Identifying Important Risk Factors for Future Mania in Individuals with Major Depressive Disorder Using Weighted Random Forest Models and the NESARC Dataset

Jialiang Hua
Advisor: Melanie M. Wall
Columbia University Mailman School of Public Health

1. Overview and Student Role

This project aimed to identify important risk factors for future mania in individuals with Major Depressive Disorder (MDD). This was accomplished by utilizing data from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) and applying the Weighted Random Forest algorithm. This study is of paramount importance as it allows for more targeted interventions, symptom surveillance, and potential preventive research for individuals with unipolar depression who are at the highest risk for a manic episode, and consequently, progression to bipolar disorder.

As the student conducting this research, my responsibilities included:

1. Reviewing relevant literature to understand the existing knowledge and research gaps in the prediction of future mania in individuals with major depressive disorder.
2. Collaborating with my faculty advisor to formulate research questions and objectives.
3. Familiarizing myself with the NESARC dataset. Learning about the structure of the data and coding of the variables.
4. Preprocessing the NESARC dataset, handling missing data, ensuring data quality, and creating descriptive statistics tables to summarize and explore the dataset.
5. Selecting appropriate methods to handle imbalanced data. Utilizing cross-validation and grid search to train three weighted random forest models with different sets of predictors. Employing the parallel computing technique to expedite the modeling process.

6. Evaluating the models' performance using AUC to examine how random forest deals with different types of variables.
7. Analyzing the models' results to identify important predictors of future mania. Comparing the top important variables of the three models. Creating partial dependence plots to study the marginal relationship between the risk of mania and the variables.
8. Comparing our results with prior studies and discussing the implications of our findings.
9. Writing the final report and communicating the research findings with faculty and peers.

2. Background

Bipolar disorder is a complex psychiatric condition characterized by alternating episodes of mania and depression.[1] Most individuals with bipolar disorder experience one or more major depressive episodes prior to their first manic episode.[2] Identifying individuals with unipolar depression who are at high risk for a manic episode, and potential progression to bipolar disorder, allows for targeted interventions, symptom surveillance, and possible preventative research.

Previous studies have attempted to identify risk factors for the transition from unipolar depression to bipolar depression. Cross-sectional studies have identified several features associated with a higher risk for bipolar disorder, including more depressive episodes and earlier onset of illness,[3-7] greater depressive severity,[5] atypical depressive symptoms,[8,9] psychotic symptoms,[6,10] and family history of bipolar disorder.[6,7] However, due to the nature of the cross-sectional study design, these studies do not necessarily distinguish features present before the first manic episode, such as psychiatric comorbidity,[4] and symptoms of bipolar depression may not remain stable across episodes.[11]

A study published in 2012 utilized the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) dataset and the logistic regression model to identify demographic, clinical and psychosocial risk factors for mania among individuals with major depressive disorder (MDD), indicating diagnostic conversion from MDD to bipolar I disorder. This study identified various risk factors for the transition from MDD to bipolar disorder, including younger age, Black race/ethnicity, less than high school education, prior psychopathology (social phobia and generalized anxiety disorder), and environmental

stressors (history of child abuse and past-year problems with one's social support group). However, the overall predictive power of these risk factors based on AUC was modest.[12]

Given the modest predictive power of the 2012 study and the limitations of logistic regression models in handling large number of variables and correlations among them, there is a need for further research employing different statistical models. One such alternative model is the random forest model, which is a robust and flexible machine learning technique that can handle large datasets, complex interactions among variables, and nonlinear relationships. By conducting a new study using the random forest model and the NESARC dataset with the same goal, we can potentially improve the prediction of bipolar disorder among individuals with MDD and identify important risk factors that may have been overlooked by previous studies. Our new study could potentially lead to more targeted interventions and preventive measures for those at a higher risk of transitioning from unipolar depression to bipolar disorder.

3. Methods

This study utilized the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), a nationally representative household survey of 43,093 participants. The survey covers a wide range of topics, including alcohol, drug, and psychiatric disorders, as well as associated risk factors and consequences. The first wave of the NESARC was conducted in 2001-2002, with a follow-up three years later comprising 34,653 of the original participants.

The analytic sample for our study included all Wave 1 participants diagnosed with Major Depressive Disorder (MDD), as defined in NESARC by lifetime MDD history without any lifetime manic or hypomanic episodes. This criterion yielded a sample of 5219 participants, with 341 developing mania and 4878 not developing mania in the survey's three-year follow-up period. Our aim was to identify the demographic, clinical and psychosocial determinants of the transition from MDD to bipolar disorder based on the NESARC dataset.

Considering the class imbalance in the dataset (mania : not developing mania = 341 : 4878), we employed the Weighted Random Forest algorithm, recognized for its effectiveness in handling imbalanced data. The traditional Random Forest algorithm aims to minimize the

overall error rate, which can overlook minority classes due to their sparse representation.[13] However, the Weighted Random Forest approach addresses this by assigning higher penalties on misclassifying the minority class,[13] enhancing the model's ability to accurately predict outcomes for underrepresented groups.

The Weighted Random Forest method follows the idea of cost-sensitive learning. We assigned a weight to each class, with the minority class given a larger weight, (i.e., higher misclassification cost). The weights assigned to each class in our study were derived from the prevalence of mania in our sample. The weight assigned to the mania class was equal to the prevalence of the majority class (not developing mania), while the weight assigned to the not developing mania class was equal to the prevalence of the minority class (developing mania) in the sample. The class weights were incorporated into the RF algorithm in two places. In tree induction procedure, class weights are used to weight Gini criterion for finding splits. In the terminal nodes of each tree, class weights were again considered. The class prediction of each terminal node was determined by a "weighted majority vote" (i.e., the weighted vote of a class is the weight for that class times the number of cases for that class at the terminal node). The final class prediction for the random forest was determined by aggregating the weighted vote from each individual tree, where the weights are average weights in the terminal nodes.[13]

The variables for our model were selected based on previous literature, specifically from the previously mentioned 2012 logistic regression model study. We included in our model participant demographic characteristics, clinical characteristics of depression, lifetime psychiatric disorders, and past-year stressors **(Table 1)**. We constructed three weighted random forest models using cross-validation and grid search for hyperparameter optimization based on the AUC metric, with each model incorporating a different set of selected predictors. The first model used only categorical variables, with all continuous variables transformed into categorical ones (tertiles). The second model incorporated both categorical and continuous variables when applicable (a change of 12 variables from categorical to continuous out of the 59 in model 1). The third model utilized processed variables created by our knowledge. The aim was to identify risk factors for a manic episode occurring during the study's three-year follow-up period.

4. Results

The Area Under the Curve (AUC) were 0.689, 0.701 and 0.688 for the categorical variable model, continuous variable model and pre-processed variable model, respectively **(Table 2)**. The continuous variable model demonstrated the highest AUC, suggesting superior predictive ability. Interestingly, the pre-processed variable model maintained a stable AUC even after reducing the number of variables from 59 to 21 through variable collapsing.

The top important variables identified by our models included age at onset of the first episode, age, annual family income, education level, and the number of lifetime depressive episodes. The continuous variable model displays higher importance scores for continuous variables compared to their categorical counterparts in the categorical variable model. **(Fig 1)**. The partial dependence plots from the categorical variable model indicated a higher risk of transitioning to mania with lower levels of education and decreased annual family income. Notably, the lower probability associated with the income group $24-19,999 compared to the income group $20,000 - $34,999 may be attributed to the small sample size in the former group, leading to a greater variation in the results. Earlier onset of the first depressive episode, a greater number of lifetime depressive episodes and black race/ethnicity (compared to white, native American and Asian) were associated with higher risks for transition from MDD to bipolar disorder **(Fig 2)**. When examining the partial dependence plots for the continuous variable model, the results echoed the findings from the categorical model. For the variables of education level, annual family income, age at onset and number of lifetime depressive episodes, the continuous variable model reflected similar trends. Additionally, a younger age was associated with a higher risk of transition from MDD to bipolar disorder **(Fig 3)**.

Furthermore, comparisons between the partial dependence plots of the categorical variable model and the continuous variable model revealed strong alignment between the two in terms of the relationship between the risk of mania and the variables of education level, age at onset, annual family income, and number of lifetime depressive episodes **(Fig 4)**. The continuous variable model, however, provided a more nuanced understanding of these relationships due to the finer range of variable values.

Lastly, scatter plots of the predicted probabilities of the models against each other showed that for true mania cases, the continuous variable model produced higher prediction probabilities than both the categorical and pre-processed variable models **(Fig 5)**. This further supports the superior predictive capacity of the continuous variable model in our study.

5. Conclusions/Discussion

The findings from our study align strongly with existing research on the risk factors for transitioning from Major Depressive Disorder (MDD) to bipolar disorder.[3-10, 12] The most important risk factors identified by our models include earlier onset of first depressive episode, younger age, lower annual family income, black race/ethnicity, lower education level, and increased number of lifetime depressive episodes. This continuity between our study and previous research supports the validity of these factors in predicting the transition from MDD to bipolar disorder.

Our study demonstrates the potential of the Weighted Random Forest approach in handling imbalanced datasets. Class weights are a critical tuning parameter to achieve desired performance, and the out-of-bag estimate of the accuracy from the random forest can be used to select these weights.[13] In this study, we used the class weights derived by prevalence, which might not be optimal. Tuning this parameter could potentially lead to enhanced performance.

Our continuous variable model showed superior predictive power, producing a higher AUC value and prediction probabilities for true mania cases compared to both the categorical and pre-processed variable models. However, the AUC is still round 0.7, indicating better-than-chance prediction, but not sufficiently high for use as a clinical screening tool.[12] The rise in variable importance for continuous variables compared to their categorical counterparts is likely to be attributed to the random forest algorithm's inherent feature selection for continuous variables.[14]

Interestingly, the AUC of the pre-processed variable model remained stable even after significant data reduction, which suggests its potential for further optimization with domain

knowledge. Collaborating with domain experts could enhance the model's performance by providing more informed guidance on variable collapsing and selection.

The consistency of our findings with a previous study that used the same NESARC dataset but a logistic regression model approach underscores the robustness of these findings across different methodologies. Future studies could build upon this work by further refining the predictive models and investigating potential interventions targeting the identified risk factors to reduce the risk of transitioning from MDD to bipolar disorder.

6. References

[1] Anderson, I. M., Haddad, P. M., & Scott, J. (2012). Bipolar disorder. BMJ (Clinical research ed.), 345, e8508. https://doi.org/10.1136/bmj.e8508

[2] Perlis, R. H., Delbello, M. P., Miyahara, S., Wisniewski, S. R., Sachs, G. S., Nierenberg, A. A., & STEP-BD investigators (2005). Revisiting depressive-prone bipolar disorder: polarity of initial mood episode and disease course among bipolar I systematic treatment enhancement program for bipolar disorder participants. Biological psychiatry, 58(7), 549–553. https://doi.org/10.1016/j.biopsych.2005.07.029

[3] Mitchell, P. B., Goodwin, G. M., Johnson, G. F., & Hirschfeld, R. M. (2008). Diagnostic guidelines for bipolar depression: a probabilistic approach. Bipolar disorders, 10(1 Pt 2), 144–152. https://doi.org/10.1111/j.1399-5618.2007.00559.x

[4] Schaffer, A., Cairney, J., Veldhuizen, S., Kurdyak, P., Cheung, A., & Levitt, A. (2010). A population-based analysis of distinguishers of bipolar disorder from major depressive disorder. Journal of Affective Disorders, Mar 9.

[5] Perlis, R. H., Brown, E., Baker, R. W., & Nierenberg, A. A. (2006). Clinical features of bipolar depression versus major depressive disorder in large multicenter trials. The American journal of psychiatry, 163(2), 225–231. https://doi.org/10.1176/appi.ajp.163.2.225

[6] Andreasen, N. C., Grove, W. M., Coryell, W. H., Endicott, J., & Clayton, P. J. (1988). Bipolar versus unipolar and primary versus secondary affective disorder: which diagnosis takes precedence?. Journal of affective disorders, 15(1), 69–80. https://doi.org/10.1016/0165-0327(88)90011-0

[7] Solomon, D. A., Leon, A. C., Maser, J. D., Truman, C. J., Coryell, W., Endicott, J., Teres, J. J., & Keller, M. B. (2006). Distinguishing bipolar major depression from unipolar major

depression with the screening assessment of depression-polarity (SAD-P). The Journal of clinical psychiatry, 67(3), 434–442. https://doi.org/10.4088/jcp.v67n0315

[8] Serretti, A., Mandelli, L., Lattuada, E., Cusin, C., & Smeraldi, E. (2002). Clinical and demographic features of mood disorder subtypes. Psychiatry research, 112(3), 195–210. https://doi.org/10.1016/s0165-1781(02)00227-5

[9] Benazzi F. (2003). Clinical differences between bipolar II depression and unipolar major depressive disorder: lack of an effect of age. Journal of affective disorders, 75(2), 191–195. https://doi.org/10.1016/s0165-0327(02)00047-2

[10] Guze, S. B., Woodruff, R. A., Jr, & Clayton, P. J. (1975). The significance of psychotic affective disorders. Archives of general psychiatry, 32(9), 1147–1150. https://doi.org/10.1001/archpsyc.1975.01760270079009

[11] Perlis, R. H., Ostacher, M. J., Uher, R., Nierenberg, A. A., Casamassima, F., Kansky, C., Calabrese, J. R., Thase, M., & Sachs, G. S. (2009). Stability of symptoms across major depressive episodes in bipolar disorder. Bipolar disorders, 11(8), 867–875. https://doi.org/10.1111/j.1399-5618.2009.00764.x

[12] Gilman, S. E., Dupuy, J. M., & Perlis, R. H. (2012). Risks for the transition from major depressive disorder to bipolar disorder in the National Epidemiologic Survey on Alcohol and Related Conditions. The Journal of clinical psychiatry, 73(6), 829–836. https://doi.org/10.4088/JCP.11m06912

[13] Chen, C. (2004). Using random forest to learn imbalanced data (Doctoral dissertation). Department of Statistics, University of California, Berkeley.

[14] Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. BMC bioinformatics, 8, 25. https://doi.org/10.1186/1471-2105-8-25

Tables and graphs:

Category	Variables
Demographic factors	age, sex, race, educational attainment
Clinical characteristics of depression	age at onset, first depression onset in the past year, presence of atypical features (3 variables), number of lifetime depressive episodes, high family history loading of depression (14 variables)
Lifetime psychiatric disorders	panic disorder, social phobia, generalized anxiety disorder, alcohol dependence (2 variables), nicotine dependence, cannabis abuse or dependence (2 variables), other substance dependence (17 variables)
Past-year stressors	social support group problems (3 variables), social environmental problems (2 variables), occupational problems (4 variables), economic problems

Table 1 *Variables included in the models*

Model	Categorical variable model	Continuous variable model	Pre-processed variable model
AUC	0.689	0.701	0.688
Mtry	3	4	1
Minimum node size	6	12	10

Table 2 *Model performance comparison*

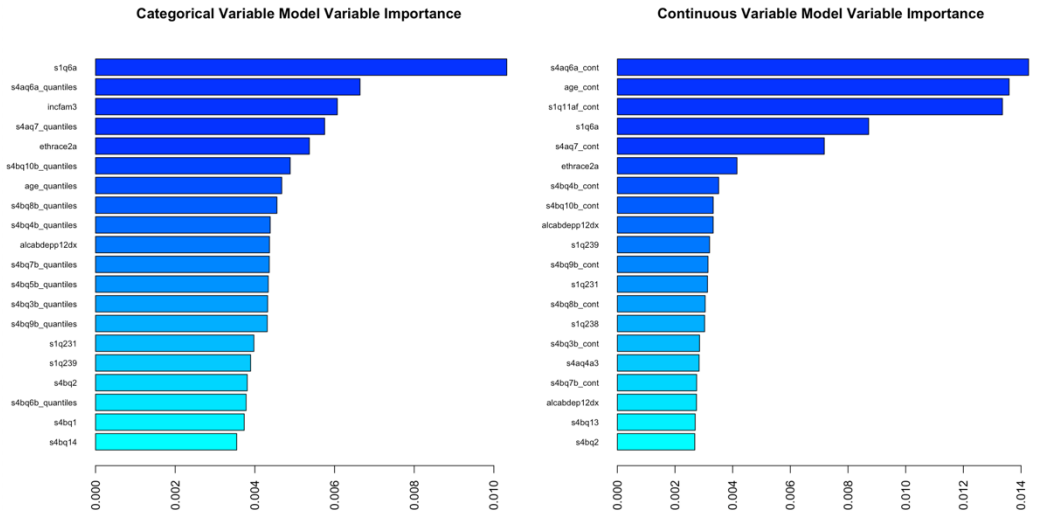

Figure 1 *Variable importance plots*

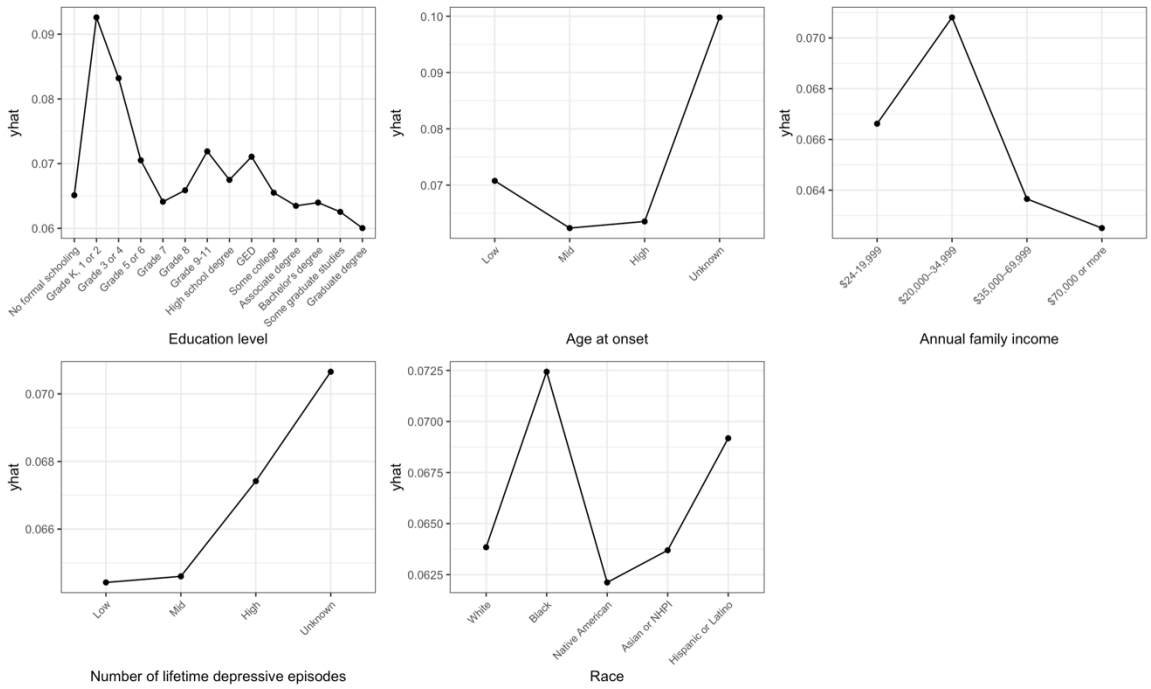

Figure 2 *Partial dependence plots of top 5 important variables in the categorical variable model*

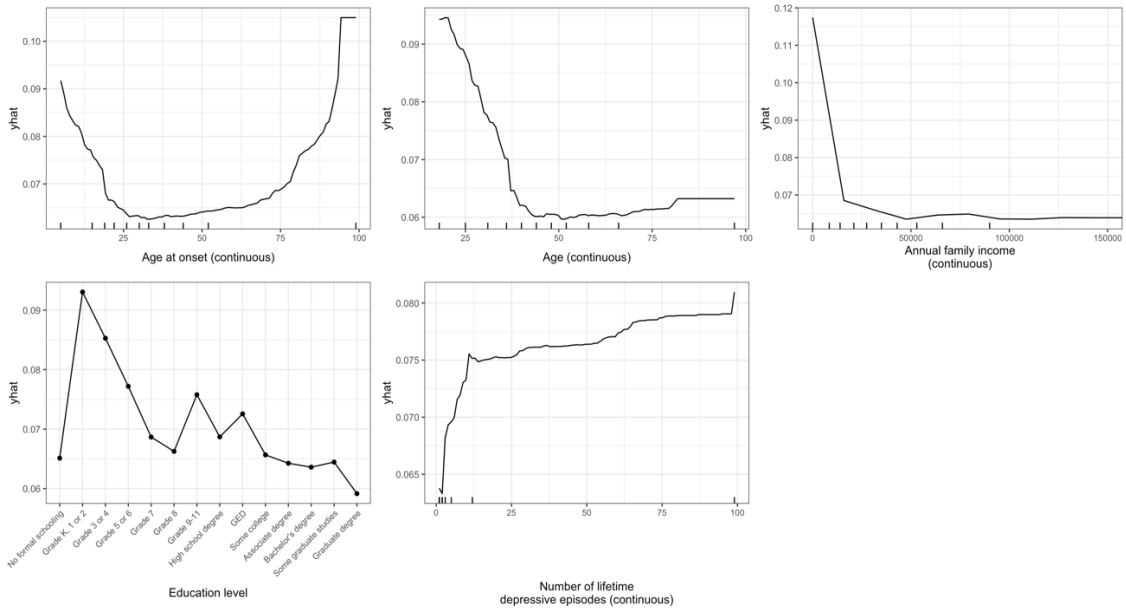

Figure 3 *Partial dependence plots of top 5 important variables in the continuous variable model*

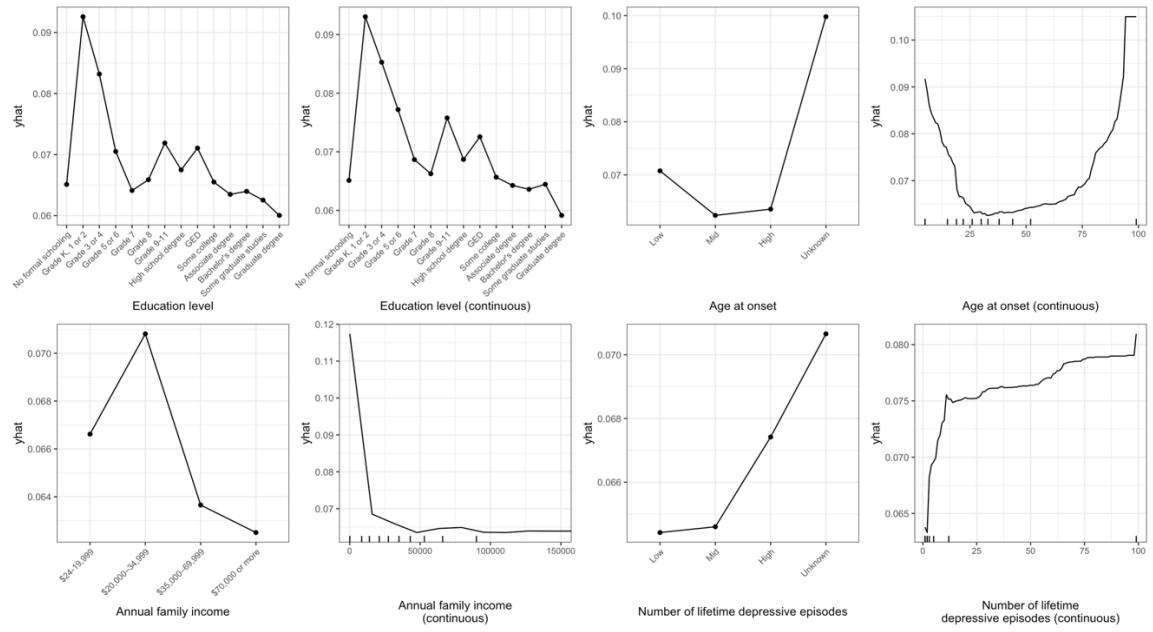

Figure 4 *Partial dependence plots of categorical variable model versus continuous variable model*

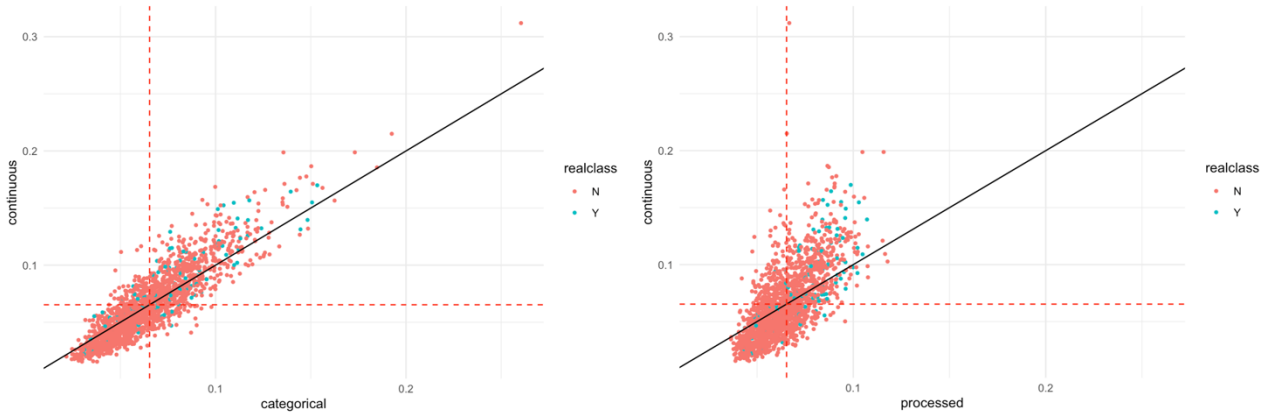

Figure 5 *Scatter plots of the models' predicted probabilities against each other*