psychother Res, 27(4), 488-500. doi:10.1080/10503307.2015.1119329

Sung, S. C. (2015). Cognitive therapy plus medication management is better than antidepressants alone for patients with severe depression. *Evid Based Ment Health, 18*(3), 95. doi:10.1136/eb-2014-102012

Szegedi, A., Jansen, W. T., van Willigenburg, A. P., van der Meulen, E., Stassen, H. H., & Thase, M. E. (2009). Early improvement in the first 2 weeks as a predictor of treatment outcome in patients with major depressive disorder: a meta-analysis including 6562 patients. *J Clin Psychiatry, 70*(3), 344-353. doi:10.4088/jcp.07m03780

Tadic, A., Wachtlin, D., Berger, M., Braus, D. F., van Calker, D., Dahmen, N., . . . Lieb, K. (2016). Randomized controlled study of early medication change for non-

improvers to antidepressant therapy in major depression--The EMC trial. Eur Neuropsychopharmacol, 26(4), 705-716. doi:10.1016/j.euroneuro.2016.02.003

Williams, J. B. (1988). A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry*, *45*(8), 742-747. doi:10.1001/archpsyc.1988.01800320058007

Zdanowicz, N., Reynaert, C., Jacques, D., Lepiece, B., & Dubois, T. (2017). Selective Serotonergic (SSRI) Versus Noradrenergic (SNRI) Reuptake Inhibitors with and without Acetylsalicylic Acid in Major Depressive Disorder. *Psychiatr Danub*, *29*(Suppl 3), 270-273. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28953776

Zhang, H., & Singer, B. (2010). Recursive partitioning and applications (2nd ed.). New York: Springer.

Zhang, Z., Zhang, L., Zhang, G., Jin, J., & Zheng, Z. (2018). The effect of CBT and its modifications for relapse prevention in major depressive disorder: a systematic review and meta-analysis. *BMC Psychiatry*, *18*(1), 50. doi:10.1186/s12888-018-1610-5

Appendix

Table 1 – Slope Estimates for Each Response Group Before and After Week 6

Estimated HAM-D Change with 1 Week increase by Response Group	Before Week 6	After Week 6
Non-Responders	-0.65 (0.06)	-0.41 (0.08)
Recurrence	-1.70 (0.10)	0.37 (0.12)
Late Sustained Response	-1.32 (0.11)	-0.98 (0.15)
Early Sustained Response	-2.14 (0.06)	-0.28 (0.08)

Table 2 – Pairwise Comparison of Baseline Clinical and Biological Measures Across Response Groups

Clinical Measures (N=265)		Early Sustained Response (N=90)	Recurrence (N=38)	Pairwise P-Value	Non-Responders (N=108)	Late Sustained Response (N=29)	Pairwise P-Value
Age	Mean (SD)	39.3 (10.7)	38.1 (12.0)	0.582	41.8 (12.0)	37.2 (11.0)	**0.064**
Previous Episodes (N-Missing = 1)	1 (vs. 2 or more)	48 (53.9%)	20 (52.6%)	0.893	56 (51.9%)	12 (41.4%)	0.317
Race (N-Missing = 17)	NH White	42 (50.6%)	19 (52.8%)	0.427	54 (54.0%)	15 (51.7%)	0.504
	NH Black	12 (14.5%)	8 (22.2%)		21 (21.0%)	4 (13.8%)	
	Hispanic	29 (34.5%)	9 (25.0%)		25 (25.0%)	10 (34.5%)	
Melancholic Type	Yes	50 (55.6%)	22 (57.9%)	0.807	57 (52.8%)	14 (48.3%)	0.667
Past AD Trials*	1 or more	3 (3.3%)	5 (13.2%)	**0.050**	9 (8.3%)	1 (3.4%)	0.688
Married	Yes	54 (60.0%)	19 (50.0%)	0.296	49 (45.4%)	17 (58.6%)	0.205
Sex	Female	43 (47.8%)	23 (60.5%)	0.187	66 (61.1%)	18 (62.1%)	0.925
High School Education*	Yes	82 (91.1%)	35 (92.1%)	0.855	96 (88.9%)	25 (86.2%)	0.690
Employed (N-Missing = 1)	Yes	50 (56.2%)	13 (34.2%)	**0.023**	48 (44.4%)	12 (41.4%)	0.768
Chronic Episodes (N-Missing = 3)	Yes	29 (33.0%)	10 (26.3%)	0.459	35 (32.7%)	9 (31.0%)	0.864
Anxiety Diagnosis	Yes	29 (32.2%)	18 (47.4%)	**0.104**	49 (45.4%)	16 (55.2%)	0.348
Family History of Depression	Yes	32 (35.6%)	11 (28.9%)	0.470	38 (35.2%)	10 (34.5%)	0.944
Baseline HAM-D	Mean (SD)	19.16 (3.77)	19.53 (3.42)	0.603	19.51 (3.46)	19.24 (3.70)	0.716
Baseline HAM-A	Mean (SD)	14.94 (4.62)	16.55 (4.74)	**0.077**	16.14 (5.00)	15.41 (5.44)	0.498
Biological Markers (N=212)		Early Sustained Response (N=74)	Recurrence (N=28)	Pairwise P-value	Non-Responders (N=90)	Late Response (N=20)	Pairwise P-value
C-Reactive Protein Log Concentration Mean (SD)		-0.13 (1.73)	0.46 (1.86)	0.131	-0.01 (1.87)	0.14 (1.44)	0.736
Interleukin-6 Log Concentration Mean (SD)		0.22 (0.76)	0.11 (0.80)	0.523	0.18 (0.77)	-0.02 (0.63)	0.270

Bold values indicate P <= 0.1

Table 3 – Logistic Regressions Modeling Recurrence for Patients with Early Response

Model 1: Inflammatory Measures only								
Variable		Beta Estimate	Exp(β)	Std. Error	Pr(>|Z|)		Correct Prediction Rate: 72.5%	
B_0	Intercept	-0.93	0.39	0.23	**<0.001**		Recurrence	Early Sustained
B_1	C-Reactive Protein at Timepoint 1 (Log Concentration)	0.28	1.33	0.16	0.081	Test Positive	0	0
B_2	Interleukin-6 at Timepoint 1 (Log Concentration)	-0.47	0.62	0.34	0.162	Test Negative	28	74
						% Correct	0%	100%

Model 2: Adjusted for Potential Confounding with Anxiety Diagnosis								
Variable		Beta Estimate	Exp(β)	Std. Error	Pr(>|Z|)		Correct Prediction Rate: 74.3%	
B_0	Intercept	-0.82	0.44	0.37	**0.027**		Recurrence	Early Sustained
B_1	C-Reactive Protein at Timepoint 1 (Log Concentration)	0.33	1.38	1.84	0.065	Test Positive	5	3
B_2	Interleukin-6 at Timepoint 1 (Log Concentration)	-0.43	0.65	-1.18	0.239	Test Negative	23	70
B_3	Employed	-0.89	0.41	-1.77	0.077	% Correct	17.9%	95.9%
B_4	One or More Previous AD Trials	1.44	4.23	1.59	0.113			
B_5	Concurrent Anxiety Diagnosis	0.48	1.62	0.99	0.323			

Model 3: Adjusted for Potential Confounding with Baseline Hamilton Anxiety Score								
Variable		Beta Estimate	Exp(β)	Std. Error	Pr(>|Z|)		Correct Prediction Rate: 69.3%	
B_0	Intercept	-1.49	0.22	0.87	0.088		Recurrence	Early Sustained
B_1	C-Reactive Protein at Timepoint 1 (Log Concentration)	0.35	1.42	0.18	**0.053**	Test Positive	3	6
B_2	Interleukin-6 at Timepoint 1 (Log Concentration)	-0.41	0.67	0.37	0.268	Test Negative	25	67
B_3	Employed	-0.90	0.41	0.50	0.072	% Correct	10.7%	91.8%
B_4	One or More Previous AD Trials	1.59	4.91	0.90	0.078			
B_5	Baseline HAM Anxiety Score	0.05	1.06	0.05	0.286			

N = 102 for these models due to missing values of inflammatory markers; Bold values indicate P <= 0.05

Table 4 – Logistic Regressions Modeling Late Sustained Response for Patients with No Early Response

Model 4: Inflammatory Measures only								
Variable		Beta Estimate	Exp(β)	Std. Error	Pr(>|Z|)	Correct Prediction Rate: 81.8%		
B_0	Intercept	-1.46	0.23	0.25	**<0.001**		Late Sustained Response	Non-Responders
B_1	C-Reactive Protein at Timepoint 1 (Log Concentration)	0.20	1.23	0.18	0.270	Test Positive	0	0
B_2	Interleukin-6 at Timepoint 1 (Log Concentration)	-0.68	0.51	0.43	0.116	Test Negative	20	90
						% Correct	0%	100%

Model 5: Adjusted for Potential Confounding								
Variable		Beta Estimate	Exp(β)	Std. Error	Pr(>|Z|)	Correct Prediction Rate: 81.8%		
B_0	Intercept	0.29	1.34	0.88	0.743		Late Sustained Response	Non-Responders
B_1	C-Reactive Protein at Timepoint 1 (Log Concentration)	0.26	1.29	0.19	0.172	Test Positive	0	0
B_2	Interleukin-6 at Timepoint 1 (Log Concentration)	-0.67	0.51	0.44	0.124	Test Negative	20	90
B_3	Age	-0.05	0.96	0.02	**0.048**	% Correct	0%	100%

N = 110 for these models due to missing values of inflammatory markers; Bold values indicate $P < 0.05$

Figure 1 – Bent-line Regression Fit Trends for Each Response Group HDRS over Time

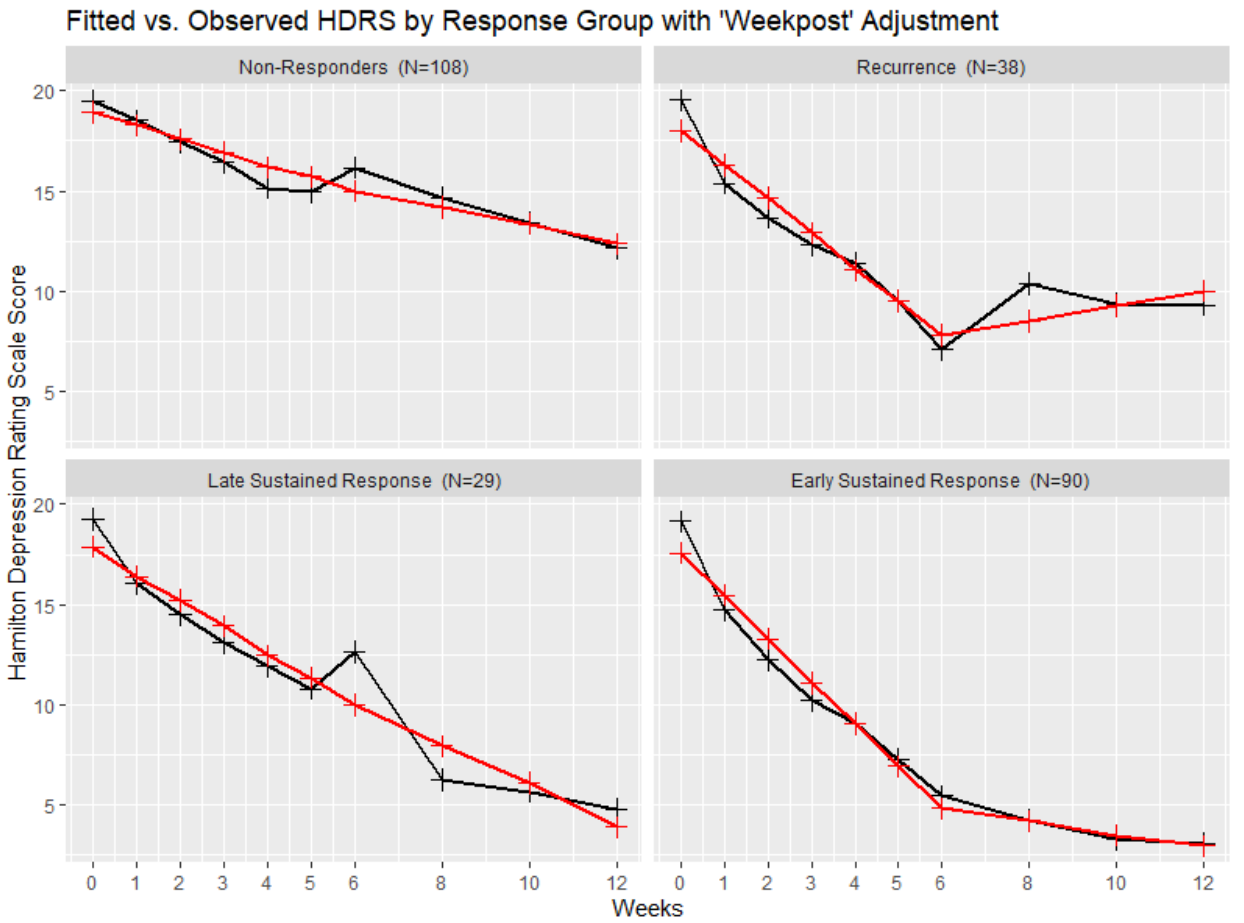
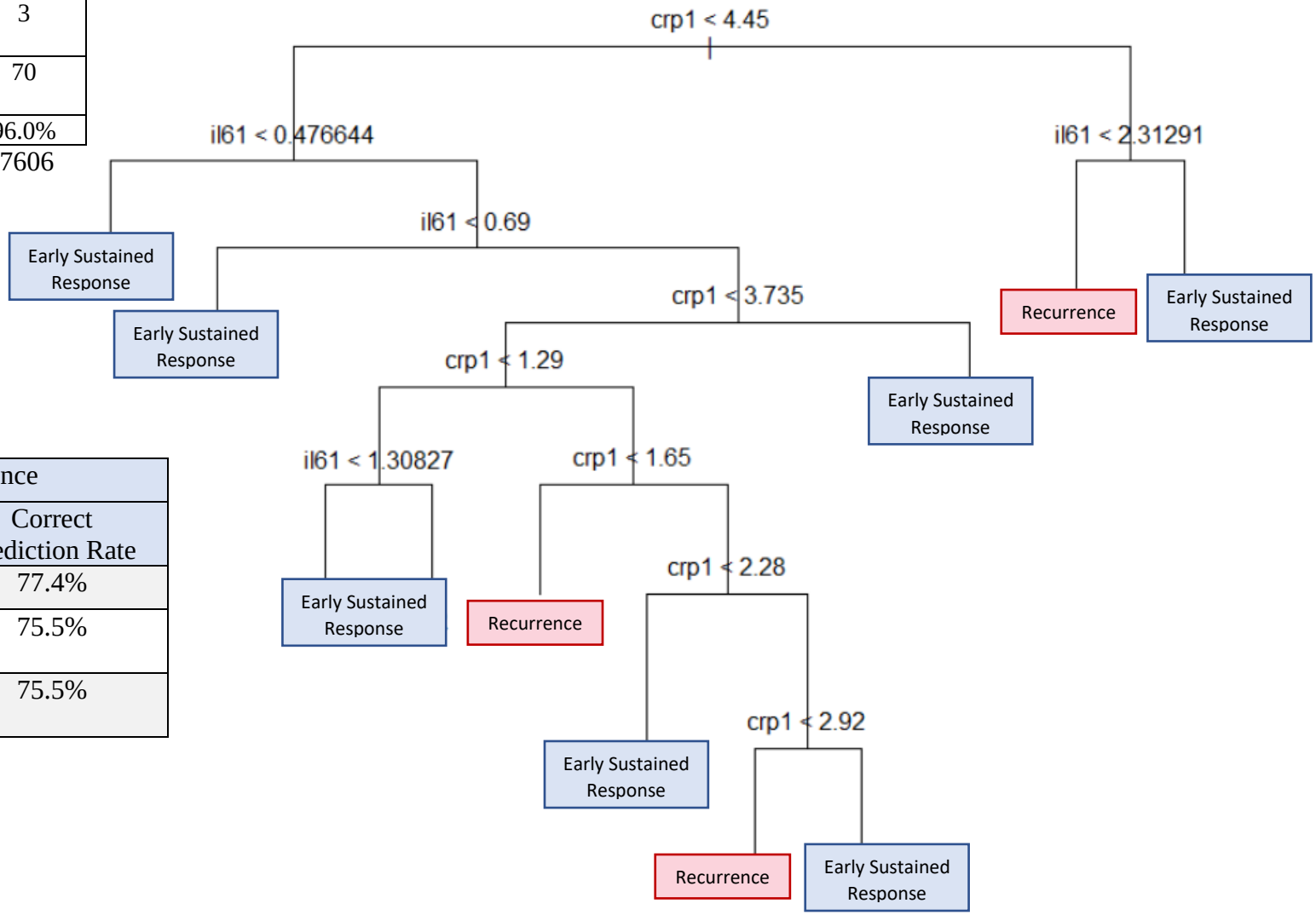


Figure 2 – Classification Tree for Early Response Group vs. Recurrence Group

Potential Predictors: *C-Reactive Protein Concentration at Baseline, Interleukin-6 Concentration at Baseline*

Tree Correct Prediction Rate: 84.2%		
	Recurrence	Early Sustained
Test Positive	15	3
Test Negative	13	70
% Correct	53.6%	96.0%

Residual Mean Deviance: 0.7606



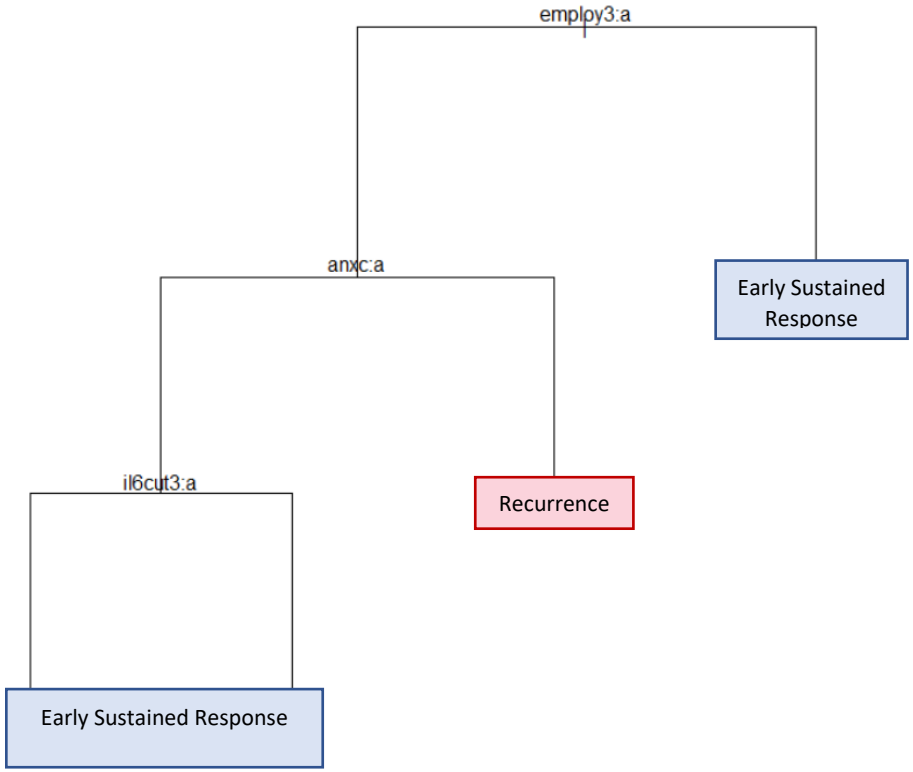
Paths to Recurrence		
CRP (mg/L)	IL6 (mg/L)	Correct Prediction Rate
> 4.45	< 2.31	77.4%
< 1.65 > 1.29	> 0.69	75.5%
< 2.92 > 1.29	> 0.69	75.5%

Figure 3 – Classification Tree for Early Response Group vs. Recurrence Group with Clinical Cutoffs Adjusting for Confounding

Potential Predictors: *C-Reactive Protein Clinical Cutoffs, Interleukin-6 Quantile Cutoffs, Employment Status, Number of AD Trials, Concurrent Anxiety Diagnosis*

Tree Correct Prediction Rate: 73.3%		
	Recurrence	Early Sustained
Test Positive	10	9
Test Negative	18	64
% Correct	35.7%	87.7%

Residual Mean Deviance: 1.119



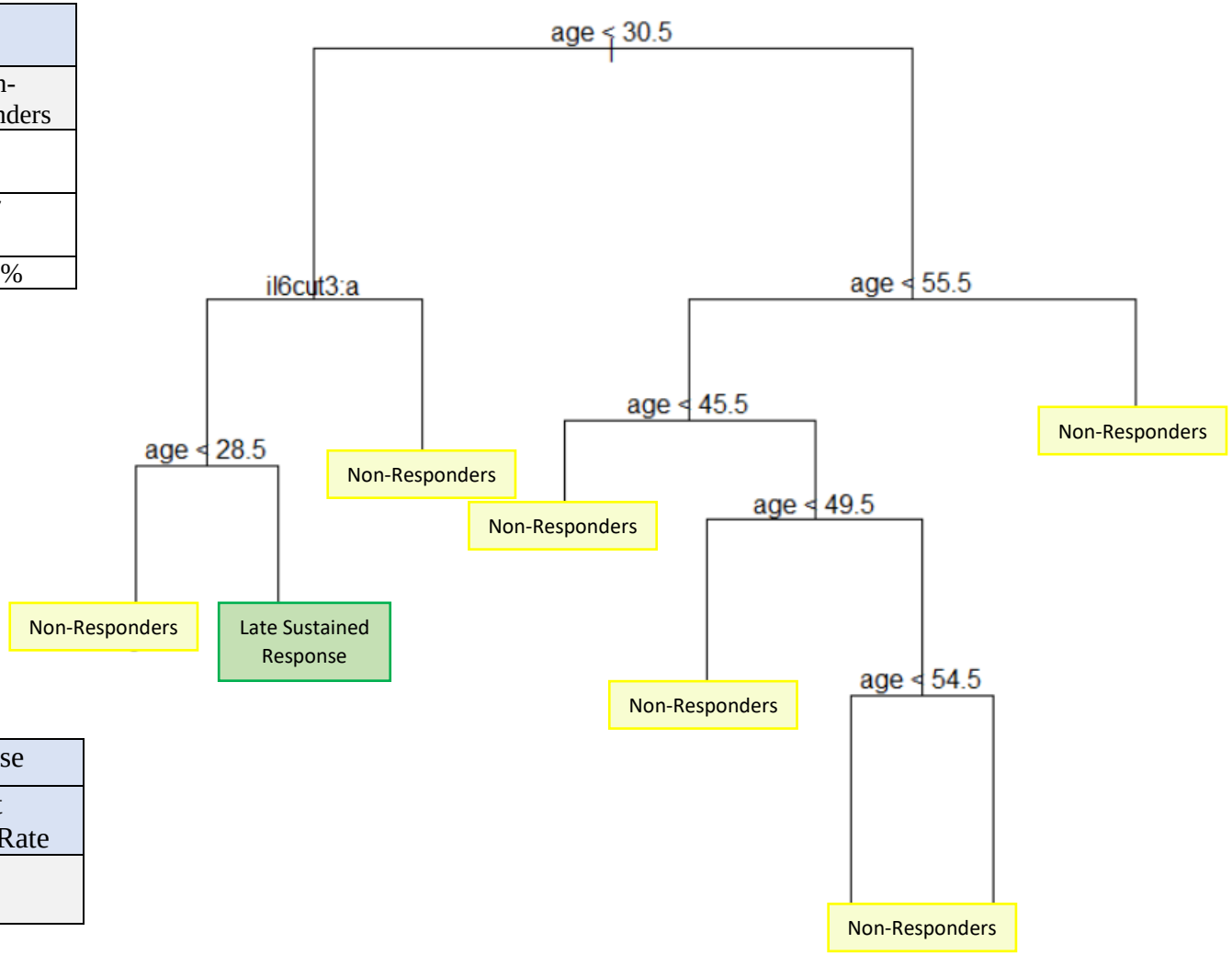
Paths to Recurrence		
Employed	Anxiety	Correct Prediction Rate
No	Yes	73.3%

Figure 4 – Classification Tree for Non-Responders vs. Late Sustained Response Group with Clinical Cutoffs Adjusting for Confounding

Potential Predictors: *C-Reactive Protein Clinical Cutoffs, Interleukin-6 Quantile Cutoffs, Age*

Tree Correct Prediction Rate: 84.5%		
	Late Sustained Response	Non-Responders
Test Positive	6	3
Test Negative	14	87
% Correct	30.0%	96.7%

Residual Mean Deviance: 0.7067



Paths to Late Sustained Response		
Age (years	IL6 (mg/L)	Correct Prediction Rate
< 31 > 28	<1.93	84.5%