This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Lara Andrea Neira Gonzalez

THE UNIVERSITY of EDINBURGH

Doctor of Philosophy
The University of Edinburgh
2017
Declaration

I declare that this Thesis has been composed by me, that the work done in the present Thesis is my own, and that the work described in this Thesis has not been submitted for any other degree or professional qualification.

Lara Andrea Neira Gonzalez
Table of Contents

Declaration

Figures and Tables

Acknowledgements .............................................................................. i

Abstract .................................................................................................. iii

Lay summary.............................................................................................. v

Abbreviations .......................................................................................... vi

1 Introduction ............................................................................................ 1

1.1 Background ......................................................................................... 1

1.1.1 Burden of Psychiatric Disorders ...................................................... 1

1.1.1.1 Social impact ........................................................................... 1

1.1.1.2 Government mental health programs ....................................... 2

1.1.1.3 Costs to medical system ......................................................... 2

1.1.1.4 Research investment ............................................................... 3

1.1.2 Psychosis ....................................................................................... 4

1.1.2.1 Definitions ............................................................................ 4

1.1.2.2 Types of psychotic disorders ................................................. 5

1.1.3 RDoC project ............................................................................... 5

1.1.4 Genetic Epidemiology .................................................................... 6

1.1.4.1 Heritability ............................................................................ 7

1.2 Genetic Epidemiology of Psychosis ................................................... 8

1.2.1 Clinical Features and Epidemiology ............................................. 8

1.2.1.1 Prevalence ............................................................................ 8

1.2.1.2 Diagnosis ............................................................................. 9
1.2.1.3 Age of onset ................................................................. 11
1.2.1.4 Environmental Factors ................................................ 11
1.2.1.5 Dopamine Hypothesis .................................................. 13
1.2.2 Risk Factors Overview .................................................... 14
1.2.3 Family and twin studies .................................................. 16
1.2.4 Linkage Analysis .............................................................. 17
1.2.5 Association Studies .......................................................... 18
  1.2.5.1 Candidate genes in Psychosis ....................................... 19
  1.2.5.2 GWAS in Psychosis .................................................... 20
1.2.6 Polygenic Risk Scores ....................................................... 29
1.2.7 Epistasis ........................................................................... 31
1.3 Issues in Big Data Omics and Machine Learning Overview ........... 33
  1.3.1 Problems of Classical Statistics ....................................... 33
    1.3.1.1 The “Small N, Large P” Problem ............................... 34
    1.3.1.2 Variable Co-dependency .......................................... 36
    1.3.1.3 Interaction Effect Detection ...................................... 36
  1.3.2 Why Machine Learning? .................................................. 37
    1.3.2.1 Dimensionality – Epistasis ...................................... 39
  1.3.3 Kernel and Ensemble Models Review ................................ 39
    1.3.3.1 Classification and Regression Trees ............................ 42
    1.3.3.2 Random Forest ...................................................... 44
    1.3.3.3 Support Vector Machines .......................................... 49
  1.3.4 Use in Genetics .............................................................. 50
1.4 Study Goals ......................................................................... 50
2 Performance of variable importance measures in Random Forest under correlation and application in PGC2 ......................................................... 53
2.1 Introduction ........................................................................................................................................................................53
  2.1.1 Previous studies .................................................................................................................................................................53
  2.1.2 Why the study is needed? .....................................................................................................................................................53
  2.1.3 Aim .........................................................................................................................................................................................55
2.2 Methods ...................................................................................................................................................................................56
  2.2.1 Random Forest .................................................................................................................................................................56
    2.2.1.1 Variable Importance Measures .........................................................................................................................................58
  2.2.2 Data Simulation .....................................................................................................................................................................63
    2.2.2.1 Simulation under $H_A$ .....................................................................................................................................................65
      2.2.2.1.1 Single association ......................................................................................................................................................65
      2.2.2.1.2 Interaction association .............................................................................................................................................65
    2.2.2.2 Simulations under $H_0$ .................................................................................................................................................65
  2.2.3 Power and 5% significance cut-off .......................................................................................................................................66
  2.2.4 Random Forests VIMs simulations ......................................................................................................................................67
2.3 Results .......................................................................................................................................................................................68
  2.3.1 Bias, coverage and correlation of simulated data ..................................................................................................................68
    2.3.1.1 Single effect association ..................................................................................................................................................69
    2.3.1.2 Interaction effect association ...........................................................................................................................................70
  2.3.2 Distributions under $H_0$ .......................................................................................................................................................72
  2.3.3 Power detecting the true signal ..............................................................................................................................................77
    2.3.3.1 Single effects association ................................................................................................................................................77
    2.3.3.2 Interaction effects association ........................................................................................................................................78
  2.3.4 Distributions of RF VIMs under $H_A$ ....................................................................................................................................85
    2.3.4.1 Single association study ................................................................................................................................................85
    2.3.4.2 Interaction association study ........................................................................................................................................88
2.3.5 Conditional PVIM with different correlation cut-off .............................. 99

2.4 Discussion ....................................................................................................... 107

2.5 Application: No significant epistasis in a 29 biomarkers pathway in PGC2 ..
.......................................................................................................................... 112

2.5.1 Data Extraction .......................................................................................... 112

2.5.2 Pathway ........................................................................................................ 114

2.5.3 Population Stratification ............................................................................ 116

2.5.4 Leave-One-Out Cross-Validation Across 39 Studies .................. 116

2.5.5 Random Forest .......................................................................................... 117

2.5.6 Likelihood Ratio Tests (LRTs) between nested models ................. 118

2.5.7 Results ......................................................................................................... 120

2.5.8 Discussion ..................................................................................................... 121

3 Bias Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors ........................................... 124

3.1 Introduction ...................................................................................................... 124

3.1.1 Aims ........................................................................................................... 125

3.2 Methods ........................................................................................................... 126

3.2.1 Data based on normal distributed variables with different variances ...

3.2.1.1 Data simulation under H_o ................................................................... 126

3.2.1.1.1 All variables follow a standard normal distribution ................ 126

3.2.1.1.1 Continuous outcome ...................................................................... 126

3.2.1.1.2 Binary outcome ............................................................................ 127

3.2.1.2 All variables follow a normal distribution with different variances ........................................................................................................ 127

3.2.1.2.1 Continuous outcome .................................................................. 127

3.2.1.2.2 Binary outcome .......................................................................... 127
3.2.1.2 Data simulation under $H_A$ ................................................................. 127

3.2.1.2.1 All variables follow a standard normal distribution ............... 128

3.2.1.2.1.1 Continuous outcome ................................................................. 128

3.2.1.2.1.2 Binary outcome ........................................................................ 128

3.2.1.2.2 All variables follow a normal distribution with different variances ................................................................. 129

3.2.1.2.2.1 Continuous outcome ................................................................. 129

3.2.1.2.2.2 Binary outcome ........................................................................ 129

3.2.2 Data based on normal distributed predictors with different cut-points . 129

3.2.2.1 Data simulation under $H_0$ ................................................................. 130

3.2.2.1.1 Continuous outcome ................................................................. 130

3.2.2.1.2 Binary outcome ........................................................................ 130

3.2.2.2 Data simulation under $H_A$ ................................................................. 130

3.2.2.2.1 All variables follow a standard normal distribution with different cut-points ................................................................. 131

3.2.2.2.1.1 Continuous outcome ................................................................. 131

3.2.2.2.1.2 Binary outcome ........................................................................ 131

3.2.2.2.2 All variables follow a normal distribution with different variances ................................................................. 129

3.2.2.2.2.1 Continuous outcome ................................................................. 129

3.2.2.2.2.2 Binary outcome ........................................................................ 129

3.2.3 Data based on normal distributed errors with different variance........ 131

3.2.3.1 Data simulation under $H_0$ ................................................................. 131

3.2.3.1.1 All variables follow a standard normal distribution ............... 132

3.2.3.1.1.1 Continuous outcome ................................................................. 132

3.2.3.1.1.2 Binary outcome ........................................................................ 132

3.2.3.2 Data simulation under $H_A$ ................................................................. 132

3.2.3.2.1 All variables follow a standard normal distribution ............... 133

3.2.3.2.1.1 Continuous outcome ................................................................. 133

3.2.3.2.1.2 Binary outcome ........................................................................ 133
3.2.4 Random Forest based on VIM_{Gini} simulation. ........................................... 136

3.3 Results .................................................................................................................. 137

3.3.1 Bias, coverage and p-values............................................................................. 137

3.3.1.1 Continuous outcome.................................................................................... 137

3.3.1.2 Binary outcome .......................................................................................... 141

3.3.2 VIM_{Gini} for normal distributed variables with and without the same variance
.......................................................................................................................................... 145

3.3.2.1 Continuous outcome.................................................................................... 145

3.3.2.1.1 Under the null hypothesis ....................................................................... 145

3.3.2.1.2 Under the alternative hypothesis........................................................... 146

3.3.2.2 Binary outcome ............................................................................................ 150

3.3.2.2.1 Under the null hypothesis ....................................................................... 150

3.3.2.2.2 Under the alternative hypothesis........................................................... 151

3.3.3 VIM_{Gini} normal distributed predictors with the same variance but with rounded to a different number of decimal places
.......................................................................................................................................... 154

3.3.3.1 Continuous outcome.................................................................................... 155

3.3.3.1.1 Under the null hypothesis ....................................................................... 155

3.3.3.1.2 Under the alternative hypothesis........................................................... 157

3.3.3.2 Binary outcome ............................................................................................ 160

3.3.3.2.1 Under the null hypothesis ....................................................................... 160

3.3.3.2.2 Under the alternative hypothesis........................................................... 162

3.3.4 VIM_{Gini} for error with different variances ..................................................... 164

3.3.4.1 Continuous outcome.................................................................................... 164

3.3.4.1.1 Under the null hypothesis ....................................................................... 164

3.3.4.1.2 Under the alternative hypothesis........................................................... 166

3.3.4.2 Binary outcome ............................................................................................ 168
3.3.4.2.1 Under the null hypothesis ................................................................. 168
3.3.4.2.2 Under the alternative hypothesis ..................................................... 169
3.4 Discussion ................................................................................................. 172
4 Detecting significantly associated interactions with schizophrenia and cognition in abnormal behaviour and pathways from the Mouse Genotype Informatics (MGI) database ........................................................................................................ 175
4.1 Introduction .............................................................................................. 175
4.2 Methods .................................................................................................... 177
4.2.1 Data and analysis .................................................................................. 177
4.2.2 Random Forest ..................................................................................... 178
4.2.3 Likelihood Ratio Tests (LRTs) between nested models ....................... 179
4.3 Results ...................................................................................................... 183
4.3.1 MGI: Aggression-related behaviour phenotype pathway ..................... 183
4.3.2 MGI: Depression-related behaviour phenotype pathway .................... 186
4.3.3 MGI: Fear/anxiety-related behaviour phenotype pathway .................. 187
4.3.4 MGI: Response to novel object phenotype pathway ............................ 191
4.4 Discussion ............................................................................................... 191
5 Conclusions and future directions ............................................................... 197
5.1 Summary of thesis ................................................................................... 197
5.1.1 Aim 1 .................................................................................................... 197
5.1.2 Aim 2 .................................................................................................... 199
5.1.3 Aim 3 .................................................................................................... 200
5.2 Strengths of the study ............................................................................. 201
5.3 Limitations of the study .......................................................................... 202
5.4 Future directions ..................................................................................... 205
Figures and Tables

Figure 1.1. Illustration of the process of RF algorithm. The exemplificative SNPs are annotated as in dbSNP database, reference SNP ID number. In the third step the split criteria is based on whether the individual carries the risk allele at the SNP chosen in that node. Note that mtry is resampled at each node. ........................... 43

Figure 2.1. Illustration of the process of RF algorithm. The exemplificative SNPs are annotated as in dbSNP database, reference SNP ID number. In the third step the split criteria is based on whether the individual carries the risk allele at the SNP chosen in that node. Note that mtry is resampled at each node. ........................... 57

Figure 2.2. Illustration of RF based in minimal depth. As an example, it shows the minimal depth of V2, V4 and V10............................................................... 62

Figure 2.3. Generation of the single association study as an example of the data simulation........................................................................................................... 64

Figure 2.4. RF VIMs, minimal depth, VIMAUC and VIMparty under H0 for V2, for two variable correlated V3 and V6, and for two independent variables V42 and V90 when r = 0.10 and N = 5................................................................. 73

Figure 2.5. RF VIMs, minimal depth, VIMAUC and VIMparty under H0 for V2, for two variable correlated V3 and V6, and for two independent variables V42 and V90 when r = 0.40 and N = 20................................................................. 74

Figure 2.6. RF VIMs, minimal depth, VIMAUC and VIMparty under H0 for V2, for two variable correlated V3 and V6, and for two independent variables V42 and V90 when r = 0.80 and N = 40................................................................. 75
Figure 2.7. RF VIMs under HA for V2, for two variable correlated V3 and V6, and for two independent variables V42 and V90 when $r = 0.80$ and $N = 40$ in the strongly-association single study. ................................................................. 86

Figure 2.8. RF VIMs under HA for V2, for two variable correlated V3 and V6, and for two independent variables V42 and V90 when $r = 0.10$ and $N = 5$ in the weakly-association single study. ................................................................. 89

Figure 2.9. RF VIMs under HA for V2, for two variable correlated V3 and V6, and for two independent variables V42 and V90 when $r = 0.40$ and $N = 20$ in the weakly-association single study. ................................................................. 90

Figure 2.10. RF VIMs under HA for V2, for two variable correlated V3 and V6, and for two independent variables V42 and V90 when $r = 0.80$ and $N = 40$ in the weakly-association single study. ................................................................. 91

Figure 2.11. RF VIMs under HA in the strongly-associated interaction study for V2 and V90 (interacting variables), for two variable correlated V3 and V6, and for one independent variable V42, when $r = 0.80$ and $N = 40$ ................................................................. 93

Figure 2.12. RF VIMs under HA in the weakly-associated interaction study for V2 and V90 (interacting variables), for two variable correlated V3 and V6, and for one independent variable V42, when $r = 0.10$ and $N = 5$ ................................................................. 96

Figure 2.13. RF VIMs under HA in the weakly-associated interaction study for V2 and V90 (interacting variables), for two variable correlated V3 and V6, and for one independent variable V42, when $r = 0.40$ and $N = 20$ ................................................................. 97

Figure 2.14. RF VIMs under HA in the weakly-associated interaction study for V2 and V90 (interacting variables), for two variable correlated V3 and V6, and for one independent variable V42 and V90 when $r = 0.80$ and $N = 40$. ................................................................. 98
Figure 2.15. VIMrawperm-CF under HA in the weakly-associated single study for $V_2$, for two variable correlated $V_3$ and $V_6$, and for one independent variable $V_{42}$ and $V_{90}$ when $r = 0.40$ and $N = 20$. .......................................................... 101

Figure 2.16. VIMrawperm-CF under HA in the weakly-associated single study for $V_2$, for two variable correlated $V_3$ and $V_6$, and for one independent variable $V_{42}$ and $V_{90}$ when $r = 0.80$ and $N = 40$. .......................................................... 102

Figure 2.17. VIMrawperm-CF under HA in the weakly-associated interaction study for $V_2$, for two variable correlated $V_3$ and $V_6$, and for one independent variable $V_{42}$ and $V_{90}$ when $r = 0.40$ and $N = 20$. .......................................................... 105

Figure 2.18. VIMrawperm-CF under HA in the weakly-associated interaction study for $V_2$, for two variable correlated $V_3$ and $V_6$, and for one independent variable $V_{42}$ and $V_{90}$ when $r = 0.80$ and $N = 40$. .......................................................... 106

Figure 2.19. Example of leave-one-out cross validation in a study design with 5 cohorts, instead with 39. .......................................................... 117

Figure 2.20. Illustration of the approach taken on the applied study in section 2.5. This illustration shows the methods taken in the study and the studies which were combined to report the final results (only the effects which were tested in all independent datasets were combined). ........................................ 116

Figure 3.1. Illustration of the data generation under HA in all different conditions when the outcome is continuous. The top one corresponds to when all predictors and the error follow a standard normal distribution. The three bottom conditions are, from left to right, when predictors follow a standard normal with different variances, when the predictors have different number of decimal places, and when the error variance is lower. The top one is going to be compared to each of the other three conditions in the approach in order to make conclusions... 131
Figure 3.2. VIM_{Gini} under H_0. The top plot illustrates the VIM when all predictors follow a standard normal distribution. The bottom plot shows the VIM when all predictors follow a normal distribution, but each one with a different variance. Continuous outcome. ................................................................. 143

Figure 3.3. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single association models, depending on which variable is associated, when all predictors follow a standard normal distribution. Continuous outcome. Each number i of the X axis corresponds to the subscript of the variable X_i. .......... 144

Figure 3.4. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a normal distribution, but each one with different variance. Continuous outcome. Each number i of the X axis corresponds to the subscript of the variable X_i. .................. 146

Figure 3.5. VIM_{Gini} under H_0. The top plot illustrates the VIM when all predictors follow a standard normal distribution. The bottom plot shows the VIM when all predictors follow a normal distribution, but each one with a different variance. Binary outcome. ................................................................. 147

Figure 3.6. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a standard normal distribution. Binary outcome. Each number i of the X axis corresponds to the subscript of the variable X_i. .................. 149

Figure 3.7. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a standard normal distribution, but with different variances (\(\Sigma=\text{diag}(50,45,40,35,30,25,20,15,10,1)\), \(\Sigma\) is the variance matrix of the predictors). Binary outcome. Each number i of the X axis corresponds to the subscript of the variable X_i. ................................................................. 150
Figure 3.8. VIM_{Gini} under H_0. The top plot illustrates the VIM when all predictors follow a standard normal distribution. The bottom plot shows the VIM when all predictors follow a normal distribution, but each one with different number of decimal places. X1 has one decimal place, X2 has two decimal places, …, X10 has ten decimal places. Continuous outcome................................................................. 153

Figure 3.9. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a standard normal distribution, but with different precision. Continuous outcome. Each number i of the X axis corresponds to the subscript of the variable X_i. . 156

Figure 3.10. VIM_{Gini} under H_0. The top plot illustrates the VIM when all predictors follow a standard normal distribution. The bottom plot shows the VIM when all predictors follow a normal distribution, but each one with different number of decimal places. X1 has one decimal place, X2 has two decimal places, …, X10 has ten decimal places. Binary outcome. ................................................................. 158

Figure 3.11. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a standard normal distribution, but each one has different number of decimal places. Binary outcome. Each number i of the X axis corresponds to the subscript of the variable X_i. ................................................................. 160

Figure 3.12. VIM_{Gini} under H_0. The top plot illustrates the VIM when the error follows a standard normal distribution. The bottom plot shows the VIM when error has a variance of 0.25. Continuous outcome................................................................. 162

Figure 3.13. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a standard normal distribution, but the error \sim N(0,0.5). Continuous outcome. Each number i of the X axis corresponds to the subscript of the variable X_i. .......... 164
Figure 3.1. VIM$_{\text{Gini}}$ under $H_0$. The top plot illustrates the VIM when the error followed a standard normal distribution. The bottom plot shows the VIM when error $\sim N(0,0.5)$. Binary outcome.

Figure 3.15. VIM$_{\text{Gini}}$ under $H_A$. The figure illustrates VIM$_{\text{Gini}}$ in the ten different single models, depending on which variable is associated, when all predictors followed a standard normal distribution, but the error $\sim N(0,0.5)$. Binary outcome. Each number $i$ of the X axis corresponds to the subscript of the variable $X_i$.

Figure 4.1. Illustration of the study design and methods.

Table 1.1. Summary of genome-wide significant findings for schizophrenia, bipolar disorder and MDD identified by GWAS (2008-2016).

Table 2.1. The nine different correlation conditions with the 3 different strengths of correlation ($r$) and 3 different number of correlated variables ($N$).

Table 2.2. Bias and coverage of V2 (associated variable) in the strongly associated study (SAC).

Table 2.3. Bias and coverage of V2 (associated variable) in the weakly associated study (WAC).

Table 2.4. Number of p-values less than 0.0001. WAC mean weakly associated continuous studies.

Table 2.5. Bias and coverage for the interactions on the strongly-associated interaction model.
Table 2.6. Bias and coverage for the interactions on the weakly-associated interaction model................................................................. 71

Table 2.7. Number of p-values less than p-value threshold $1 \times 10^{-5}$ on the weakly-associated interaction model study................................................................. 72

Table 2.8. Power of detecting $V_2$ in the single strongly-associated study (SAC), VIMs, mtry=39 and mtry=27 Mindepth................................................................. 78

Table 2.9. Power of detecting $V_2$ in the single weakly-associated study (WAC), VIMs, mtry=39 and mtry=27 Mindepth................................................................. 80

Table 2.10. Power of detecting $V_2$ in the strong interaction study (SAC), VIMs, mtry=39 and mtry=27 Mindepth................................................................. 81

Table 2.11. Power of detecting $V_{90}$ in the strong interaction study (SAC), VIMs, mtry=39 and mtry=27 Mindepth................................................................. 81

Table 2.12. Power of detecting $V_2$ in the weak interaction study (WAC), VIMs, mtry=39 and mtry=27 Mindepth................................................................. 82

Table 2.13. Power of detecting $V_{90}$ in the weak interaction study (WAC), VIMs, mtry=39 and mtry=27 Mindepth................................................................. 84

Table 2.14. Power of VIMrawperm-CF in detecting $V_2$ under all correlation conditions with the three different cut-offs in the weak single association study. .......... 100

Table 2.15. Power of VIMrawperm-CF in detecting $V_2$ under all correlation conditions with the three different cut-offs in the strong single association study .......... 100
Table 2.16. Power of VIMrawperm-CF in detecting $V_2$ under all correlation conditions with the three different cut-offs in the weakly-associated interaction study. .. 103

Table 2.17. Power of VIMrawperm-CF in detecting $V_{90}$ under all correlation conditions with the three different cut-offs in the weak interaction association study. .. 103

Table 2.18. Power of VIMrawperm-CF in detecting $V_2$ under all correlation conditions with the three different cut-offs in the strong interaction association study. ... 104

Table 2.19. Power of VIMrawperm-CF in detecting $V_{90}$ under all correlation conditions with the three different cut-offs in the strong interaction association study. ................................................................. 104

Table 2.20. Sample size for all 39 cohorts and the number of cases and controls... 114

Table 2.21. Molecular function of the 29 biomarkers and the Genes selected for the study ..................................................................................................................... 115

Table 2.22. Results for the most independent SNPs. The combined p-value is across all 39 independent datasets, taking into account the effect direction and the sample size. The range of the variance explained in percentage ($\%R^2$) is also across all 39 independent datasets ......................................................... 121

Table 3.1. Summary of the approach taken under $H_0$ and $H_A$ in the four different conditions. The case when all variables follow a standard distribution with same variance and same precision and when the error follow a standard normal is used as the reference to compared to the other three different conditions................................................................. 132
Table 3.2. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under $\text{H}_0$, when all predictors and the error followed a standard normal distribution. Continuous outcome. .............................................. 138

Table 3.3. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under $\text{H}_0$, when all predictors followed a normal distribution but with different amounts of variance. The error followed a standard normal distribution. Continuous outcome.............................................................. 138

Table 3.4. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under $\text{H}_0$, when all predictors followed a standard normal distribution but with different number of decimal places. The error followed a standard normal distribution. Continuous outcome. ................................. 139

Table 3.5. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under $\text{H}_0$, when all predictors followed a standard normal distribution. The error followed a normal distribution with 0.5 standard deviation. Continuous outcome................................................................. 139

Table 3.6. Bias and coverage under $\text{H}_A$, when all predictors and the error followed a standard normal distribution. Continuous outcome. All models were significant before and after correction. ................................................................. 140

Table 3.7. Bias and coverage under $\text{H}_A$, when all predictors followed a normal distribution, each one with different variance. The error followed a standard normal distribution. Continuous outcome. All models were significant before and after correction. ................................................................. 140

Table 3.8. Bias and coverage under $\text{H}_A$, when all predictors followed a standard normal distribution, each one rounded with different number of decimal places. The error
followed a standard normal distribution. Continuous outcome. All models were significant before and after correction. ................................................................. 140

Table 3.9. Bias and percentage of coverage under $H_A$, when all predictors followed a standard normal distribution. The error followed a normal distribution with standard deviation of 0.5. Continuous outcome. All models were significant before and after correction. ................................................................. 140

Table 3.10. Bias, coverage, number of $p$-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under $H_0$, when predictors and the error followed a standard normal distribution. Binary outcome. ......................................................... 142

Table 3.11. Bias, coverage, number of $p$-values less than 0.05 and less than Bonferroni correction under $H_0$, when all predictors are normally distributed with different variances. The error is standard normal distributed. Binary outcome. .......... 142

Table 3.12. Bias, coverage, number of $p$-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under $H_0$, when all predictors follow a standard normal distribution, each one rounded with different number of decimal places. The error is standard normal distribution. Binary outcome. ......................... 142

Table 3.13. Bias, coverage, number of $p$-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under $H_0$. The error followed a normal distribution with 0.5 standard deviation. Binary outcome. ......................................................... 143

Table 3.14. Bias and coverage under $H_A$, when all predictors and the error followed a standard normal distribution. Binary outcome. All models were significant before and after correction. ........................................................................................................ 143

Table 3.15. Bias and coverage under $H_A$, when all predictors followed a normal distribution, each one with different variance. The error followed a standard
normal distribution. Binary outcome. All models were significant before and after correction.

Table 3.16. Bias and coverage under $H_A$, when all predictors followed a standard normal distribution, each one rounded different number of decimal places. The error followed a standard normal distribution. Binary outcome. All models were significant before and after correction.

Table 3.17. Bias and percentage of coverage under $H_A$, when all predictors followed a standard normal distribution. The error followed a normal distribution with a standard deviation of 0.5. Binary outcome. All models were significant before and after correction.

Table 3.18. Median of the number of unique values of the variable $X_i$ under $H_0$. Each variable $X_i$ has $i$ number of decimal places. Continuous outcome.

Table 3.19. Median of the number of unique values of the variable $X_i$ under $H_0$. All predictors follow a standard normal distribution with the same number of decimal places. Continuous outcome.

Table 3.20. Median of the number of unique values of the variable $X_i$ under $H_A$. All predictors follow a standard normal distribution with the same number of decimal places. Continuous outcome.

Table 3.21. Median of the number of unique values of the variable $X_i$ under $H_A$. Each variable $X_i$ has $i$ number of decimal places. Continuous outcome.

Table 3.22. Median of the number of unique values of the variable $X_i$ under $H_0$. All predictors follow a standard normal distribution with the same number of decimal places. Binary outcome.
Table 3.23. Median of the number of unique values of the variable $X_i$ under $H_0$. Each variable $X_i$ has $i$ number of decimal places. Binary outcome. ........................................ 161

Table 3.24. Median of the number of unique values of the variable $X_i$ under $H_A$. All predictors follow a standard normal distribution with the same number of decimal places. Binary outcome. .................................................. 162

Table 3.25. Median of the number of unique values of the variable $X_i$ under $H_A$. Each variable $X_i$ has $i$ number of decimal places. Binary outcome. ......................... 162

Table 3.26. Summary of $VIM_{Gini}$ behaviour on the three different studies compared to when all variables and error followed a standard normal distribution under $H_0$. ................................................................. 168

Table 3.27. Summary of $VIM_{Gini}$ behaviour on the three different studies compared to when all variables and error followed a standard normal distribution under $H_A$. ................................................................. 168

Table 4.1. Number of genes and SNPs by pathway. ........................................ 177

Table 4.2. 2-way interactions with $p$-value $< 0.05$ in psychosis and IQ in aggression pathway. ........................................................................................................ 184

Table 4.3. 2-way interaction with $p$-value $< 0.05$ in psychosis and verbal IQ in aggression pathway. ........................................................................................................ 184

Table 4.4. 3-way interaction with $p$-value $< 0.05$ in psychosis, IQ and verbal IQ in aggression pathway. ........................................................................................................ 185

Table 4.5. 2-way interaction with $p$-value $< 0.05$ in psychosis and IQ in depression pathway. ........................................................................................................ 186
Table 4.6. 3-way interaction with p-value < 0.05 in psychosis and IQ in depression pathway. ....................................................................... 186

Table 4.7. 2-way interactions with p-value < 0.05 in psychosis, IQ and verbal IQ in fear/anxiety pathway. .................................................................................................................. 188

Table 4.8. 3-way interaction with p-value < 0.05 in psychosis and IQ in fear/anxiety pathway. .................................................................................................................. 189

Table 4.9. 3-way interaction with p-value < 0.05 in psychosis and verbal IQ in fear/anxiety pathway. .................................................................................................................. 189

Table 4.10. 3-way interaction with p-value < 0.05 in psychosis, IQ verbal IQ in fear/anxiety pathway. .................................................................................................................. 190
Acknowledgements

First and foremost I would like to dedicate this Thesis to my family. To my parents, Jose Neira and Maria del Mar Gonzalez, to my grandmothers, Perpetua and Maruja, to my partner and fiancé, Israel Camarena, to my brother and his wife, Jesus Neira and Isabel Gasalla, to my niece, Carlota, to my aunt and my uncle, Loly Neira and Javier Garcia, and to my godson and cousins, Xoel, Naila and Paola Garcia. This thesis could not have finished without their support.

I need to thank the many people who helped me during my Ph.D. To my first supervisor, Kristin Nicodemus, for helping me with the thesis, sharing her experience, expertise and knowledge, for the good moments, and for having such an amazing group, who are my friends, Ksenia, Elvina, Joeri and Alex. Thanks you all. I would also like to thank my second and third supervisors, Kathy Evans and David Porteous for helping me to finish my Thesis project and spend their time meeting with me. In addition, to Cathy Abbott, who helped me to finish my Thesis.

I would like to thank my flatmates in Edinburgh and in Dublin for being such an amazing family and for always helping me, to Annalaura and Marisol (my sisters), to Gary (my brother), and to Tom (my brother in law). Also, to my best friends in Edinburgh, Victor, Erola and Nefeli, you were always there for me. Thanks to “Las primas”, Jose, Maria Isabel, Aleix, Carles, Heidi, Oscar, Marc, Christina and Hans for all the fun, good moments, and made me forget the PhD at times. To my ballroom society friends, Maciej, Angus, Emilia, Nick, Foteini, Steven… you all know who you are. I would also like to thank to Jose Alberto and Tania, you were there whenever I needed it. To Sara, Lulu, Cecy, Andrea, Aline, Diana, Memo, Toty, Edgar, Mimy, Jose, Annika, Carlos, all O’Neills and TCHPC people for all good moments in Dublin.

Thanks to my group of friends “Las Pavitas” for growing up with me and making my life wonderful. In special to my sister Debora, who lived with me for several years and supported me to finish my career; Nahir, who supported me in everything and helped me to achieve my goals and for being there listening to me in the good and bad moments; and Sandrita, for being an amazing friend and for coming to visit me.
wherever I lived, always giving brightness to my life. Thanks to Maria for also visiting me to both cities and sorry for have not being able to be with you as much as I would have liked it because I had to finish a project to submit an abstract. Lastly, to Cristina Núñez, Nelita and Alfonso for all your support.

I am so grateful to all the amazing people in my life who were always there, in the good and bad times. I am so lucky for having you all.
Abstract

Psychotic disorders such as schizophrenia and bipolar disorder have a strong genetic component. The aetiology of psychoses is known to be complex, including additive effects from multiple susceptibility genes, interactions between genes, environmental risk factors, and gene by environment interactions. With the development of new technologies such as genome-wide association studies and imputation of ungenotyped variants, the amount of genomic data has increased dramatically leading to the necessary use of Machine Learning techniques. Random Forest has been widely used to study the underlying genetic factors of psychiatric disorders such as epistasis and gene-gene interactions. Several authors have investigated the ability of this algorithm in finding single and interaction effects, but have reported contradictory results. Therefore, in order to examine Random Forest ability of detecting single and interaction effects based on different variable importance measures, I conducted a simulation study assessing whether the algorithm was able to detect single and interaction models under different correlation conditions. The results suggest that the optimal Variable Importance Measures to use in real situations under correlation is the unconditional unscaled permutation variable importance measure. Several studies have shown bias in one of the most popular variable importance measures, the Gini importance. Hence, in a second simulation study I study whether the Gini variable importance is influenced by the variability of predictors, the precision of measuring them, and the variability of the error. Evidence of other biases in this variable importance was found. The results from the first simulation study were used to study whether genes related to 29 molecular biomarkers, which have been associated with schizophrenia, influence risk for schizophrenia in a case-control study of 26476 cases and 31804 controls from 39 different European ancestry cohorts. Single effects from ACAT2 and TNC genes were detected to contribute risk for schizophrenia. ACAT2 is a gene in the chromosome 6 which is related to energy metabolism. Transcriptional differences have been shown in schizophrenia brain tissue studies. TNC is expressed in the brain where is involved in the migration of the neurons and axons. In addition, we also used the simulation results to examine whether interactions between genes associated with abnormal emotion/affect behaviour influence risk for psychosis and
cognition in humans, in a case-control study of 2049 cases and 1794 controls. Before correcting for multiple testing, significant interactions between CRHR1 and ESR1, and between MAPT and ESR1, and among CRHR1, ESR1 and TOM1L2, and among MAPT, ESR1 and TOM1L2 were observed in abnormal fear/anxiety-related behaviour pathway. There was no evidence for epistasis after Bonferroni correction.
Lay Summary

Psychotic disorders such as schizophrenia and bipolar disorder are highly inheritable. But it is difficult to know which genetic components are related to these illnesses as each single component gives a low contribution. Therefore, adding the effects from different “mutations” or genes as well as the interaction between them may better explain these disorders. Machine Learning techniques, which are novel mathematical algorithms that belong to the field of artificial intelligence, are adequate tools which serve to investigate these genetic components. In two of the chapters of my thesis, I studied the Machine Learning technique Random Forest and its ability to detect the interaction between variables. This was performed through a study which simulated real situations. In addition, I applied two real studies. In the first, I researched if interactions between genes have an important impact on schizophrenia. In the other, I tested whether interactions are importantly involved with psychotic disorders. For this investigation, I considered genes that were previously proven to be related with abnormal emotions and effect behaviour in mice. The result was that no important interactions were found. However, in the first applied study, important contributions from single genes were obtained.
Abbreviations

A2M: Alpha-2 Macroglobulin

ACAT2: Acetyl-coenzyme A acetyltransferase 2

ADHD: Attention Deficit Disorder Association

AKT1: Serine/Threonine Kinase 1

ANK3: Encoding Ankyrin 3

APA: American Psychiatric Association

AUC: Area Under the Curve

BDNF: Brain Derived Neurotrophic Factor

BP: Bipolar Disorder

CACNA1C: Encoding Calcium Channel, Voltage-dependent, L type, α1C Subunit

CART: Classification and Regression Tree

CBS: Cystathionine-Beta-Synthase

CCDC68: Coiled-coil Domain Containing 68

CIF: Conditional Inference Forest

CIT: Citron rho-Interacting Serine/Threonine Kinase

CNNM2: CBS Domain Divalent Metal Cation Transport Mediator 2

COMT: Catechol-O-Methyltransferase

CRHR1: Corticotropin Releasing Hormone Receptor 1

CSMD1: CUB and Sushi Multiple Domains 1

DAOA: D-amino Acid Oxidase Inhibitor
DGKH : Diacylglycerol Kinase Eta

DISC1 : Disrupted in Schizophrenia 1

DRD2 : Dopamine Receptor D2

DSM : Diagnostic and Statistical Manual of Mental Disorders

DTNBP1 : Dystrobrevin Binding Protein 1

ELAVL2 : ELAV Like RNA Binding protein 2

ERBB2 : Erb-b2 receptor tyrosine kinase 2

ERBB4 : Erb-b2 Receptor Tyrosine Kinase 4

ESR1 : Estrogen Receptor 1

FDR : False Discovery Rate

FEZ1 : Fasciculation And Elongation Protein Zeta 1

GLMMs : Generalized Linear Mixed Models Framework Mainly

GRM3 : Glutamate Metabotropic Receptor 3

GWAS : Genome-Wide Association Study

H₀ : Null hypothesis

Hₐ : Alternative hypothesis

HIST1H2BJ : Histone Cluster 1 H2B Family Member J

HWE : Hardy-Weinberg Equilibrium

IQ : Intelligence Quotient

ISC : International Schizophrenia Consortium

KCCA : Kernel Canonical Correlation Analysis

KEGG : Kyoto Encyclopedia of Genes and Genomes
LD : Linkage Disequilibrium

LHPP : Phospholysine Phosphohistidine Inorganic Pyrophosphate Phosphatase

LOD-score: Logarithm of Odds score

LR : Logistic Regression

LRT : Likelihood Ratio Test

MAD1L1 : Mitotic Arrest Deficient like 1

MAF : Minor allele Frequency

MAPT : Microtubule Associated Protein Tau

MCLR : Monte Carlo Logic Regression

MDA : Mean Decrease Accuracy

MDD : Major Depressive Disorder

MDG : Mean Decrease Gini

MDR : Multifactor Dimensionality Reduction

MGI : Mouse Genotype Informatics

MHC : Major Histocompatibility Complex

MIR-137 : MicroRNA 137

ML : Machine Learning

MQ : Mental Health Research Charity

MRC : Medical Research Council

MSP : Multiple Span Probability

NCAN : Neurocan Gene

NDE1 : NudE Neurodevelopment Protein 1
NDEL1 : NudE Neurodevelopment Protein 1 Like 1
NIHR : National Institute of Health Research
NLP : Non-Parametric Linkage Models
NMP16 : Encoding Matrix Metallopeptidase 16
NRG1 : Neuregulin
NRGN : Neurogranin
NT5C2 : 5'-Nucleotidase, Cytosolic II
ODZ4 : Protein Odd oz/ten-m Homolog 4
OOB : Out-Of-Bag
PAFAH1B1 : Platelet Activating Factor Acetylhydrolase 1b Regulatory Subunit 1
PCGEM1 : Prostate-Specific Transcript 1
PGBD1 : Pterin-4 Alpha-Carbinolamine Dehydratase 1
PGC : Psychiatric Genetics Consortium
PGC2 : Psychiatric Genomics Consortium 2
PRS : Polygenic Risk Score
PC : Principal component
PS: Population stratification
PVIMs : Permutation Variable Importance Measures
RDoC : Research Domain Criteria
RF : Random Forest
RFE : Recursive Feature Elimination
ROC : Receiver Operating characteristic
SIRT1 : Sirtuin 1
SMS : Smith-Magenis syndrome
SNP : Single Nucleotide Polymorphisms
STRING : Search Tool for the Retrieval of Interacting Genes/Proteins
SVM : Support Vector Machines
TCF4 : Transcription Factor 4
TNC : Tenascin C
TOM1L2 : Target Of Myb1 Like 2 Membrane Trafficking Protein
TRANK1 : Tetratricopeptide Repeat and Ankyrin Repeat Containing 1
TRIM 26 : Tripartite Motif Containing 26
UTR : Untranslated region
VIM : Variable importance measure
VIM_{Gini} : Gini variable importance measure
VIM_{rawperm-RF} : unconditional unscaled permutation variable importance measure
VIM_{Breiperm-RF} : Breiman scaled permutation variable importance measure
VIM_{Liawperm-RF} : Liaw scaled permutation variable importance measure
VIM_{rawperm-CF} : Conditional permutation variable importance measure
VIM_{AUC} : unconditional permutation variable importance measure based on AUC
VIM_{party} : unconditional unscaled permutation variable importance measure from CIF
VWF : Von Willebrand Factor
WHO : World Health Organization
WTCCC : Welcome Trust Case Control Consortium

WTCCC2 : Welcome Trust Case Control Consortium 2
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

1. Introduction

1.1. Background

1.1.1. Burden of Psychiatric Disorders

Great scientific progress has been made in the field of neuroscience in recent decades; a field which is particularly important as it strives to better our understanding of the functioning of the brain and its effects on millions of different processes. The aim of psychiatric research is to use our knowledge to improve the standards of living and help those with psychiatric conditions, this includes better understanding of aetiology, better diagnoses and better personalized treatment.

It is still incredibly difficult to give each individual a correct diagnosis; this is partially due to subjectivity of diagnosis. Furthermore, many psychiatric conditions have similar and overlapping symptoms as well as the fact that, in order to diagnose each condition, one only needs to have a subset of these symptoms (American Psychiatric Association 2013). This means that people with identical symptoms may be given a different diagnosis by different clinicians and people with the same diagnoses may only have some, or even no, symptoms in common. Another problem is that the economic and social burden suffered by those affected and those close to them, as well as the social stigma that is associated with having a psychiatric illness, can contribute to developing and worsening of the disease (Muntaner et al. 2004).

1.1.1.1. Social impact

Mental illness does not only affect those who suffer from these conditions, but also affects those in their social environment on different levels (Allen et al. 2014). People with mental illness require special care from health professionals and people in their
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

social circle (affecting the lives of their friends and relatives) as well as other financial resources associated with treatment. This produces a burden not only on the affected individuals and those close to them but also on health care and government in general.

1.1.1.2. Government mental health programs

Despite the increased scientific interest and growing contributions to psychiatric research, there is still a long way to go to reduce the social impact and also help patients with economic healthcare implications, and a dramatic rise in drug costs (Markram, 2013). An appropriate allocation of economic resources is required by governments as well as the general public (in the form of private donations) for psychiatric research and awareness in their populations (Gustavsson et al., 2011).

Among the factors that affect both the economy and patients with mental illness, health-related programs have been shown not to have the desired performance due to poor implementation practices, financing and the current state of development in many countries. Murawiec and Krysta (Murawiec and Krysta 2015) point out that in European countries there is a gap between good legislation and poor implementation, some of the health reforms are largely aspirational and severely underfunded for the expected results. They also suggest that, in order to achieve these desired improvements, the government needs to be more involved in policy implementation.

1.1.1.3. Costs to medical system

It has been reported that the costs associated with mental health are the greatest health-related financial and social burden in Europe. The economic costs incurred include direct and indirect treatment costs, welfare spending, and productivity losses (Wykes et al. 2015).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

In the UK the following statistics have been reported by Fineberg et al. (2013). Roughly 45 million cases of brain disorder, including even headache, were recorded in 2010 at a healthcare cost of over €100 billion. The diseases that affect patients the most were headache, anxiety, sleep, somatoform and mood disorders. By medical expenditure, dementia ranked the highest at more than €22 billion per year; followed by psychotic illnesses; mood disorder; and addiction disorders €16,717 million; €19,238 million and €11,719 million respectively; and anxiety being the lowest in this group, with a cost of €11,687 million. However, if we break down the costs incurred per person, dementia, psychotic, mood, anxiety and addiction disorders are amongst the lowest, with less than €3000 spent per patient, with the exception of psychosis with more than €5000 per person. The costs per subject can be divided between around 50% for both indirect and direct costs (50% indirect costs, 25% direct non-medical and 25% direct healthcare costs), whereby direct costs comprise direct non-medical and direct healthcare costs (Fineberg et al. 2013), and indirect cost such as loss of productivity and the time spent by care givers involved in the process. These figures give us an overview of the scale of the problem and the number of individuals affected by psychiatric ill-health in the UK as well as the scope of financial resources allocated for this category of health of UK.

1.1.1.4. Research investment

Despite the financial burden and the social constraints that mental illnesses create, there is not enough investment in research of these psychiatric disorders - which could help find better treatment, aid general understanding and efficient diagnosis of the diseases, and eventually improve the quality of life of patients (Joyce 2014). MQ is a large UK mental health charity founded by Lord Dennis Stevenson and Sir Mark Walport in 2009, which was formed precisely for the purposes described above. MQ is currently allocating around £20 million every year for research funding in many scientific areas that may contribute to either: treatment; diagnosis; support methods; or general understanding of mental ill-health (2016).
A recent report by MQ (2015) provides a general picture of the UK research budget. It indicates that there is a major disparity between investment provided for mental health research, accounting for 5.5% of the UK budget, and funding for cancer, which is almost quadrupled at 19.6%. The approximate yearly research expenditure per patient in the UK is £9.75 for mental disorders. This amount is dwarfed by the £1,571 spent on a patient with cancer (National Cancer Research Institute 2013).

In the UK, charitable funding is pivotal in medical research, accounting for over a third of the total. Over the total funding in medical research, there is also a gross disparity in the charitable funding provided to cancer and mental health research, with 3.1% destined to mental health compared to 30% for cancer. According to the MQ report, the three major charitable contributors of mental health research are the Wellcome Trust, Medical Research Council (MRC) and National Institute of Health Research (NIHR), which provide 33.5%, 26.6% and 25% of the total charitable budget for mental health (of the 100% of mental health) respectively.

1.1.2. Psychosis

1.1.2.1. Definitions

Psychosis is a condition defined by a group of symptoms which may appear regularly or infrequently with a duration that may also vary depending on the type and state of the psychotic disease (Lawrie et al. 2016). These symptoms influence the behaviour and cognition of the affected individual and typically take the form of hallucinations and delusions, and also other problems of thought and emotion (American Psychiatric Association 2013). There are five domains or symptoms of psychosis: i) hallucinations (e.g. hearing or seeing something that is not real); ii) delusions (a belief that it is contradictory to the reality and it is strongly maintained, e.g. they think that someone wants to kill them when actually there is no reason to think so); iii) disorganized thought/speech (their thoughts are not connected and when they speak, they show that...
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

disconnection); iv) disorganized or abnormal motor behaviour (including catatonia, or absence of natural movement); and v) negative symptoms (degradation of normal function, including self-neglect, the inability to feel pleasure).

There is a group of psychiatric illnesses which are characterized of these symptoms called psychotic diseases such as schizophrenia and non-affective psychotic disorders, and affective psychoses which include schizoaffective disorder, bipolar disorder (BP) and major depressive disorder (MDD) with psychosis, with schizophrenia and BP being the most common (Tandon et al. 2012).

1.1.2.2. Types of psychotic disorders

Even though the symptoms described above are key in schizophrenia and other non-affective psychotic disorders, affective psychoses cannot be characterised by these symptoms because they are secondary traits (George 2014). In fact, psychotic diseases can be classified according to the way the symptoms described above feature in the illnesses. Schizophrenia, schizoaffective disorder, schizophreniform disorder and brief psychotic disorder are characterised by hallucinations, delusions and also disorganised speech with a typical age of onset somewhere in late adolescence or young adulthood (Gogtay et al. 2011). Patients with affective psychotic diseases such as BP, who suffer severe mood swings from manic moods to depressive moods, may also have delusions and hallucinations in their extreme states. Although it is not common, they are also present in people suffering from severe MDD with a prevalence of 0.4% (Ohayon and Schatzberg 2002).

1.1.3. RDoC project

Previous studies in the fields of behavior and psychiatric diseases have led to more research that now also takes into account cognitive, memory and executive functions (Zanello et al. 2009; Mancuso et al. 2011). The US National Institute of Mental Health
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

(NIMH), another mental health research organization, is supporting the Research Domain Criteria (RDoC) project (Insel 2012). The underlying genetic contribution to psychiatric disorders is very complex with no single gene having a sufficiently significant effect to explain the heredity of these diseases. Therefore, the RDoC initiative aims to study behavioural phenotypes, many of which may overlap between mental illnesses, as opposed to symptomology alone (Simmons and Quinn 2014). One of the main goals of the RDoC project is to study the positive and negative valence systems, cognition, social processes and arousal and regulatory systems; as well as their relation to genomic, molecular, cellular, circuit, physiological and behavioural factors.

In other words, RDoC is a research framework for new ways of studying mental disorders. It integrates many levels of information (from genomics to self-report) to better understand basic dimensions of functioning underlying the full range of human behavior from normal to abnormal (Simmons and Quinn 2014).

1.1.4. Genetic Epidemiology

A very challenging task is understanding the genetic and molecular architecture of psychiatric disorders, particularly factors which lead to higher prevalence of dysfunction in general. From the genetics point of view, there are mainly two different types of neuropsychiatric disorders: monogenic and complex (multifactorial or oligogenic); but there are also chromosomal abnormalities (changes in chromosome structure or number such as aneuploidy like down syndrome). Monogenic disorders are caused by a single gene and are, therefore, disorders with Mendelian patterns of inheritance. However, their clinical manifestations can be affected by other genes and environmental circumstances (Weatherall 2000). In contrast, oligogenic or complex disorders are more common and it is more difficult to study the underlying factors of the illnesses. Complex disorders such as psychotic disorders develop when several
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

genetic and environmental elements are present and interactions occur between them (Hannan 2013).

1.1.4.1. Heritability

Heritability is a statistic which indicates the amount of genetic variation found in a phenotypic feature distinctive among individuals of a population (Akey et al. 2001). It is important to note that heritability of a disease is characteristic of an entire population and is not a measure of probability of a single individual having that illness.

Roughly speaking, a phenotypic trait can be defined by the sum of genetic and environmental effects as follows: \( P = G + E \), where \( P \) is the phenotype under study, \( G \) and \( E \) refer to the genetic and environmental effects respectively (Tenesa and Haley 2013). \( G \) covers additive and interaction genetic components and genetic dominance; and \( E \) is constituted by the shared environment of the relatives in a family and the environmental effect that does not take into account the relatedness between individuals. Then, the total phenotypic variance can be explained by the sum of the variances of its components, \( S^2_P = S^2_G + S^2_E \).

There are two different types of heritability, the narrow-sense and the broad-sense heritability, denoted as \( h^2 \) and \( H^2 \) respectively. \( h^2 \) accounts for additive genetic difference, measuring the proportion of genes linked to the phenotype carried from the parents. On the other hand, the broad-sense heritability, defined as \( H^2 = \frac{S^2_G}{S^2_P} \), explains the degree in which the phenotype is determined by the individual’s genotype (Visscher et al. 2008).

Finding no heritability for the trait is not a demonstration that genes are irrelevant; rather, it demonstrates that, in the particular population studied, there is no genetic variation at the relevant loci. In other populations or other environments, the trait might be heritable (Griffiths et al. 2000).
1.2. Genetic Epidemiology of Psychosis

1.2.1. Clinical Features and Epidemiology

1.2.1.1. Prevalence

Most epidemiological studies on psychosis focus mainly on schizophrenia and BP since these are the most common disorders that feature psychotic symptoms. The percentage of people who already have, at a certain moment or during a period, a disease is called prevalence. There are different ways to calculate the prevalence.

\[
\text{Point prevalence} = \frac{\text{Number of cases at that point}}{\text{Number of population at risk at that point}} \times 100
\]

\[
\text{Period prevalence} = \frac{\text{Number of cases with the disease at some point over that period}}{\text{Number of population at risk over that period}} \times 100
\]

\[
\text{Lifetime prevalence} = \frac{\text{Number of cases who had the disease over their lifetime}}{\text{Number population at risk (from beginning of time period)}} \times 100
\]

Psychotic illnesses occur 10 times less than psychotic-like symptoms in the general population (van Os et al. 2001; Nuevo et al. 2012). Only a few general population studies have been carried out and Bogren et al. (2009) have estimated the prevalence of all psychotic disorders together. They have shown a 50-year period prevalence (1947–1997) of 4.2% using the Lundby cohort, which is a prospective, longitudinal cohort study on a sample consisting of 3,563 subjects over the period between 1,947
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

and 1997. And a lifetime prevalence in 1997 of 2.8% for any psychotic disorder, it was calculated including individuals at age 40 years or over, although the normal age range of this studies is 18 and 65 years old. As the study was applied considering healthy individuals of an entire population, which does not correspond with the current sociodemographic structure, the fact of including surviving individuals from the original population in the lifetime prevalence of 1997 makes the age corresponds to over 40 years old in the Lundby cohort (Bogren et al. 2009).

Taking into account specific psychotic disorders, the prevalence of schizophrenia has been estimated as 1% (McGrath et al. 2008), whilst the prevalence of BP is approximately 4% and MDD varies between 10% and 15% in the UK population (Ketter 2010); (Smith et al. 2013). MDD with psychosis has a 0.4% prevalence in the general population of several countries in Europe, but the prevalence dramatically increases to 18.5% considering patients with MDD in Europe without psychosis (Rothschild 2013).

1.2.1.2. Diagnosis

The Diagnostic and Statistical Manual of Mental Disorders V (DSM-V) of the American Psychiatric Association (APA) defines the mental disorder classification with the specific symptoms and criteria of each psychiatric disorder for their clinical diagnosis. It is the guide for mental health professionals in many countries in the world including the United States and the United Kingdom (American Psychiatric Association 2013), but mostly in the United States. The current classification of psychotic disorders covered in the chapter Schizophrenia Spectrum and Other Psychotic Disorders of DSM-V (American Psychiatric Association 2013) has undergone only a few changes from the last version of DSM-IV (Widiger and American Psychiatric Association. Task Force on DSM-IV. 1994).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

The diagnoses are based on how many, how long and how the presence of symptoms affect the individual. Schizoaffective disorder is considered as an independent disease, and both schizophrenia and schizoaffective disorder diagnoses have been shown to be appropriate and consistent (independent diseases) with the symptoms criteria in DSM-V (Regier et al. 2013). Bizarre delusions were symptoms assigned before to schizophrenia. Now with DSM-V, patients who have them are likely to be diagnosed with delusional disorder. In addition, although several researchers argued that catatonia should be classified as an independent disease (Fink and Taylor 2008), it is still considered within the domain of psychosis (Heckers et al. 2010). Bizarre delusions were symptoms assigned before to schizophrenia. Now with DSM-V, patients who have them are likely to be diagnosed with delusional disorder. In addition, although several researchers argued that catatonia should be classified as an independent disease (Fink and Taylor 2008), it is still considered within the domain of psychosis (Heckers et al. 2010). Bizarre delusions were symptoms assigned before to schizophrenia. Now with DSM-V, patients who have them are likely to be diagnosed with delusional disorder. In addition, although several researchers argued that catatonia should be classified as an independent disease (Fink and Taylor 2008), it is still considered within the domain of psychosis (Heckers et al. 2010). Bizarre delusions were symptoms assigned before to schizophrenia. Now with DSM-V, patients who have them are likely to be diagnosed with delusional disorder. In addition, although several researchers argued that catatonia should be classified as an independent disease (Fink and Taylor 2008), it is still considered within the domain of psychosis (Heckers et al. 2010). Bizarre delusions were symptoms assigned before to schizophrenia. Now with DSM-V, patients who have them are likely to be diagnosed with delusional disorder. In addition, although several researchers argued that catatonia should be classified as an independent disease (Fink and Taylor 2008), it is still considered within the domain of psychosis (Heckers et al. 2010).

The classification of mood disorders has experienced a more significant change. In DSM-V bipolar disorders are not included in the depressive disorders, they have their own chapter which was set up between psychotic disorders and the depressive disorders. In terms of depressive disorders, DSM-V includes three new diseases: disruptive mood dysregulation disorder, persistent depressive disorder, and premenstrual dysphoric disorder.

DSM-V considers a schizophrenia diagnosis as the presence of two or more symptoms for at least one month, where the patient must present with either hallucinations, delusions or disorganized speech. Negative symptoms and disorganized or catatonic behaviours can also be considered for a diagnosis (American Psychiatric Association 2013).

BP is the second most common disorder featuring psychosis. According to DSM-V (American Psychiatric Association 2013), Bipolar I is defined by the occurrence of a minimum of one high mood episode (manic), whereas Bipolar II is defined as having
both low and high mood stages (depressive and hypomanic episodes respectively) which do not reach a manic episode. The diagnosis of MDD according to DSM-V is described as a patient with five or more of the following symptoms: weight changes, sleep disturbances, abnormal motor function, fatigue, feelings of worthlessness or guilt, cognitive deficits, suicidal ideation, a depressed mood, and anhedonia, where at least one of the latter two must be experienced almost every day for at least two weeks. There can be up to 227 combination of symptoms for the clinical diagnosis of MDD (S.-C. Park et al. 2016); (Zimmerman et al. 2015), some being more prevalent than others (Zimmerman et al. 2015). Psychotic depression is considered a subtype of MDD and essentially it does not have to be considered only as severe illness; patients can be affected by both mood-congruent and mood-incongruent psychotic symptoms (Rothschild 2013).

1.2.1.3. Age of onset

The typical age of onset of psychosis is somewhere in late adolescence or young adulthood (Gogtay et al. 2011). The presentation of psychotic symptoms in non-diagnosed children has been shown to be highly distressing as well as being predictive of suicide and self-harm, and the onset (typically during adolescence) of schizophrenia, and other disorders, such as BP and MDD (Armando et al. 2010); (Polanczyk et al. 2010); (Kelleher et al. 2013); (Fisher et al. 2013); (Kelleher et al. 2014).

1.2.1.4. Environmental Factors

Studies conducted on monozygotic and dizygotic twins, on a sample of 2,232 British children in a study from 5 to 12 years old showed a substantially higher psychotic symptoms concordance in monozygotic twins, with 41% compared to 22% concordance rate in dizygotic twins (Polanczyk et al. 2010), suggesting that psychosis is linked to genetic factors. However, they suggested that 57% of the variance was
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

explained by the environmental factors. A further study (Stepniak et al. 2014) was conducted to discover which environmental factors had an impact on the schizophrenia severity determinants, including 750 males with either schizophrenia or schizophreniform disorder. The research concluded that pre-adult cannabis use, mild parietal neurotrauma and perinatal complications were each strong predictors of age of onset. Further, while the study concluded migration, urbanicity and general psychotrauma could not individually result in higher risks of schizophrenia, the observed effect was such that exposure to multiple such factors could lead to early age of onset of schizophrenia-spectrum disorders. However, as they only performed the study on males, these environmental factors may not be related with an earlier age of onset in females.

In comparison with these findings, a follow-up study looked at a number of factors such as social class and status, place of birth, season of birth and immigration status. It largely replicated the results controlling for gender, family history of psychosis and diagnosis (O’Donoghue et al. 2015). This study suggests that cannabis use ($Z=-5.9$, $P=0.001$) and obstetric complications ($Z=-2.24$, $P=0.03$) were the primary risk factors, resulting in around 6 years and 2.7 years younger age of onset by cannabis use and obstetric complications respectively. However, they also concluded that social class, place of birth and time of birth were also factors that could increase risk of the age of onset, but only as part of the aforementioned cumulative effect. These results were replicated in an independent British survey sample with a size of 8,580 self-respondents (Johns et al. 2004). Psychotic symptoms were independently associated with several factors such as lower IQ, cannabis and alcohol dependence.

Caspi et al in 2005 reported a significant gene by environment interaction concerning cannabis use and *catechol-O-methyltransferase (COMT)* genotype in the SNP rs4680 for psychosis, in a sample of 803 people. The COMT gene has a substitution of Valine (Val) to methionine (Met) in a SNP at codon 158. The authors suggested that development of psychosis in adults can be due to a functional polymorphism in the
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

*COMT* gene that moderates the impact of cannabis use in adolescents ($b = 2.21$, $p$-value = 0.002 in Val/Val individuals (carriers of homozygote Val alleles); $b = 2.63$, $p$-value < 0.001 in Val/Met individuals (carriers of heterozygote genotype); not significant in Met/Met individuals) (Caspi *et al.* 2005).

A later family-based study (Nicodemus *et al.*, 2008) showed evidence of gene-environment interactions. Including 116 probands diagnosed with schizophrenia-spectrum disorder, they found significant interactions between obstetric complications and 4 different genes over 13 under study in cases. The interactions involved 3 SNPs in *Serine/Threonine Kinase 1 (AKT1)* (minimum LRT $p$-value = 0.012, OR = 3.89, 95% CI = (0.83, 18.2)), 2 SNPs in *Brain derived neurotrophic factor (BDNF)* (minimum LRT $p$-value = 0.011, OR = 0.15, 95% CI = (0.032, 0.73)), one in *Dystrobrevin Binding Protein 1 (DTNBPI)* (LRT $p$-value = 0.025, OR = 9.49, 95% CI = (1.23, 73.3)) and one in *Glutamate Metabotropic Receptor 3 (GRM3)* (LRT $p$-value = 0.035, OR = 3.39, 95% CI = (0.95, 12.17)). In the general population there is no significant correlation between the environmental factor and variants within the gene, which is an assumption (the authors assume) of gene-environment interaction using only cases. So, in order to know if the assumption was reasonable, the authors tested in controls the variants that were significant in gene-environment interactions in cases. They did not observe evidence of these interaction effects in controls which supports the assumption of the independence in controls ($N = 134$) (Nicodemus *et al.* 2008). As they used a family-based study design, the authors were not concerned about difference in maternal recall between cases and controls (Nicodemus *et al.* 2008).

### 1.2.1.5. Dopamine Hypothesis

When modelling the onset of psychosis, neither epidemiological nor prodromal studies have been successful (Broome *et al.* 2005). It is also necessary to bear in mind our current knowledge regarding neurochemical causes of such symptoms. For instance, the role of dopamine dysregulation in psychosis was established by a study showing
increased dopamine release in response to amphetamine challenge (Laruelle et al. 1996); which is a psychopharmacological test to study if patients are likely to suffer psychosis after dopamine agonist ingestion. More pertinently, the volume of dopamine released correlated with the presence of positive symptoms in patients, as expected, this also positively affected the effectiveness of dopamine blockers in treating these symptoms (Laruelle et al. 1996); (Abi-Dargham et al. 2000).

1.2.2. Risk Factors Overview

As one of the leading causes of disability (WHO | World Health Organization, 1946), disorders that have psychosis symptoms or are characterized by them represent a serious challenge to health. Genetic and environmental factors have been related to psychosis. Furthermore, the genetic epidemiology of psychiatric disorders often indicates complex models in which gene-environment interactions have a significant impact (Cristóbal-Narváez et al. 2016). The aetiology of psychosis consists of a complex combination of factors ranging from environmental stressors (Cryan and Dinan 2012); (Kavanagh et al. 2015), genetic predispositions (Bruenig et al. 2014); (Sullivan, Daly, and O’Donovan 2012), and neurodevelopmental abnormalities (Eisenberger and Cole 2012). Childhood trauma, affecting around 5% of children, has been identified as one of the strongest environmental risk factors in these disorders as well as worsening pre-existing conditions (Polanczyk et al. 2010). Moreover, prenatal factors, obstetric complications and drug abuse have been shown to play a relevant role to the development of schizophrenia (Weinberger 1987); (Cannon et al. 2002); (Chen et al. 2005). The idea that psychotic disorders might be attributable to a collection of single major genes has undergone multiple tests using comparisons of the observed recurrence risks in various classes of relatives and those predicted by this type of genetic model. Rather than confirming the monogenic signal, these studies suggested that the mode of inheritance for these disorders is liable to be either oligogenic, polygenic, or a mixture of genes with different effect sizes (O’Rourke et al. 1982); (Risch 1990); (Craddock et al. 1995); (Culverhouse et al. 2002). This makes
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

polygenic risk score (PRS) models (additive effects of several genes); and epistatic models, where several genes interact with one another, quite likely to influence psychosis as the likelihood of association between a major single gene model and psychosis is quite low. Moreover, the contribution of gene-environment interactions has been shown to have risk for psychosis (Nicodemus et al. 2008); (Cristóbal-Narváez et al. 2016).

Many genome-wide association studies (GWAS) have been performed to discover genes that have some impact in psychotic disorders. Several genes have been related to the two major disorders featuring psychosis: schizophrenia (S. M. Purcell et al. 2009); (Ripke et al. 2011, 2014) and BP (Sklar et al. 2011), showing a strong genetic similarity between both. Research on MDD has not been significantly linked to any genotype before 2015. To date 3 studies found genome-wide significant associations with MDD (N. Cai et al. 2015); (Power et al. 2017). In most cases the absence of genome-wide significant evidence has been thought to be caused by small sample sizes, therefore, the use of meta-analysis has been increased in the scheme of genetics, or due to the complex genetic models that underlie the aetiology of psychotic disorders. Therefore, in 2014 the Psychiatric Genetics Consortium 2 (PGC2) performed the largest-ever schizophrenia GWAS and found 108 GWAS-significant common susceptibility variants which confer risk in developing schizophrenia (Ripke et al., 2014). In addition, the authors found a significant polygenic impact on schizophrenia of a large number of small allelic effects, taken together having a bigger contribution that any single variant (R² around 18%). Unfortunately, the increment in risk by the polygenic models is still moderate in several studies (Ripke et al. 2011), although in the next study the model explains 18.4% (Ripke et al. 2014).

This, therefore, supports the hypothesis that psychosis risk may be influenced by epistasis (gene-gene interactions), as the heritability of psychotic disorders is high (heritability of schizophrenia is around 80% (Gejman et al. 2010) and BP is also quite high, between 60% and 85% (Barnett and Smoller 2009)). There is not a uniquely
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

most-powerful study design nor statistical technique for the detection of epistasis, but due to the vast number of markers, machine learning (ML) methods have become quite attractive and reasonable to apply in such studies. The choice of the most-powerful ML technique to be applied is unknown and thus research to lay the foundation for the use of ML in GWAS is critical to provide guidelines on their use.

1.2.3. Family and twin studies

In order to observe evidence of the heritability of psychotic disorders, several researchers have performed family and twin studies (Shih et al. 2004). These types of studies seek to determine the risk of acquiring a particular disease if related individuals also have it. The so-called twin study is one of the clearest ways to investigate heritability. These studies look for a similar feature between both monozygotic and dizygotic twins (identical and non-identical, respectively); as identical twins have the same genome and the non-identical share approximately 50% of genetic components in a majority of features like normal siblings (Polderman et al. 2015).

For instance, a recent sibling study (Pettersson et al. 2016) found that several psychiatric traits are influenced by the same genetic factors. This study selected a sample of adults living in Sweden (total adults 3,475,112) who had been diagnosed with at least one psychiatric disorder. With the objective of maximizing the probabilities of a similar shared environment, they included the two oldest siblings in each family, with no more than 5 years of difference. Hence, the final samples consisted of 1 466 543, 129 715 and 141 298 pairs of full siblings, maternal half-siblings and paternal half-siblings. The different mental diseases under study were: schizophrenia and schizoaffective disorders; depression and BP; drug abuse; ADHD; anxiety; and alcohol use disorder, together with these disorders the authors also considered convictions of violent crimes, confirming previous twin and family studies that suggested shared genetic factors among some psychiatric disorders (Cardno and Owen 2014).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

1.2.4. Linkage Analysis

A different type of analysis in genetic epidemiology is called genetic linkage analysis which aims to detect regions in the genome disconnected by a few gametic division events or meiosis, that are likely co-segregated, and that are related with a disorder or trait in related people. In other words, genetic linkage analysis is a technique to identify the area of the chromosome of genes influencing diseases or traits (Teare et al. 2006). There exist two different types of linkage analysis which are parametric and nonparametric analysis. Parametric linkage analysis determines the relation between a phenotype and a genotype when there is a specific genetic model for the phenotype. It can be applied if there is enough information from the parameters such as inheritance mode and genome from several participants from informative families (families where one parent has a heterozygous disease allele or where the siblings have distinct phenotypes due to the presence of at least two alleles in the family (Laird and Lange 2006)).

On the other hand, nonparametric linkage analysis should be employed if there is no information from the genetic model of the phenotype. Nonparametric tests are usually called model-free tests as they are based on fewer assumptions. In fact, the outcome or genetic model is not assumed to follow a normal distribution, no assumptions on the trait allele frequencies or on the mode of inheritance; but they require to use a marker model based on the observed marker data on the family members. They can test for an increase of sharing among patients with a particular phenotype. In general, in statistics, parametric tests are based on the assumption that the sample or parameters under study follows a known distribution in contrast to non-parametric tests which do not require any information about the distribution of the data, and so does not require any such assumptions (Vickers 2005). Parametric linkage analysis is based on the logarithmic odds score known as LOD-score method, which measures the genetic distance between two loci. However, when the model of the disease is unknown, model-free methods are considered, some of them are an extension of the LOD-score
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

methods and other estimates excessive allele sharing among patients with the disease, the later are known as non-parametric linkage models (NLP) (Sham et al. 2000).

Parametric linkage analysis is a powerful model for studying major gene disorders (Korkeila et al. 1991). When studying complex disorders non-parametric tests are more suitable as the disease model is unknown (Sham et al. 2000). Although linkage analysis has been effective in detecting evidence for Mendelian traits or diseases, they have not been powerful tests to study the underlying genetics of psychiatric disorders; even though they have been useful in another complex disorders such as Alzheimer disease and dementia (Guerreiro et al. 2012).

For instance, a study with 18 cohorts (>1,929 affected individuals) found regions in chromosomes 8p, 13q and 22q (Badner and Gershon 2002) with significant relation or linkage with schizophrenia, by applying one method called Multiple Span Probability (a method to combine \( p \)-values, it is an extension of Fisher’s \( p \)-value method). These authors also observed significant linkage with BP in regions of chromosomes 13q and 22q (MSP < 0.001) (2002) in a meta-analysis of 11 studies, including chromosomal regions that showed evidence considering a \( p \)-value < 0.01 (1,228). Another study using the same cohorts, but instead applying a meta-analysis based on the combination of summary statistics from each cohort (linkage statistics or \( p \)-values), the authors performed the analysis using the original genotype data from each independent study. They performed a non-parametric linkage analysis to study each cohort individually as well as the combined dataset from the 11 studies. They found evidence after correcting for multiple testing at chromosomes 6q and 8q (McQueen et al. 2005).

1.2.5. Association Studies

In comparison to linkage studies, association studies attempt to detect significant differences of frequencies of alleles in different individuals, currently with GWAS and previously with candidate gene studies (Korkeila et al. 1991). These individuals may be affected individuals by a disease (cases) and individuals without the disease
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

(controls), or individuals with different phenotypes for a trait such as quantitative phenotype like height or gene expression. Association studies are applied in order to identify candidate genes or genome regions that have an impact on a particular disorder or a trait, looking for any association between genetic variation and a phenotype. For instance, in cases-control association studies (the most common), higher allele frequency in affected individuals can be interpreted as meaning that the variant increases the risk of the phenotype (Lewis and Knight 2012). As before, these studies are generally performed in one of two ways: the candidate gene or a GWAS. The first one aims to identify a particular gene as well as the gene features like particular allele variations or SNPs (Ardlie et al. 2001), while GWAS try to detect different genes (at the same time looking at hundreds or thousands SNPs) in the genome which have risk for the phenotype.

1.2.5.1. Candidate genes in Psychosis

Before the GWAS era studies have made inroads in identifying candidate genes that can increase risk for psychotic disorders in patients (Harrison and Owen 2003); (Owen, Williams, and O’Donovan 2004) including those that regulate the glutamate system which is strongly linked to the regulation of dopamine levels, particularly neuregulin 1 (NRG1), dysbindin (DTNBPI), and disrupted in schizophrenia 1 (DISC1). Another gene catechol-O-methyltransferase (COMT); so identified is responsible for the breakdown of dopamine in the prefrontal cortex (Egan et al. 2001); (Malhotra et al. 2002); (Rosa et al. 2004). In addition, genes which have shown association are D-amino acid oxidase inhibitor (DAOA) and Dopamine Receptor D2 (DRD2) (Ross et al. 2006); (Straub and Weinberger 2006); (Riley and Kendler 2006); (Serretti and Mandelli 2008); (Nick Craddock and Sklar 2009); (Parsons et al. 2007).

The idea that psychotic disorders might be attributable to single major genes has been refuted (O’Rourke et al. 1982); (Risch 1990); (Craddock et al. 1995). More recent data suggests that the additive effect from PRS is to be preferred (PGC, Ripke et al., 2014).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Here, GWAS play an important role. However, GWAS have not found significantly associated variants in these genes of any psychotic disorder, with the exception of \textit{DRD2}. Some authors support the idea that the lack of a relevant genome-wide association is related to the common disease/rare variant hypothesis (Porteous \textit{et al.} 2014) (rare variants that have relatively high penetrance are playing an important role on the genetic susceptibility to common diseases), and that GWAS might be hiding highly penetrant mutations, which are contributing for the risk of the disorder (Gibson 2012). Penetrance is the proportion of people having the disease conditionate of having the variant; in other words, is the probability of having the disease when the individual has the variant or mutation. Under the rare alleles with high penetrance model, disease-causing alleles have a frequency of less than 1\%, their effect on individuals with this variants are modified by other loci or the environment, although rare alleles would be largely responsible for the disease. In the case of schizophrenia for instance, if each of the rare variants from a collection that are attributable to disease explain most of the risk in only affected individuals, the effects of these variants will not be detected by standard GWAS procedures (MAF > 1\%), which detect the effects of these alleles in the general population (Gibson 2012).

\subsection*{1.2.5.2. GWAS in Psychosis}

During the last decade, with the improvement of high-throughput genotyping technologies, GWAS became a key way to find candidate genes associated with psychotic disorders, such as schizophrenia, BP and MDD. Before 2009, year of the first GWAS publication of PGC, only one study (O’Donovan \textit{et al.} 2008) found significant associations in schizophrenia passing the GWAS threshold \(p\)-value \(5 \times 10^{-8}\) (Bonferroni corrected \(p\)-value). The authors found in a meta-analysis for schizophrenia case-control status the first GWAS significant result (\(p = 9.96 \times 10^{-9}\)), between a marker in \textit{zinc finger protein 804A (ZNF804A)} and the phenotype which included cases (\(n = 9,173\)) with schizophrenia and BP, and controls (\(n=12,834\)), the total number of SNPs studied was 362,532. One of the reasons why other studies were not able to
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

find a genome-wide significant association could be because of low statistical power due to small sample sizes; in addition, many studies were unsuccessful in finding significant results on replication studies (independent replication datasets).

In 2009, the International Schizophrenia Consortium (ISC) (Purcell et al., 2009) studied the genome-wide association of around 1 million of SNPs with schizophrenia in study with 3,322 affected individuals and 3,587 healthy individuals using the Cochran-Mantel-Haenszel test and logistic regression. The most significant marker ($p$-value = $4.79 \times 10^{-8}$) was in notch 4 ($NOTCH4$), a non-immune system gene in the major histocompatibility complex (MHC) (Mokhtari and Lachman 2016).

The same day, a larger GWAS was published including this latter one with the aim of testing the genetic risk of 1,500 markers (Stefansson et al. 2009) in schizophrenia using the generalized likelihood ratio test for association and combined studies using the Mantel–Haenszel model. The authors reported 7 genome-wide association SNPs, 5 of them in the MHC region: one in histone cluster 1 H2B family member J ($HIST1H2BJ$) ($p$-value = $1.1 \times 10^{-9}$); two SNPS which were the most significant ones in protease, serine 16 ($PRSS16$) with $p$-values $1.3 \times 10^{-10}$ and $1.4 \times 10^{-12}$; and the other two in pterin-4 alpha-carbinolamine dehydratase 1 ($PGBD1$) and the $NOTCH4$, with $p$-values $8.3 \times 10^{-11}$ and $2.3 \times 10^{-10}$ respectively. Again, results confirmed the relation between schizophrenia and genes in the MHC region. The other two loci involved in significant associations were in neurogranin ($NRGN$) ($p$-value = $2.4 \times 10^{-9}$) and in transcription factor 4 ($TCF4$) ($p$-value = $4.1 \times 10^{-9}$). The expression of the first affects only the brain, and has been related with working memory showing less activation in cingulate cortex (Krug et al. 2013) and alterations in left superior frontal (Rose et al. 2012) and the second to Pitt-Hopkins syndrome, a disorder characterized by mental delays and severe motor (Brockschmidt et al. 2007).

The Schizophrenia Working Group of the Psychiatric GWAS Consortium (PGC) (Ripke et al. 2011) have shown genome-wide significant associations between 7 loci
and schizophrenia in their first GWAS report performing standard logistic regression and combining different samples. The study tested the GWAS significance of 1,252,901 autosomal SNPs for risk of schizophrenia in a large sample of 9,394 affected individuals and 12,462 healthy people from 17 different cohorts, resulting in 5 novel genome-wide significant variants, the most significant was mapped to the microRNA 137 (MIR-137) gene (p-value = 2.65x10^{-6}; OR = 1.11%; 95% CI (1.07–1.16)), the second at cyclin and CBS domain divalent metal cation transport mediator 2 (CNNM2) and 5'-nucleotidase, cytosolic II (NT5C2), and the other three to the encoding matrix metallopeptidase 16 (NMP16) gene, the encoding CUB and Sushi multiple domains 1 (CSMD1) gene and the prostate-specific transcript 1 (PCGEM1). One month later, another study confirmed a significant association with tripartite motif containing 26 (TRIM 26) and coiled-coil domain containing 68 (CCDC68) (Steinberg et al. 2011). Moreover, they performed the same analysis to find genome-wide association factors related to schizophrenia, schizoaffective disorder and BP, in this case the sample included 16,374 cases with and 14,044 controls. Their findings suggested that three genes were associated with schizophrenia and BP, encoding calcium channel, voltage-dependent, L type, α 1C subunit (CACNA1C), the region containing encoding inter-α (globulin) inhibitors H3 and H4 (ITIH3 - ITIH4) and encoding ankyrin 3 (ANK3) genes previously associated to BP (Ferreira et al. 2008); (Scott et al. 2009); (Green et al. 2010).

Furthermore, the most recent schizophrenia GWAS was published by the Schizophrenia Working Group of Psychiatric Genomics Consortium 2 (PGC2) (Ripke et al. 2014). This time they performed the largest GWAS including 36,989 cases and 113,075 controls, and more than 9 million markers. This study resulted in 108 loci genome-wide significant associations with schizophrenia, 83 discovered for first time. The most significant SNP (p-value = 3.48 x 10^{-31}) was on chromosome 6, in the MHC region.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

In addition, researchers have also investigated the genetic factors underlying the psychotic disorder BP, however, just a few studies successfully discovered genome-wide significant SNPs. The first reporting genome-wide significant association was performed by Baum et al. (2008). The study found a significant SNP, rs1012053 (p-value = 1.5 x 10^{-8}) in diacylglycerol kinase eta (DGKH) which increases risk for BP. The study included two independent European samples, the original sample included 461 cases and 562 controls, and the replication one had 772 affected patients and 876 healthy individuals (Baum et al. 2008). Then, two more genes were significantly associated (calcium channel, voltage-dependent, L-type, alpha 1C subunit (CACNA1C) and ankyrin 3 (ANK3) with BP (Ferreira et al. 2008). The study performed a meta-analysis of 3 studies with 4,387 cases and 6,209 controls, including the Welcome Trust Case Control Consortium (WTCCC, 2007) sample, which was carried out by Ferreira et al. (Ferreira et al. 2008), the study used logistic regression to model the BP risk among 1.8 million variants. Later, in 2011 the PGC Bipolar Disorder Working Group published the largest GWAS to date in BP, which found 2 significant loci applying logistic regression, one of them novel (Sklar et al. 2011). The study reported extra evidence for the association of CACNA1C and a new genome-wide associated gene protein odd oz/ten-m homolog 4 (ODZ4) in a replicated sample of 11,974 individuals with bipolar and 51,792 controls, as well as they confirmed the strong role that CACNA1C has in schizophrenia and BP in a study combining the a sample within PGC study (7,481 cases and 9,250 contols) and the independent sample used for replication, suggesting an important relation with psychosis phenotype. Furthermore, in the same year another case-control study, 2,411 patients and 3,613 controls, found another significant variant for BP in the neurocan (NCAN) gene, (Cichon et al. 2011) with a p-value 3.02×10^{-8}.

The most recent BP GWAS (Hou et al. 2016a) studied the association between more than 9 million autosomal genetic variants in two stages. Fist, the sample contained 7,647 cases and 27,303 controls and more than 60 markers were observed to be GWAS-significant involving two genes, one of those was near the tetratricopeptide
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

repeat and ankyrin repeat containing 1 (TRANK1), all rest were in the gene mitotic arrest deficient like 1 (MAD1L1) gene. Both genes had previously shown evidence for association with BP (Chen et al. 2013) and the latter had shown also evidence in SZC and BP combined study (Ruderfer et al. 2014). Then, they found 2 new candidate loci, erb-b2 receptor tyrosine kinase 2 (ERBB2) gene and a region near the ELAV like RNA binding protein 2 (ELAVL2) (p-values $4.53 \times 10^{-9}$ and $5.87 \times 10^{-9}$ respectively), in a meta-analysis including 9,784 cases and 30,471 controls (Hou et al. 2016b); (Hou et al. 2016a).

The other psychiatric disorder which might develop psychosis is MDD, the first genome-wide significant loci was reported in 2015 by the CONVERGE consortium (CONVERGE consortium 2015), although it was the forefront of many studies previously. This GWAS included only Han Chinese women with severe recurrent MDD including 10,640 (5303 women with recurrent depression and 5337 healthy women); in this way the authors ensured homogeneity in the sample, increasing the statistical power compared to other studies aiming to detect genetic risk factors in MDD. Two SNPs showed GWAS significance over more than 6 million tested, one near sirtuin 1 (SIRT1) gene and the another in phospholysine phosphohistidine inorganic pyrophosphate phosphatase (LHPP). Recently, a MDD GWAS meta-analysis reported 15 new loci replicated across 3 cohorts from European ancestry (Hyde et al. 2016). One cohort included 75,607 self-reported MDD and 231,747 controls, the other one was the PGC MDD data (9,240 MDD cases and 9,519 controls) (Ripke et al. 2013), and the 23andMe with 45,773 cases and 106,354 controls.

The most recent research with GWAS significant associations in MDD was performed by the PGC Major Depressive Disorder Working Group (Power et al. 2017). They studied associations with both the early-onset and late-onset MDD as well as to the intermediate age at onset from 1,235,109 autosomal SNPs in 9 cohorts with a sample of 8,920 cases and 9,521 controls, dividing the affected individuals in eight groups by age at onset. To study GWAS associations with early-onset and late-onset MDD, they
performed sequential GWAS analysis adding cases applying logistic regression using PLINK. In the early-onset MDD study, they started analyzing GWAS associations including cases in the earliest onset versus all controls, then they considered the second group early-onset and they performed GWAS using the combined cases against all controls; they studied continue testing GWAS associations until all cases were under study. The analysis for the late-onset MDD was performed following same process but now starting for the latest onset subset. And then, the authors performed a GWAS analysis including the four intermediate group of age at onset to study whether the two earliest or latest groups of age of onset introduced heterogeneity to the affected individuals. All of these GWAS analysis were performed considering four different cases, all cases, only affected males, only affected females, and only patients with recurrent MDD. Therefore, because of the large amount of tests, the GWAS p-value threshold after multiple testing decreased to \( p < 9.5 \times 10^{-10} \). They excluded SNPs that were highly significant without a specific effect of Age at onset. The authors found only one GWAS significant intergenic SNP rs7647854 on the chromosome 3 that was found in the half oldest onset group of cases against controls (\( p\)-value = \( 3.4 \times 10^{-11} \)). The association of this SNP was tested for replication in nine independent cohorts including 6,107 cases (individuals with half oldest age at onset) and 124,230 controls were showed a significant association with MDD (\( p\)-value = \( 7.5 \times 10^{-4} \)). Moreover, the SNP showed significant association in meta-analysis including the validation sample and each replication cohort (\( p\)-value = \( 5.2 \times 10^{-11} \), OR = 1.16, 95% CI: (1.11, 1.21)).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>STUDY</th>
<th>DISEASE(S) OF INTEREST</th>
<th>FINDINGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baum et al. (2008)</td>
<td>BP</td>
<td>Reported genome-wide significant association between bipolar disorder and DGHK in two independent samples.</td>
</tr>
<tr>
<td>O’donovan et al. (2008)</td>
<td>Schizophrenia and BP</td>
<td>Reported strong evidence for association between ZNF804A and schizophrenia, attaining genome-wide significance when both schizophrenia and bipolar disorder were considered.</td>
</tr>
<tr>
<td>Ferreira et al., (2008)</td>
<td>BP</td>
<td>First study to report genome-wide significant associations with bipolar disorder implicating CACNA1C and ANK3.</td>
</tr>
<tr>
<td>Stefansson et al. (2009)</td>
<td>Schizophrenia</td>
<td>Identified genome-wide significant association with schizophrenia at loci in the major histocompatibility complex (MHC), NRGN, and TCF4.</td>
</tr>
<tr>
<td>Steinberg et al. (2011)</td>
<td>Schizophrenia</td>
<td>Provided evidence in support of association between schizophrenia and NRGN and TCF4. Identified two novel loci associated with schizophrenia at CCDC68 and VRK2</td>
</tr>
</tbody>
</table>
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>Study</th>
<th>Disorder/Combination</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cichon et al. (2011)</td>
<td>BP</td>
<td>Identified a genome-wide association between bipolar disorder and NCAN.</td>
</tr>
<tr>
<td>PGC (Ripke et al., 2011)</td>
<td>Schizophrenia and schizophrenia and BP combined</td>
<td>Seven schizophrenia-associated loci identified, five of which were novel and mapped to six genes (MIR137, PCGEM1, CSMD1, MMP16, CNNM2 and NT5C2). Confirmed association between schizophrenia and TRIM26 and CCDC68. Also reported association between schizophrenia and bipolar disorder and CACNA1C, ANK3 and ITIH3-ITIH4, all previously associated with bipolar disorder.</td>
</tr>
<tr>
<td>PGC (Sklar et al., 2011)</td>
<td>BP and schizophrenia and BP combined</td>
<td>Confirmed evidence for association between bipolar disorder and CACNA1C. Identified a novel susceptibility locus at ODZ4. Reported association between schizophrenia and bipolar disorder combined and NEK4, CACNA1C and a multi-gene region spanning ITIH-1, -3 and -4.</td>
</tr>
<tr>
<td>Chen et al. (2013)</td>
<td>BP</td>
<td>Genome wide significant association with bipolar disorder reported near TRANK1, LMAN2L and PTGFR. Also</td>
</tr>
</tbody>
</table>
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Condition</th>
<th>Association</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ruderfer et al. (2014)</td>
<td>BP and schizophrenia</td>
<td>Identified a novel association between both disorders and (PIK3C2A), as well as five previously-identified loci (TRANK1, MHC, MAD1L1, and CACNA1C)</td>
</tr>
<tr>
<td>PGC (Ripke et al., 2014)</td>
<td>Schizophrenia</td>
<td>108 genome-wide significant loci consisting of intergenic regions, single genes and multiple genes. The top hit was a broad 400 kb region on chromosome 6, within the MHC</td>
</tr>
<tr>
<td>Converge Consortium (Cai et al., 2015)</td>
<td>MDD</td>
<td>First report of genome-wide significant association in MDD. Two genome-wide significant loci identified on chromosome 10 located 5’ of SIRT1 and within an intron of LHPP.</td>
</tr>
<tr>
<td>Hou et al. (2016)</td>
<td>BP</td>
<td>Two novel genome-wide significant loci were identified: ERBB2 and an intergenic region on chromosome 9. Also reported association between MAD1L1 and bipolar disorder only.</td>
</tr>
<tr>
<td>Hyde et al. (2016)</td>
<td>MDD</td>
<td>15 novel genome-wide significant loci were identified in a meta-analysis of a</td>
</tr>
</tbody>
</table>
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

previous GWAS of MDD by the PGC (Ripke et al., 2013b) and consumer genomic data from 23andMe.

| PGC (Power et al., 2017) | MDD | rs7647854 showed risk for MDD (GWAS significant). The SNP is on chromosome 3. |

Table 1.1. Summary of genome-wide significant findings for schizophrenia, bipolar disorder and MDD identified by GWAS (2008-2016). Table summarises GWAS of psychiatric disorders in which genome-wide significant results have been reported based on a p-value threshold of 5 x 10^-8. “Study” column provides the references to each study, column labelled “Disease(s) of Interest” refers to the disease or diseases under investigation in each study while the column labelled “Findings” summarises the genome-wide significant disease-associated findings of each study.

1.2.6. Polygenic Risk Scores

Because of the small effects of each significant variant from GWAS, researchers aimed to look for additive effects that may be involved in psychotic disorders, in other words, they attempted to investigate PRS which may explain variation of psychotic disorders to use them for classification studies as well as for predicting particular phenotypes. Essentially, a PRS is a variable constituted by the linear combination (additive) of different variants, which showed risk of the phenotype considering a particular p-value threshold in a reference GWAS. After taking those SNPs, the PRS is constructed by the sum of the number of risk alleles in the study weighting them by the logarithm of their effect (OR) in the reference study (Purcell et al. 2009); (Dima and Breen 2015). Different studies have found an association between PRS and schizophrenia, where the precision of the results and their prediction in schizophrenia improve as the sample size increases. These association would be expected as the additive effects include the effect of multiple SNPs considering even those who had a p-value greater than the GWAS p-value threshold or even greater than 0.05 on the study of reference, as the value of p-value is arbitrary.
In 2009, Purcell et al. (ISC; (International Schizophrenia Consortium et al. 2009)) found significant cumulative risk for schizophrenia in a case – control study. The study considered a sample of 2,176 affected males and 1,146 affected females, and 1,642 male controls 1,945 female controls, and included 74,062 autosomal SNPs. The PRS was constructed using a $p$-value threshold of 0.5 and it involved 37,655 SNP, its significance resulted in a small $p$-value = $9.4 \times 10^{-19}$ and its contribution accounts for about 3% of schizophrenia risk (Nagelkerke’s pseudo $R^2$). They also showed significant additive effect associated with BP from that PRS (Purcell et al. 2009). In 2011, one study used the later study as reference (PRS associated with both schizophrenia and BP) and the PRS was significantly associated with BP (Sklar et al. 2011). In addition, a PGC study (Ripke et al. 2014) found evidence of additive effects of a PRS ($p$-value threshold 0.05) for the risk of schizophrenia using a larger sample size, which explains approximately 18% of the variance of schizophrenia. The PGC also studied polygenic associations with both early-onset and late-onset MDD using logistic regression including covariates (study indicators and 20 principal components) in different case-control studies using all cases against all controls, cases in the two earliest-onset octiles against all controls, cases in the two latest-onset group against all controls, and two earliest-onset groups against the two latest-onset cases (Power et al. 2017). The PRSs tested in the study were for schizophrenia and BP (9,379 cases and 7,736 healthy controls, 6,990 cases and 4,820 controls respectively), Alzheimer’s disease (3,177 cases and 7,277 healthy controls) and coronary (22,233 cases and 64,762 controls). Significant associations were found between the PRSs for schizophrenia ($R^2 = 0.67\%$, $p$-value = $3.0 \times 10^{-19}$, including early-onset cases vs controls; $R^2 = 0.14\%$, $p$-value = $3.9 \times 10^{-5}$ late-onset cases vs controls) and BP ($R^2 = 0.41\%$, $p$-value = $1.4 \times 10^{-12}$ early-onset cases vs controls; $R^2 = 0.16\%$, $p = 1.9 \times 10^{-5}$) with early- and late-onset MDD. Furthermore significant association between the PRS for coronary artery disease and MDD was detected ($R^2 = 0.05\%$, $p$-value = 0.01 early-onset cases versus controls; $R^2 = 0.05\%$, $p$-value = 0.01 late-onset cases versus controls; $R^2 \leq 0.01\%$, $p$-value = 0.76 early-onset cases versus late-onset cases).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Due to the small amount of variance explained by polygenic risk factor models in psychosis, this led researchers to investigate the impact of interaction effects which may contribute more to psychotic disorders.

### 1.2.7. Epistasis

Epistasis refers to gene-gene interactions, there are two types: biological and statistical (Cordell 2002). Biological epistasis means the physical interactions between genes, such as when an allele at one locus can mask or modify the allele effect in one or more loci affecting a particular phenotype. Statistical interaction is the phenomenon that happens when the effects of two or more genes have an effect either significantly higher than or lower than the additive effect, and the interaction between 2 or more loci contributes to variation in the phenotype, and the effect of the interaction is statistically significant different between individuals such as affected and healthy individuals in case-control studies.

Psychotic disorders are genetically complex and the small effects per SNP may interact with one another (Cordell and Clayton 2005), which has made difficult the creation of models which to be successful should account for epistasis (Andreasen et al. 2012). Several authors have attempted to model epistasis using ML approaches. For instance, Nicodemus et al. (2010a) tested epistasis for risk in schizophrenia between DISC1 (12 SNPs), citron rho-interacting serine/threonine kinase (CIT) (19 SNPs), NudE Neurodevelopment Protein 1 Like 1 (NDEL1)(1 SNP), NudE Neurodevelopment Protein 1 (NDE1)(3 SNPs), Fasciculation And Elongation Protein Zeta 1 (FEZ1) (13 SNPs) and Platelet Activating Factor Acetylhydrolase 1b Regulatory Subunit 1 (PAFAH1B1) (2 SNPs). They performed the study using three different ML techniques random forest (RF), generalized boosted regression and Monte Carlo logic regression (MCLR) as well as likelihood ratio tests (LRTs) for nested models. Their findings showed interactions between genes related with psychosis, between NDEL1/CIT 4.44 (LRT p-value = 0.00013; OR=4.4; 95% CI (2.22, 8.88)), DISC1/CIT (LRT p-value =
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

0.007; OR = 3.07; 95% CI (1.37, 6.98)), and two between SNPs in CIT (LRT p value = 0.038; OR = 2.16; 95% CI (1.04, 4.46) and LRT p-value = 0.0030; OR = 2.90; 95% CI(1.45, 5.79)). (Nicodemus et al. 2010a). They validated two of these interactions in an independent neuroimaging study of healthy controls. The authors tested the interactions between NDELI/CIT and DISC1/CIT by performing N-back task in a BOLD working memory test in healthy controls to exclude a confounder variable because of the performance, for example, schizophrenia patients would not perform the task better but they would show higher activation in the brain. The interactions showed significant less efficient cognitive processing prefrontal in healthy controls although they did not find replication in independent genetic datasets, the replication database did not have the same genotypes as in the GWAS.

Later in the year, Nicodemus et al. (Nicodemus et al. 2010b) published a genetic and neuroimaging study suggesting significant epistasis between 3 other genes for schizophrenia risks using three ML algorithms, RF, conditional inference forest (CIF) and MCLR. Two hundred and ninety six affected individuals and 365 healthy individuals were under study and LRTs of logistic regression models were performed to test the significance of the interaction between NRG1, erb-b2 receptor tyrosine kinase 4 (ERBB4) and AKT1 (LRT p-value = 0.042, OR = 27.13). The interaction was replicated in a functional neuroimaging study (fMRI) with a sample size of 114 individuals. This study showed that the interaction between NRG1, ERBB4 and AKT1 was significant in the brain function in healthy individuals, associated with working memory, impacting in a less efficient processing of the dorsolateral prefrontal cortex (Nicodemus et al. 2010b). Andreasen et al. (2011) deployed a combined ML technique to find significant interactions with schizophrenia, their results suggested 17 interacting SNPs mapped to 5 genes: phosphodiesterase 4B (PDE4B), reelin (RELN), ERBB4, DISC1, NRG1, some of the SNPs confirmed previous relations to the disease (Andreasen et al. 2012).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

In addition, in 2013 a GWAS found significant gene-gene interactions that contributes risk for BP (Judy et al. 2013). The study included 3,849,034 genotypes as well as 2,191 affected individuals and 1,434 unaffected people to test for 2-way interactions between ANK3 and each interacting gene identified by STRING, a database of predicted protein-protein interaction (von Mering et al. 2003) using regression and permutation procedures. The authors showed both biological evidence (STRING) and statistical evidence ($p$-value $= 3.18 \times 10^{-8}$; permuted $p$-value $= 0.005$) for epistasis between ANK3 and Potassium Voltage-Gated Channel Subfamily Q Member 2 ($KCNQ2$).

### Issues in Big Data Omics and Machine Learning

#### Overview

Big Data is a recent and very used term in real studies, but it is still an unclear and confused term. The term of Big Data considers the three “vs”: volume, velocity and variety (Walesby et al. 2017). The size of the data, the time spent when it is generated and the different forms where the data is stored or available characterise the Big Data term. In this section, Big Data term refers to the volume of data and when there is more variables (p) than observations (n). The Omics terms refers to genetic data.

#### 1.3.1. Problems of Classical Statistics

Over the last decade, in psychiatric genetics, the amount of data available for analysis has dramatically increased, leading to high dimensional databases with more features (p) than observations (n) where variables are correlated and where variables may interact, and classical statistical approaches may lead to overfitting (Iniesta et al. 2016) (larger p than n, correlation between variables and interaction between variables are discussed below). Furthermore, small sample size is a real problem which leads to other detrimental situations such as low statistical power, an increase of false positives and false negatives as well as of the effect size estimation and low reproducibility (Button et al. 2013); (Colquhoun 2014). In fact, the single effect from each variant in
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

psychiatric disorders is low (low effect size), so in order to have statistical power the required sample size is large (p should be large).

One of the reasons for such large databases is the continuous reduction of time and cost of sequencing technologies by a factor of 1 million in less than 10 years (Mardis 2011). In psychiatric genetics, statisticians, bioinformaticians and biologists study different types of data to investigate the molecular biology of illnesses, such as gene expression data, both microarrays and RNA-seq; protein data, metabolomics data, and single nucleotide polymorphisms (SNPs) (Lee et al. 2013); (Martins-de-Souza 2014); (Mostafavi et al. 2014); (Jansen et al. 2016).

1.3.1.1. The “Small N, Large P” Problem

Nowadays, most data sets coming out of modern genetic techniques are high-dimensional, so the number of observations (n) is not similar to the number of variables, features or predictors (p), and in most omics analysis n is lower than p (n < p). To deal with high dimensional data is not easy. It poses statistical challenges as classical approaches could give a different model each time you run it. They would fit the data equally well, but would predict terribly because there is no way to check that the solution they got this time is better than last time (Donoho and Stodden 2006). In big datasets there is information available regarding many predictors, but some predictors have some impact on psychiatric disorders and others are completely not useful. Including too many predictors in our models, the data are going to be overfitted. In other words, the statistical model has a great performance on the data used to develop it, but it will predict future observations quite poorly because the model would take into consideration variables that are not important and should have been dropped, as they introduce noise when predicting new observations. Therefore, working in high dimensions one have to be careful (Hawkins 2004).

Furthermore, in classical statistics, p-values and confidence intervals are the main tool to draw conclusions and determine evidence for rejection of the null hypothesis. In
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics, GWAS and genome sequencing have increased to the point that all data collected are analysed considering a large amount of information and, therefore, they attempt to have many genetic features (Storey and Tibshirani 2003). Then, to analyse the data, hypothesis tests are performed on many (million, thousands …) of genetic traits with the aim of rejecting as many hypothesis tests as possible, and in this way confirm with a high probability (usually greater than 95%) the statistical significance of the genetic features; while avoiding errors of saying that the feature is significant when actually is not, false positives (Gondro et al. 2013). So, this leads to a large number of simultaneous test where the probability of having a significant feature by chance increases as the number of test increases, this is called the curse of dimensionality (Dudoit et al. 2008).

To deal with the multiple testing problems, there have been several methods proposed in the literature, but the most common used in GWAS is Bonferroni correction (Bland and Altman 1995), which is the most conservative. The method assigns significance to those feature which have p-value less than the ratio between the significance level and the number of variables tested on the study. Therefore, in psychiatric genetics because of the large number of traits are tested to have risk for a phenotype, the p-value threshold by Bonferroni correction is quite small which makes it difficult to find genetic contributions to disorders. In fact, the GWAS p-value threshold is small (5 x 10^{-8}) and it is based on a study with 100,000 SNPs (Dudbridge and Gusnanto 2008). Hence, look for associations is not straightforward because is more likely to find false positives, whereas finding true signals is difficult because of lots of variables which provides low p-value thresholds (e.g. 5 x 10^{-6} with 10,000 SNPs). To solve those kind of issues, ML techniques such as, but not limited to, feature selection or regularization (they use L1-norm or L2-norm in the cost function that reduces the optimal values of the model parameters and thus prevent the model from overfitting such as Lasso or ridge regression) have become very popular and useful in analysing high dimensional data (Meinshausen and Bühlmann 2010); (Iniesta et al. 2016).
1.3.1.2. Variable Co-dependency

Genetic markers can be in Linkage Disequilibrium (LD) which shows a correlation pattern in the genome, so in high-dimensional genetic databases it is common to have correlated features. In the area of psychiatry, many different methods to analyse data have been employed such as multiple linear or logistic regression (Tse et al. 2015); (Watson et al. 2014); (Nery et al. 2007).

When fitting the above models to the data, the phenomenon called collinearity or multicollinearity should be taken into account. It occurs when two or more covariates are strongly correlated with each other. Regression models that suffer from collinearity might inflate the effect of the coefficients estimates as correlation may cause false positives and false negatives results when testing for their relevance (De et al. 2013). Therefore, when a model is fitted to the data, researchers should be aware of collinearity and test whether there is variable inflation. To avoid these problems, backwards feature selection using nested models and Chi-squared tests are normally used. Also, there have been publications studying the performance of ML algorithms under correlation conditions such as RF (Nicodemus and Malley 2009); (Nicodemus et al. 2010c); (Nicodemus 2011).

1.3.1.3. Interaction Effect Detection

In genetics, epistasis or interaction effects between two or among several SNPs on phenotypes are one of the challenges to study genetics risks of complex disorders, as the effect of SNPs explain only a small percentage of the heritability of such disorders (Moore et al. 2010). Several authors have used statistical models to study the significance of interaction effects such as penalized multivariate regression models. Park & Hastie (2008) proposed an extension of logistic regression (LR) using L2-regularization to detect interaction models, both gene-gene and gene-environment interactions. The performances of penalized LR and multifactor dimensionality
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

reduction (MDR) algorithms were compared in a simulation study. The results showed higher power in the penalized LR than the MDR detecting interaction models. Furthermore, the authors also compared both models with FlexTree in 2 real datasets, hypertension and bladder cancer data. Penalized LR showed the highest specificity (true negative rate), although it was low in the hypertension data; and higher sensitivity (true positive rate) and specificity in the bladder data. Their model was stable even with a high number of parameters (Park and Hastie 2008). More recently, Bien et al (2013) proposed an algorithm based on a set of convex constraints that are added to the lasso to capture weak interaction models, and they implemented their model in the R package hierNet (Bien et al. 2013).

Also, in genetics like in GWAS, authors study the association of millions of SNPs with a phenotype, if they attempt to test the effect of interactions, the number of pairwise interaction effects to study is much larger \( \frac{n(n-1)}{2} \) which aggravates the multiple testing problem and which also presents a computational challenge. For example, in psychiatric genetics where variables are correlated and the character of diseases is complex suggesting non single association factors, several authors have applied ML techniques to study the effect of the interactions on a particular phenotype such as schizophrenia as discussed in section 1.2.7 (Nicodemus et al. 2010a); (Nicodemus et al. 2010b); (Andreasen et al. 2012).

1.3.2. Why Machine Learning?

Humans show natural tendency to perform complex actions unconsciously following practice, for example writing or playing a musical instrument. ML tries to do the same, learning from the data to predict new findings (Michie et al. 1994). Statistical ML can be divided into two main areas: supervised and unsupervised algorithms. Supervised learning models train the data first find an association with an outcome; regression and classification belong to this group. On the other hand, unsupervised learning techniques do not have labels or outcomes, they try to find a particular signal in the
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

data or detect associations within data instead; clustering methods take place in this area (Ayodele 2010). Regression models aim to predict a quantitative outcome, such as logical memory, brain volume or intelligence quotient (IQ) in psychotic patients (Leeson et al. 2009). Otherwise, classification aims to predict a categorical response like having a psychiatric disorder or not, having a high score on social impairment or not or belonging to a subtype of genes. To fit a model both use a training set of observations from the sample, for example 2/3 of the sample; in this way, we can calculate the training error, but as the training error will be lower with more features in the model, the model has to be applied in an independent dataset with the observations used to fit the model not included to study the model performance. In this way the model performance is evaluated on the test set, estimating the test error rate on a future observation. In high-dimensional data, when p>>n, as said in the section above, we have to be careful to not overfit the data. To do that we must rely on test error.

In addition, as explained on the section above, in “Big Data” studies the number of features is large, and in the most cases like in omics data, databases have many noise variables which will easily increase the risk of overfitting, and the difficulty deploying a model that will work well on future observations, such as genes or SNPs that are not associated with the outcome or phenotype. But it is necessary to detect a signal or true features that are useful to explain the outcome under study. This is addressed by feature or variable selection (Guyon and Elisseeff 2003); (Kohavi and John 1997) that will reduce the dimensionality of the data, and hence delete noisy variables, select the relevant ones, and reduce the probability of overfitting to the training set.

There are three main different ML techniques that have been applied to select variables: filter, ensemble and wrapper algorithms. Filter algorithms use a ranking based on the probability of each variable to predict an outcome, the best subset of features form the input to the algorithm (Yu and Liu 2004). Ensemble algorithms (Saeys et al. 2008) might be applied following this filtering purpose, for example RF
measures the importance of each feature to be associated with an outcome, and gives a ranking by the importance. Wrapper algorithms choose a set of features to construct a model that might be significant and they test its efficacy (how well they explain the outcome), then the group of features is changed to compute its efficacy again. Finally, the best model is chosen (Kohavi and John 1997).

### 1.3.2.1. Dimensionality – Epistasis

In high-dimensional studies like GWAS, dimensionality is a problem as it involves a large number of SNPs (millions) taken from thousands of individuals where the outcome is a particular phenotype (like a trait or having a disease) and the variables or features are the genotypes (Kooperberg, LeBlanc, and Obenchain 2010); (Kruppa, Ziegler, and König 2012). The problem is even worse when looking for epistasis between genotypes, interaction effects between them, which makes detecting association a harder challenge and which aggravates the problem of multiple-hypothesis testing correction.

### 1.3.3. Kernel and Ensemble Models Review

As explained in subsection 1.2.5, GWAS are designed to detect common SNPs. Thus, the variation between affected and non-affected samples can be contrasted, concerning a specific gene which is associated with that disease (Hirschhorn and Daly 2005). This is addressed mostly in high-dimensional datasets by employing statistical, ML and computational techniques. It has been showed that single associated SNPs do not have a strong effect which contributes to the risk of disease (low effect size), so much effort has been recently done to focus on studying combined effect of multiple disease-associated SNPs or interaction effects on the risk of disease. In fact, the combination of several genes working together, which usually interact between them, has relevance in complex diseases (Cordell 2009). Hence, these investigations aim to detect different interaction relations among genes with some environmental factors, which may
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

increase the risk of developing the diseases.

As gene-gene interactions may not be linear, kernel and ensemble methods play an important role in their detection. Kernel algorithms such as support vector machines (SVM) have reached an outstanding importance to address nonlinear association between variables (predictor and response) in supervised learning situations (Wang et al. 2015).

Although nonlinear relations are also tested in regression and classification it is difficult to identify before analyzing a functional relation between predictors and response, mostly in terms of multivariate predictors. A very important feature for this case is the kernel trick (kernel functions can work in a higher-dimensional space, without building its representation, so we can determine in the original space a nonlinear decision boundary by the transformed linear decision boundary in higher dimensions) because this does not require the exact same formula of nonlinearity prior to the analysis of it. In recent years, there has been research applying statistical kernel techniques in order to determine the effects of epistasis in complex diseases. For example, Larson & Schaid (2013) proposed a kernel regression method based on generalized linear mixed models framework mainly (GLMMs) to detect pair-wise gene-gene interactions when the response is binary using score-based variance component tests. The authors performed a genetic simulation to examine the behaviour of the tests in interaction models, and they compare their approach to other three methods for detecting epistatic models, SNP-SNP logistic regression, principal component (PC) analysis based on logistic regression (PC-LR) and kernel canonical correlation analysis (KCCA). They showed that the epistatic effects with or without main effects were significant even in main effects tests. Their approach outperformed the other models in detecting interaction model with main effects (Larson and Schaid 2013).

Ensembles algorithms are defined by collections or “ensembles” of base learners,
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

which can be recursively partitioned trees, regression models, etc. Base learners should be able to classify better than coin-tossing (50%) on average. So, given a training dataset D (based on \((X_1, Y_1), \ldots, (X_n, Y_n)\), \(X\) is the matrix of predictors which have \(n\) observations and \(y\) the outcome with \(n\) observations) ensemble learning algorithms estimate the function \(f\) which better relates \(X\) and \(Y\) having a base procedure or base learner. For instance, a classification tree or a regression tree. The base learner can be run several times \((b \in \{1, \ldots, B\})\) from different input data (reweighted original data) in order to have different \(f\) estimations \((f_1, f_2, \ldots, f_B)\), then linear combinations of each individual estimation are considered to build an ensemble based function, such as the average (Bühlmann 2012).

Individual trees are unstable (explained in next subsection), but regression stable. The use of multiple trees improves stability and potentially reduce the variance without increasing the bias of the predicted values (Dietterich 2000a). Ensemble algorithms have been proved to be efficient methods (Dietterich 2000b). Ensemble algorithms have become a primary technique for SNP-SNP interaction identification (Zhang and Bonney, 2000); (Huang et al., 2004). For analyzing these epistatic effects, RF (Breiman, 2001) have become a primary tool (Cordell, 2009). RF is explained in more detail in a later subsection.

Other ML algorithm which is also ensemble-based is the gradient boosting machines. Gradient boosting machine uses a regression function that minimizes some loss function, in case-control studies it is the deviance, which is similar in concept to minimizing squared error in a linear regression. To minimize the loss function, the algorithm uses a stagewise expansion trees (other learners could be used). The cost for misclassifying an observation is updated in each iteration, and the cost on previously misclassified observations is up-weighted (Friedman 2000).

Through the application of an ensemble method, in particular RF, this thesis have the objective of finding subsets of markers which can reveal possible causal mechanisms.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

and causal variants for complex disease.

1.3.3.1. Classification and Regression Trees

In order to interpret complex patterns in high-dimensional data, Breiman et al. (1984) deployed the Classification and Regression Tree (CART) method.

Suppose there is a given training data:

\[ D = \{ y_i, x_{i1}, x_{i2}, \ldots, x_{in} \\mid i = 1, \ldots, n \} = \{(Y, X_j)\mid j = 1, \ldots, N\} \]

where \( y_i \) is the ith observation of the outcome, response variable; \( x_{ij} \) is the value at the observation ith of feature j; \( X_j \) is the vector constituted by all observations of the feature j, N is the number of features or predictors; thus, n is the total number of observations. Using D, the aim is to create a function which makes the best predictions of \( y \) given \( X_j \), \( y=f(x,\theta) \) where \( \theta \) is the function's parameter set. When \( y \) is categorical the model aims to find the discrete category of a new observation which is called classification, and regression when \( y \) is continuous.

A tree-based algorithm creates a classification tree using the predictors. The classification tree (Breiman et al., 1984) is built by repetitively partitioning the data D into subsets of observations which are more homogeneous. The variable with most discrimination score is chosen to divide the dataset into subsets depending on the splitting rule, and partitioning is recursive until the data at one node cannot improve the discrimination or another stopping criterion is met, such as the sample size of each node must be larger than \( N \) or and prune the tree process, in order to predict efficiently and avoid overfitting.

The output is a tree model with the respective branches defined by the splitting rules and the response frequency at the nodes. Formally, a tree model with T terminal nodes is as follows:
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

\[ \hat{y} = f(x) = \sum_{m=1}^{T} k_m I_{\hat{R}_m}(x) \]

where I is the indicator function, if x is the region \( \hat{R}_m \) I is equals to 1, otherwise is 0. Each tree divides the input space in independent regions, the model can be defined by the sum of all regions.

To quantify the error prediction, trees can use different loss functions, the most used being the mean squared error and the impurity or information gain, in regression and classification trees respectively. The tree-building starts calculating a score with all variables in a single region R. Then, each split rule \( s_i \) is based on the Boolean operator OR, like having the genotype AA OR Aa/aa, is tested on each variable for partitioning R into the left and the right regions, Rl and Rr, and the scores of each side region are calculated, \( e(Rl) \) and \( e(Rr) \). The improvement score at each \( s_i \) is considered as the decrease in overall error:

\[ \hat{I}(x_i, s_i) = \hat{e}(R) - \hat{e}(Rl) - \hat{e}(Rr) \]

The model selects the variables and the region with the best fit improvement recursively until the variables of one node are homogenous, in other words, cannot reduce the impurity function I. Also, the minimum node size, the number of terminal nodes and the maximum node size can be also specified as stopping rules.

It is important to mention that the most common split criterion to account for the decrease in the node impurity is the Gini index (Breiman 1993; Zhang and Singer 1999; Sutton 2005). The Gini index can be defined by

\[ \text{Gini index} = 1 - \sum_{j=1}^{N} P_j^2 \]

where \( P_j \) is the relative proportion of the categorical label j in a node.
In GWAS CART trees seek to predict both classes, cases and controls, but Gini impurity gives them an equal importance to misclassification rates. Due to greedy search strategy, small sample fluctuations can result in a high variance, which reduce the CART predictive ability (Breiman et al., 1984). This problem is aggravated in high-dimensional datasets where data are noiser and predictors have less information leading to overfitting. Moreover, very deep trees without a tree size stopping rule can lead to an inefficient prediction as an error in upper splits is propagated and has an impact in all splits.

As the size of tree might be difficult to determine, estimations have shown good misclassification predictions by dividing the original dataset into training and independent test samples (Sutton 2005) or, in small datasets, by using cross-validation (Breiman et al., 1984).

1.3.3.2. Random Forest

Tree-based ensembles algorithm combine many trees which leads to better predictions than with single CART trees. Breiman (2001) developed an ensemble algorithm called Random Forest (RF) which solves the large data problem by modelling many classification trees based on bootstrap subsamples and selecting random predictor variables to build single trees and at the end average all multiple trees (Breiman 2001). The bootstrap subsamples can be with replacement (bootstrapping) or without replacement (subsampling), bootstrapping was used on the original study (Breiman 2001). However, RF showed to be biased when using bootstrapping even under the null hypothesis, while using subsampling RF is reliable as it is unbiased (Strobl et al. 2007b). So, when applying RF in real applications subsampling should be used.

First, RF randomly divides the training dataset D into two independent sets called in-bag data and “out-of-bag” OOB data (the in-bag usually includes the 63.2% of random observations). Second, a specific number of random variables are selected which is
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

called mtry ($X_{mtry}$ C X, X_mtryN predictors matrix) in each split of the tree, then a classification tree $f_b$ is built using the random in-bag sample and a subset of random variables in each split ($X_{mtry}$), the tree $f_b$ is grown until the stopping rules are fulfilled. Once the $f_b$ tree structure is built, RF takes the independent OOB observations of the selected variables and applies the estimated tree to these observations to obtain a prediction ($f_b(X_{mtry})$). Then, RF permutes the variables on the OOB observations, losing the actual association with the outcome, and takes the “null” prediction of the node ($f_b(X_{mtry}^*)$), the error rate at that individual tree is extracted as the difference between both predictions ($f_b(X_{mtry}) - f_b(X_{mtry}^*)$). Finally, RF builds a large number (B) of trees ($b \in \{1, \ldots, B\}$) following the same strategy to finally average the outcomes from all the forests or accumulate the impurity reduction. This prediction rate of each variable is a way to measure the importance of each variable, which allows the model to detect the most predictive variables. Thus, the variable importance indicates how much overall the original association improves the prediction over the “null” one; variables with the highest values correspond to the most relevant variables, and they can be calculated to detect the smallest set of predictor variables to ensure a good prediction performance (Strobl, Malley, and Tutz 2009); (Hastie, Tibshirani, and Friedman 2009).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 1.1. Illustration of the process of RF algorithm. The exemplificative SNPs are annotated as in dbSNP database, reference SNP ID number. In the third step the split criteria is based on whether the individual carries the risk allele at the SNP chosen in that node. Note that mtry is resampled at each node.

RF is one of the algorithms that allows variables to have more than two labels, so in genetics variables do not need to be transformed and the original form of genotypes can be remain as AA, Aa and aa, coded as 0,1,2. Also, in order to avoid overfitting when modelling the classic CART trees algorithms, it is necessary to prune the trees, nevertheless, in RF pruning is not required as the OOB observations are not used to fit the trees, their predictions are considered as accuracy estimations. The optimal mtry and the number of classification trees in the forest are not estimated from the data and need to be set up by the users as well as estimated by cross-validation.

RF (Breiman, 2001) have been used for analyzing gene-gene interactions (Cordell, 2009) due to its ability to analyze several SNPs together in a nonlinear approach (McKinney et al., 2006). RF have been also very useful when identifying disease
associated SNPs because of their use as a filter (Bureau et al., 2003). Furthermore, there has been research on RF behavior under correlation conditions. For instance, Nicodemus & Malley (2009) examined the RF, CIF and MCLR variables importance measure (VIM) in a case-control simulation study including correlated predictors. The results of the study showed that CIF and MCLR outperformed RF in detecting the association of the “causal” variables when these ones were correlated with other variables at effect sizes found in complex studies diseases. They also showed that RF based on permutation VIMs had a better behaviour than RF based on Gini index (Nicodemus and Malley 2009). In addition, Nicodemus et al. (2010c) studied the performance of the permutation variable importance measures (PVIMs) in RF and CIF using synthetic data with correlated and uncorrelated variables. The authors showed that at the first split CIF and RF based on the unscaled and scaled PVIMs selected more frequently the correlated predictors than the uncorrelated ones. But, across all splits under the null and under the alternative hypotheses, the models selected slightly more the uncorrelated predictors, with the exception of unscaled PVIMs under the null that showed very small inflation for the correlated predictors. Moreover, unscaled PVIMs outperformed scaled PVIMs under predictor correlation. The study suggested that RF is more suitable than CIF to apply in high-dimensional studies such as GWAS (Nicodemus et al. 2010c). The next year, Nicodemus (2011) performed a case-control simulation study with uncorrelated binary variables to examine whether the VIM stability and rankings were affected by differences in category frequencies comparing the mean decrease accuracy (MDA) and the mean decrease Gini (MDG) measures. The study suggested that MDA measure is less sensitive to category frequencies than MDG. Furthermore, the author performed a genetic case-control study investigating the stability of ranking in presence of correlated predictors, which showed that MDG were less stable than MDA when the correlation between predictor is strong, and that MDG might be less suitable to apply under correlation conditions (Nicodemus 2011).

In 2009, RF was compared to the CART and to the logistic regression in a simulation study considering 99 different situations depending on missing data, sample size,
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

minor allele frequencies amongst others and in models involving interactions with and without marginal effects (García-Magariños et al. 2009). The study suggested that RF outperforms CART and LR when detecting interactions models mainly when the model is without main effects. Moreover, Schwarz et al (2010) developed a software package that fast-implements the RF algorithm called Random Jungle (Schwarz et al. 2010). In their study they showed that the software outperforms computationally other RF-implementations and they applied RF to detect associations from 275,153 SNPs (single and interactions) in a Crohn’s disease case-control (501 cases / 505 controls) study, where they found significant associated SNPs and SNP-SNP interactions (Schwarz, König and Ziegler, 2010). Recently, Wright et al (2016) studied whether RF based on different VIMs was able to detect interaction effects with and without marginal effects in a genetic simulation study (Wright et al. 2016). The results of the study showed that RF are able to detect SNP-SNP interactions, moreover, RF based on Gini index was more able to detect interactions than the permutation VIMs, and also that VIMs had a better performance capturing models with only main effects than with only interaction effects (Wright et al. 2016).

Because of the efficiency of RF in detecting genetic factors, it has been studied its performance under conditions that may be present in real situations which could affect the behaviour of the model and, therefore, result in spurious results, with poor prediction ability. In 2007 Strobl et al. found a bias in the Gini importance in RF in a simulation study when the predictors have different categories, preferring the ones with more classes. The study also showed that subsampling should considered rather than bootstrap (Strobl et al. 2007b). Moreover, another study (Archer and Kimes 2008) performed a simulation to examine the capability of RF in detecting the true variable among 800 variables (continuous) with 20 different strengths of correlation between 0 and 0.95 in increments of 0.05, where the association is with a binary outcome, having similar conditions to microarray studies. The results of the study showed that RF is a useful algorithm to capture single predictor variables even under correlation conditions and is unbiased in producing classifiers. Therefore, the authors suggested the use of RF in microarray studies (Archer and Kimes 2008). In 2008, a conditional permutation
VIM was proposed as an alternative of the permutation unscaled VIMs when variables are correlated, as permutation VIMs are affected by prediction correlation when a "causal" predictor is correlated with other variables (Strobl et al. 2008). Moreover, Meng et al (2009) proposed alternative models of RF to cope with SNPs in LD, the results of their study suggested that the modified RF by building the tree considering SNPs that are not in LD may be more suitable to apply when there exist SNPs in LD (Meng et al. 2009). Nicodemus et al. (2011) investigated the performance of RF also when predictors have different categories. The authors suggested that SNPs with minor allele frequencies are preferred by the Gini VIM (Nicodemus 2011).

1.3.3.3. Support Vector Machines

One extensively studied kernel technique is SVM, which is a supervised ML technique that assigns classes to objects (Boser et al. 1992). The algorithm was developed by Vapnik and has been widely used for both classification and regression as well as density estimation. SVMs try to find a hyperplane $w \cdot x + b = 0$ ; $x_i \in R^n$. The hyperplane separates the points $x_i$, in order to have all $x_i$ from the same class or label in the same plane side, that fulfilling to $g(x) = sign(w \cdot x + b)$, a decision rule. SVMs selects the hyperplane $w \cdot x + b = 0$ that best separates the points with different classes. Then, the hyperplane or support vector $w$ should minimize the risk of misclassifying a new data point if it is far from the observations; therefore, SVMs maximize the distance from the hyperplane or support vector to the closest $x_i$ (Sweilam et al. 2010).

The algorithm have been performed in genetic studies to classified gene expression profiles in cancer (Guyon et al. 2002). The authors propose a gene selection approach using Recursive Feature Elimination (RFE) in Support Vector Machine to classify genes for cancer diagnosis (leukaemia and colon) and discover drugs in a case-control gene expression study. They showed that the genes selected by their approaches are significant related with cancer in biological terms and the classification had more power than the baseline method (Guyon et al. 2002).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

1.3.4. Use in Genetics

ML is used as a primary tool to detect interactions between genes due to the limits when employing classical statistical techniques which may overfit when analysing big sets of data as well as presenting problems in finding such gene-gene interactions (Koo et al. 2013). Recently, Lu et al (2014) evaluated the performance of several ML algorithms, both supervised (SVM, penalized regression with different penalties, and permanental classification) and unsupervised (sparse graphical models and spare PC analysis), in finding common and rare variants using SNP data from Generic Analysis Workshop18, blood pressure traits and rare variants determined by imputation and sequence analyses. The authors examined the different models in two simulations and four real studies. The results of the study suggested that supervised and unsupervised methods outperform classical statistical techniques (Lu et al. 2014). Therefore, there has been an increase in the use of both techniques (supervised and unsupervised) in the bioinformatics field (Bhaskar et al. 2006).

Moreover, ML has potentially improved the detection of gene-gene interactions. Even though we do not consider Neural Networks in the present study, they have also been used to detect epistasis (Ritchie et al. 2007) (Motsinger-Reif et al. 2008). RF has been widely used to detect epistasis as explained in subsection 1.3.3.2. As an extra example, a study in 2004 (Lunetta et al. 2004) performed a simulation experiment to check the behaviour of RF in detecting SNP-SNP interactions in GWA studies when interactions are present. The study showed that RF has more power to detect interaction effects than the Fisher’s exact test (Lunetta et al. 2004).

1.4. Study Goals

The primary goal of my thesis is to examine RF based on different VIMs and minimal depth under predictor correlation considering three different strengths of correlation (10%, 40%, 80%) and three different number of correlated predictors (5, 20, 40) when
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

both the predictors and the outcome are continuous in four association scenarios in a simulation study. First, I study their behaviour when a single predictor (correlated with other predictors) is associated with the outcome with both weak and strong effects. Second, I study their ability in detecting weak and strong 2-way interaction models with main effects (one predictor is correlated with other predictors and the other one is uncorrelated). In addition, I investigate the performance of RF based on Gini importance in a simulation study considering independent normal distributed continuous variables and both a continuous outcome and a binary outcome, when the variance of the variables is different, when the precision of the variables is different, and when the error variance has different variance. In all situations the effect size is the same.

Although there are other machine learning techniques that have shown to be suitable for detecting interactions as explained in subsection 1.3.1.3, this study is not aimed on investigating the ability of different ML techniques in detecting single and interaction effect and make a comparison between them. Instead, the study is focused on the comparison of different RF VIMs in order to make conclusions about which should be used on real applications when using RF.

The aims of the present study were: 1) to analyse the performance of RF based on different variable important measures (VIMs) so as to identify which one has the best performance in situations where variables are correlated and present a weak association with the outcome under study such as in psychiatric genetics. 2) the Gini variable importance measure (VIM) is widely used even though it has shown to prefer predictors with more categories. To date, there have not been research on how Gini VIM performs when having continuous predictors with different variances, but there was research considering binary ones (Nicodemus 2011). Hence, the second chapter of this Thesis attempts to investigate whether Gini VIM is affected by the amount of variance of the predictors and by the precision of the predictors having the same effect on the outcome, as well as by the error variance. 3) To use the results from Chapter 1
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

to apply the most powerful RF VIM to a psychosis case-control study as well as to a cognition study considering both IQ and verbal IQ in order to detect interactions between two and three different markers. We studied epistatic effect among human genes that have been related with abnormal/affect behaviour in mouse models.

**Hypothesis:** The ability of RF to detect main effects when predictors are correlated has been investigated as discussed in the Introduction, but the studies performed mainly simulations where the association was with binary outcomes. In addition, its performance in detecting interaction models under predictor correlation has been less studied and there has not been so much research considering both continuous predictors and continuous outcomes. This study seeks to test which VIMs are more powerful in continuous data and check that RF is still appropriate to apply when dealing with continuous outcomes.

As discussed in the present Introduction, Gini importance has been shown to be biased when the predictors had different number of categories. Considering continuous variables and both continuous and binary outcomes, I expect to see a similar behaviour of Gini importance, an inflation when the variance of the variable is higher as well as an inflation when the number of cut-points is higher (higher precision).

Psychotic disorders have an oligogenic aetiology, complex models such as epistasis and PRSs may identify associations if the main effect from single genes does not have a significant contribution on the disease. In addition, PRSs have not explained much variation of the diseases in several studies. Therefore, this study will try to detect interaction effects which may demonstrate more variance explained of psychosis than polygenic effects, and gives us a simple interpretation of the biological system of the disease.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

2. Performance of variable importance measures in Random Forest under correlation and application in PGC2

2.1. Introduction

2.1.1. Previous studies

Over the last decade, machine ML algorithms have increasingly been used in different backgrounds such as genetics, neuroscience, and finance (Jordan and Mitchell 2015); (Patel et al. 2015); (Libbrecht and Noble 2015). In genetics, with the introduction of increasingly large GWAS, these techniques have become necessary due to the high dimensionality of data (Kooperberg et al. 2010). The challenge of managing big data with more variables than observations makes ML attractive to researchers (Kruppa et al. 2012).

RF is a supervised ML technique, which measures the importance of each variable associated with an outcome (Breiman 2001). There are different ways to measure that importance; in other words, there are different VIMs: the Gini variable importance (Breiman et al., 1984), permutation (Breiman 2001), scaled, conditional (Strobl et al. 2008), minimal depth (Ishwaran et al. 2008); (Ishwaran et al. 2010), and area under the curve VIM (AUC VIM) (Janitza et al. 2013). In this study, the performance of this algorithm with regard to its ability under correlation conditions was examined.

2.1.2. Why the study is needed?

During last decade, there has been research on which VIM has the best performance for detecting true positives instead of false positives, when looking at main effects (Strobl et al. 2007b); (Nicodemus and Malley 2009); (Nicodemus et al. 2010c); (Calle and Urrea 2011); (Nicodemus 2011). Also, the capability of RF VIMs to detect interaction effects has also been studied (Yang et al. 2010); (Goldstein et al. 2011); (Boulesteix et al. 2012b); (Boulesteix et al. 2015); (Wright et al. 2016).
An interaction effect happens when the effect of one variable depends on other variable, or in other words, a value of one variable changes the effect of the other variables and vice versa. For instance, the effect that SNP1 has on disease depends on the values a SNP2, in this way SNP2 modifies the phenotype of SNP1. However, a different term of interaction is conditional dependence. Conditional dependence happens when the association between two predictors depends on the values of a third predictor, in this way, the third predictor does not affect the association between a response variable or outcome and a predictor, but it affects the association between the other two predictors. For example, the association between SNP1 and SNP2 depends on the values or the number of risk alleles of SNP3.

RF is supposed to measure interactions due to its natural architecture in recursive trees, which provide certain dependency in a hierarchical way through the forest (Breiman 2001). Also, Boulesteix et al. (2015) suggested that an interaction effect can be detected when the tree growing process stops on one side and continues on the other, when the effects of the two child nodes are different but the variable selected is the same, or when the variables selected in a split are different on both sides; in other words, it is feasible that interactions exist between variables if, after the split, the two branches behave differently. In addition, RF VIMs are supposed to be able to detect interaction effects (García-Magariños et al. 2009).

Nevertheless, a recent study (Wright, Ziegler and König, 2016) showed the difficulty of using RF to capture interaction effects which do not include the marginal or main effects; in fact, with the natural construction of RF, it is not easy to distinguish between interaction, marginal or chance fluctuations under H0. Furthermore, a previous study claimed that this method may not be able to capture interactions in Big Data without having strong marginal effects (Winham et al. 2012). However, in 2004 Lunetta et al. (2004) showed the ability of RF to detect SNP interactions, and suggested that RF performance would be better than the Fisher Exact test for interactions. Furthermore, a year later, another study agreed with the previous one, showing that SNP pairs were the ones with the highest importance (Bureau et al. 2005). Thus, the performance of
the different RF VIMs in presence of interactions effects has been investigated, and they showed contradictory results.

As has been explained in the Introduction, psychiatric disorders are not Mendelian where genetic factors are correlated and have a low effect on the disease. Hence, it is crucial to research which RF VIM can deal with such situations, as well as cover more complex variants such as interactions.

2.1.3. Aim

The first part of this research was focused on testing which VIM was the best to be applied in the area of psychiatric genetics, to detect single effects where variables (SNPs) are weakly associated with the outcome, and are also correlated with each other due to LD. This study is the first to evaluate several VIMs in combination, including AUC VIM and the permutation conditional VIM, and to compare their behaviour with others. These two VIMs were investigated in their proposed study comparing their behaviour with the unscaled PVIM. One of the most recent studies compares maximal subtrees with other VIMs considering two values of the mtry (number of variables randomly chosen to be part of the pool of variables to be selected to split the tree) but only using the joint importance to detect interactions, here I compared minimal depth with two different mtry to see if that affects minimal depth’s behaviour under correlation conditions when having also single effects. Furthermore, the second aim of my study was to perform a simulation, also covering correlation conditions between variables, to discover the capacity of different RF VIMs when capturing an interaction signal.

Implementations of RF with different VIMs were applied to a simulation study, and their behaviour was examined in the following ways: detecting single and interaction effects using continuous (N(0,1)) data simulated under H₀ and two conditions under Hₐ: weakly and strongly associated. The simulation study considered variations in the (a) number of correlated predictors and (b) strength of correlation between predictors.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

One of the most powerful VIM from the simulation study was applied to a schizophrenia GWAS case-control study examining a genetic pathway based on 29 molecular biomarkers (Chan et al. 2015), with the aim of finding single and interaction SNP effects.

2.2. Methods

2.2.1. Random Forest

RF builds a large collection of recursively-partitioned trees. Specifically, RF seeks to improve the variance reduction of bagging (Bootstrap Aggregation, a succession of trees taking a bootstrap sample of the data including all variables (Breiman 1996)) by reducing the correlation between the trees, without increasing the variance. This is achieved in the tree-growing process through random selection of the input variables and observations (Tin Kam Ho 1998). A subset from the original sample, for example, 63.2% of observations, is randomly selected to build each tree. These are called the “in bag” samples, and the Gini VIM is based on these observations and resulting tree (Breiman 2001). The remaining 36.8% observations are called OOB observations for that tree, which are used to estimate error and variable importance for permutation VIMs (Breiman 2001). Hence, RF is an ensemble consisting of multiple classification or regression trees that are grown using a subsample of given data randomly chosen in each split and without pruning (Figure 2.1) (Breiman 2001). One of the attractive features of RF is that the importance of each predictor variable can be estimated and can be used for ranking variables for high-dimensional data settings such as gene-gene relationships (Winham et al. 2012).

In this study, I included a new extension of RF originally designed for use with right-censored survival data (Ishwaran et al., 2008), Random Survival Forest (RSF), which also may be applied to binary or continuous outcomes. Here, I applied RSF to non-survival data to study its performance in continuous situations as well as the AUC and conditional permutation variables importance measures. The AUC PVIM computes
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

the area under the ROC curve (probability of detection true signals against the probability of false positives, measures the true positive rate as a function of false positive rate) before and after permutation instead the error rate (Janitz et al. 2013). The conditional permutation variable importance measures the difference between the error rate before and after permuting the predictor but taking into account the correlation pattern between predictors, permuting in different sets of a correlation grid (Strobl et al. 2008).

Figure 2.1. Illustration of the process of RF algorithm. The exemplificative SNPs are annotated as in dbSNP database, reference SNP ID number. In the third step the split criteria is based on whether the individual carries the risk allele at the SNP chosen in that node. Note that mtry is resampled at each node.
2.2.1.1. Variable Importance Measures

The two fundamentally different VIMs in RF are the Gini importance (VIM$_{\text{Gini-RF}}$) and the permutation importance (PVIM) (Breiman 2001). In the VIM$_{\text{Gini-RF}}$, at each split in each tree $b$, with $b \in \{1, \ldots, \text{ntree}\}$, the improvement in the split-criterion is the importance measure attributed to the splitting variable, and is accumulated over all the trees in the forest separately for each variable $i$, with $i \in \{1, \ldots, N\}$ and where $N$ is the total number of predictors. In the permutation-based VIMs, RF also uses OOB samples to measure the predictive ability of each variable. When the $b^{th}$ tree is grown, the OOB samples are passed down the tree, and the prediction accuracy is recorded at each split. Then the values for the $i^{th}$ variable are randomly permuted in the OOB samples, and the accuracy is again computed. The decrease in accuracy as a result of this permutation is averaged over all trees, and is used as a measure of the importance of variable $i$ in the RF, which is called PVIM (Figure 2.1).

The VIM$_{\text{Gini-RF}}$ (computed in the in-bag sample) of a predictor variable $X_i$ is the total decrease in impurity $\Delta I$, where the reduction in impurity is given by

$$\Delta i_k = i(k) - p(k_l)i(k_l) - p(k_r)i(k_r)$$

where $i(k)$ is the impurity in the node $k$, and $p(k_l)$ and $p(k_r)$ are the probabilities that the variable falls in either the left node $k_l$ or the right node $k_r$, respectively (Breiman et al., 1984; Zhang, 1999). $i(k)$ is measured in Gini impurity in binary outcomes and for the continuous outcomes is typically the mean residual squares. Thus,

$$\Delta I = \sum_k \Delta i_k$$

The standard PVIM, VIM$_{\text{rawperm-RF}}$, as was explained above, is based on the OOB error rate since it is the difference in the mean OOB error rate before and after permuting the values of the predictor $X_i$, i.e:
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

\[ V_{IX_i}^{ER} = \frac{1}{|T|} \sum_{t \in T} A_t - A_t^* \]

where \( T \) is the size of the forest, \( t \) is each tree, \(|.|\) is the number of elements in a set, \( A_t \) and \( A_t^* \) are the prediction accuracy before and after permuting the values of \( X_i \) respectively.

An alternative permutation-based VIM is based on a modification of the forest by creating each tree with only uncorrelated variables (Meng et al. 2009) proposed by Meng, but in this study I considered as the Meng VIM (VIM\textsuperscript{Mengperm-RF}) the one implemented in random jungle (Schwarz, König and Ziegler, 2010), which is not the same one that Meng proposed. In random jungle, this VIM is the same as the VIM\textsuperscript{rawperm-RF}, except the average is taken across all the trees in the forest containing that predictor instead of across all trees, hence the VIM\textsuperscript{Mengperm-RF} is:

\[ V_{IX_i}^{ER} = \frac{1}{|T_{X_i}|} \sum_{t \in T_{X_i}} A_t - A_t^* \]

where \( T_{X_i} \) the total number of trees in which the variable \( X_i \) appears. The results from this VIM are not shown due to they are virtually identical to the VIM\textsuperscript{rawperm-RF}, as expected. Other studies have proposed different VIMs based on the VIM\textsuperscript{rawperm-RF}, such as scaled PVIMs, which divide the PVIM by its empirical standard error over all the trees in the forest. More precisely,

\[ V_{IX_i}^{SER} = \frac{V_{IX_i}^{ER}}{S^2 \sqrt{|T|}} \]

I compared two scaled PVIMs, Breiman’s (VIM\textsuperscript{Breiperm-RF}) and Liaw’s (VIM\textsuperscript{Liawperm-RF}), the difference between them is the variance estimator. The estimator of VIM\textsuperscript{Breiperm-RF} is defined as
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

\[ s^2 = \frac{1}{T} \sum_{t \in T} N_{OOB,t} (A_t - A_t^*)^2 - (VI_{X_i}^{ER})^2 \]

where \( N_{OOB,t} \) is the number of samples in OOB of the tree \( t \). And the estimator of \( \text{VIM}_{\text{Liawperm-RF}} \) is

\[ s^2 = \frac{1}{T} \sum_{t \in T} (A_t - A_t^*)^2 - (VI_{X_i}^{ER})^2 \]

Hence, the difference between them is that \( \text{VIM}_{\text{Breiperm-RF}} \) takes into account the observations in OOB in tree, ensuring as much variability in the individual trees as possible.

A novel VIM was recently proposed to account for correlation between variables, called the conditional PVIM (\( \text{VIM}_{\text{rawperm-CF}} \)) which is the same as the \( \text{VIM}_{\text{rawperm-RF}} \) but conditioned on \( r \), a value of correlation between the predictor of interest and all other predictors in the matrix (Strobl et al. 2008). So, this PVIM differs from \( \text{VIM}_{\text{rawperm-RF}} \) in \( A_t^* \), in the way to calculate the OOB prediction accuracy after permutation. Here, the aim is to conditionally permute the values of \( X_i \) in groups of \( Z \), observations which do not break the pattern of correlation between the variable \( X_i \) and the others. For that, before calculating \( A_t^* \), it extracts, for all \( Z \) to be conditioned on, the cutpoints that split this variable in the current tree and create a grid by means of bisecting the sample space in each cutpoint. Hence, the OOB prediction after permutation within the grid defined by the variables \( Z \) is called \( A_t|Z^* \), and the VIM of variable \( X_i \) is derived as follows:

\[ VI_{X_i|Z}^{ER} = \frac{1}{|T|} \sum_{t \in T} A_t - A_t|Z^* \]

The \( \text{VIM}_{\text{rawperm-CF}} \) takes into account the correlation between the permuted variables and the other predictors and just permute \( X_i \) within groups which have some dependency structure with \( X_i \) (Strobl et al. 2008).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

The VIM\textsubscript{AUC} computes the ROC area under the curve (AUC), for each predictor after and before permuting a predictor instead of the prediction accuracy, which is used in the VIM\textsubscript{rawperm-RF} where the AUC is the probability to detect a random true value as a true positive rather than a false positive (Janitza \textit{et al.} 2013). It is defined as:

\[ VI_{x_i}^{AUC} = \frac{1}{|T|} \sum_{t \in T} AUC_t - AUC_t^* \]

where \( AUC_t \) and \( AUC_t^* \) denotes the AUC computed from the OOB observations in the tree \( t \) before and after randomly permuting predictor \( X_i \), respectively (Janitza \textit{et al.} 2013).

Novel VIMs from RSF are based on a tree concept referred to as "minimal depth" which measures the variable importance in terms of its splitting behaviour relative to the root node, \textit{i.e.}, the variables that split close to the root node will have a stronger effect on the outcome. The minimal depth is directly associated with the maximal subtree and can be explained precisely in terms of that. The definition is presented as follows:

For each variable \( X_i \), call \( T_{X_i} \) an \( X_i \)-subtree of \( T \) if the root node of \( T_{X_i} \) is split using \( X_i \). Call \( T_{X_i} \) a maximal \( X_i \)-subtree if \( T_{X_i} \) is not a subtree of a larger \( X_i \)-subtree (Ishwaran \textit{et al.} 2010).

The shortest distance from the root of the tree to the root of the closest maximal subtree of \( X_i \) is the minimal depth of \( X_i \) (Figure 2.2). Smaller values of minimal depth imply stronger association of the variable with the outcome.

Maximal subtrees can be easily applied to all ensembles of trees, without depending on the type of the outcome. Hence, they can be applied to popular applications like regression (continuous data) and classification (binary data). So, maximal subtrees can be used in place of (or in addition to) VIMs. One advantage to using minimal depth is that it is independent of the way prediction error is measured (Ishwaran \textit{et al.} 2010).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

is important to clarify that larger minimal depth implies less association with the outcome. In contrast, larger VIM values indicate stronger importance.

Figure 2.2. Illustration of RF based in minimal depth. As an example, it shows the minimal depth of V2, V4 and V10.

In this study, the package party is used to perform the VIM_{AUC} and the unscaled permutation VIM, called VIM_{party} in this study. This package builds RF based on standard Forest, as well as on unbiased conditional inference trees. CIF uses the Pearson $\chi^2$ test $P$-value corrected for multiplicity, which is unbiased when predictors have different numbers of categories rather than on CART classification trees (Hothorn et al. 2006). Note that this is not the VIM_{rawperm-RF}. Also, note that in the next sections when I report results of these two different unconditional PVIMs, these were applied using CIF and not on the regular RF.
2.2.2. Data Simulation

To perform the data simulations I used R version R-3.0.1. I simulated data where the outcome is only associated with a single predictor, $V_2$; and the additional 99 variables were null data with no association between the predictors and the outcome. In addition, under the same conditions, I performed data simulations where the outcome is associated with the main effects and the interaction of two variables, $V_2$ and $V_{90}$, but $V_2$ is correlated with the others and $V_{90}$ is completely independent of the others.

To deal with the simulations of continuous data, I programmed a function called rmvnormc based on the function rmvnorm in the package mvtnorm (Genz and Bretz 2009) to generate correlated multivariate standard normal data, and independent continuous variables which were randomly generated from a standard normal $N(0,1)$ distribution. The rmvnormc allows to include both the variance matrix and the correlation matrix as input parameters for the generation. I created two different types of continuous data based on how the variable $V_2$ was associated with the outcome, strongly associated or weakly associated, as well as for the interaction study considering, $V_2$, $V_{90}$, and the interaction between them. Furthermore, I created the simulations under the null based in no association. For each synthetic data simulation, I generated 500 replicates of 100 variables, which were distributed standard normal $N(0,1)$, with different number of correlated variables (correlated with $V_2$), 5, 20 and 40, with different strength of correlation $r = 0.1$, 0.4 and 0.8, and remaining variables independent of each other and $V_2$ (Figure 2.3). Thus, I performed 45 different simulations, 36 (Table 2.1) for both the single association and the interaction association studies (alternative hypotheses), and 9 for the null one. The bias and the 95% coverage were also calculated in each different situation to assess the accuracy of the simulations (Table 2.2 and Table 2.3 for the single association, Table 2.5 and Table 2.6 for the interaction association). Furthermore, the correlation of the synthetic data was extracted to confirm that the correlation pattern was consistent (Appendix Table A.2 and Table A.3). The significance of the generating models was calculated using LRTs in a nested model, considering the generating model as the full model, and the
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

model with only the intercept as the reduced one. The package lmtest (Zeileis and Hothorn 2002) was used to extract the LRT p-values.

\[
\begin{array}{cccccccccc}
  r & 80 & 80 & 80 & 40 & 40 & 40 & 10 & 10 & 10 \\
  N & 5 & 20 & 40 & 5 & 20 & 40 & 5 & 20 & 40 \\
\end{array}
\]

Table 2.1. The nine different correlation conditions with the 3 different strengths of correlation (r) and 3 different number of correlated variables (N).

![Diagram](image-url)

Figure 2.3. Generation of the single association study as an example of the data simulation.
2.2.2.1. Simulation under $H_A$

2.2.2.1.1. Single association

I performed the simulations as follows:

$$y = \beta_1 \ast V_2 + \epsilon$$

where $\beta_1$ was set to 1 in the strongly-associated case and to 0.05 in the weakly associated case. The error was set to $\epsilon \in N(0,0.05)$ for the strongly-associated case and $\epsilon \in N(0,0.5)$ for the weakly-associated case.

2.2.2.1.2. Interaction association

The different interaction effects models includes main effects and an interaction effect:

$$y = \beta_1 \ast V_2 + \beta_2 \ast V_{90} + \beta_3 \ast V_2 \ast V_{90} + \epsilon$$

where $\beta_1 = \beta_2 = \beta_3 = 1$ and $\epsilon \sim N(0,0.05)$ in the strongly-associated study and $\beta_1 = \beta_2 = 0.033$ and $\beta_3 = 0.09$, and $\epsilon \sim N(0,0.5)$ in the weakly-associated study.

In this study, we only considered this type of interaction model: main effects and interaction between two variables. One of the limitations of this study is the lack of other types of interaction models, which is discuss in section 5.2.

2.2.2.2. Simulations under $H_0$

The different conditions under the null, using all of the 9 different correlation and number of variables correlated conditions, were generated as the following model:

$$y = 0 \ast V_2 + \epsilon$$
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

So,

\[ y = \varepsilon \]

where \( \varepsilon \sim N(0,0.5) \).

2.2.3. Power and 5% significance cut-off

Under the null, VIM distributions for all variables are expected to be the same and around zero (zero with the exception of \( VIM_{\text{Gini-RF}} \) and minimal depth) (Strobl et al. 2007b); (Boulesteix et al. 2012a). Basically, the medians of all VIMs and minimal depth should be similar or uniform for all the non-influential variables. The non-parametric Wilcoxon test was used to determine if the VIM medians and minimal depth were different between the correlated and the uncorrelated variables.

To study the Power, i.e. to approximate the probability of rejecting the null hypothesis when is not true (detect true signals). First, I had to study the different VIMs under the null, and considering a significance level of 5% (\( \alpha = 0.05 \)), to extract the cut-off at that level for each VIM in each iteration (VIM outputs per null dataset). The scores for each VIM in each iteration were sorted in decreasing order, and the minimal depth (both values of mtry) in increasing order. Then, the fifth (5% of 100 importance score, 100 variables in each dataset) maximum value was extracted and considered as the cut-off for that iteration under \( H_0 \). 500 cut-offs were taken into consideration.

Once the cut-offs were extracted, the VIMs and minimal depth under \( H_A \) were considered. For each RF alternative output (one per 500 datasets), I compared whether the VIM and minimal depth score of the true variable (true signal) was greater than or equal to (lower or equal in minimal depth) all 500 cut-off values divided by 500 (for each real dataset the rate of detecting the signal; number of times the model detects the true signal/500). Finally, I averaged the number of times the true signal was detected across all 500 alternative datasets (the mean of the detecting rate). I studied the power of detecting \( V_2 \) in the single association study as well as the power of capturing \( V_2 \) and
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

V90 in the interaction association study. The power was assessed in two different ways. The first one was considering the null synthetic data generated, and the second way was based on null databases created by permuting the outcome of the corresponding alternative databases, as it is the usual approach in real studies.

2.2.4. Random Forests VIMs simulations

I investigated the performance of RF using different VIMs in conditions simulated under H0 and HA (including weakly and strongly associated conditions when a single predictor V2 is associated with the outcome as well as when an interaction term is involved). To apply the three different versions of RF to these different datasets, I used randomjungle Centos 64 Bit Version (Build 2.0.0) (Schwarz, König and Ziegler, 2010), which is an implementation of RF; and to assess the VIM_AUC and the VIM_party used the R package party version 1.0-18 (Hothorn et al. 2015), which calculates standard and conditional forest for cforest. The R package randomForesSRC version 4.6-12 (RSF) (Ishwaran and Kogalur 2007) was used to carry out analyses using RSF. The package allows us to apply RF for survival, regression and classification as well as extract the maximal subtree information to use as a VIM.

For all implementations of RF used to calculate the VIMs, I chose samples without replacement and I set the number of trees to ntree = 1000 and, the mtry equal to 39, which is the size of randomly chosen variable sets at each split. This was a slightly larger than the default N/3=33.33 (where N is the number of variables), because the default mtry is not optimal when there are many noise predictors (Segal 2004). For the conditional VIM the Pearson's correlation coefficient cut-off was set to 0.75 when the correlation simulated between the predictors was 0.8, and to 0.35 when the correlation between predictors was 0.4, and when the simulated correlation between predictors was 0.10 the cut-off was set to 0.05. Minimal depth was performed considering two different situations mtry = 39 and mtry=27 to see how the mtry value effects correlated predictors since a study suggested that mtry should be large in high-dimensional studies (Ishwaran et al. 2010). Moreover, another study showed that minimal depth
performs better with a large mtry for strong associated variables, but a large value could be unfavorable for weak associations (Ishwaran et al. 2011).

2.3. Results

2.3.1. Bias, coverage and correlation of simulated data

I extracted the bias (difference between the expected and the observed value) in order to know if the generated or simulated data, from the different models, were similar to the expected data from models illustrated in subsections 2.2.2.1 and 2.2.2.2. Furthermore, I extracted the 95% coverage, the number of times the true value (for instance $\beta_1 = 1$ (true value) in the strongly-associated case and 0.05 in the weakly-associated case) was contained within the 95% observed confidence intervals (confidence intervals estimated using the general linear regressions from the simulated data) in percentage. Also, I checked the number of p-values less than the Bonferroni corrected p-value (as in real situations because of multiple testing) from general linear regressions, to ensure the simulations were generated correctly and if $V_2$, the true signal, was detected using the same regression generating models in the single association study; and the full model in the interaction simulation study. In addition, I calculated the median of the correlation between the variables correlated, and between the variables independent and all others.

First, I checked the simulations under $H_0$, the bias ranged from -0.00071 to 0.0011, and the ninety-five percent coverage was appropriate between 93.2 and 97.4 (Appendix Table A.1). In addition, the correlation pattern between variables was consistent to the real strength when generating the data. The correlation between correlated variables was always around 0.10, 0.40 and 0.80 when the correlation considered was 0.10, 0.40 and 0.80 respectively (Appendix Table A.2), no matter the number of correlated variables. The values taking into account the correlation between independent variables and all others were always around 0 (Appendix Table A.3) indicating independency in the simulated data.
2.3.1.1. Single effect association

As expected, under $H_A$ the bias was minimal, around 0, with a range from -0.00006 to 0.00012 in the strongly-associated study and from -0.00139 to 0.00135 in the weakly-associated one (Table 2.2 and Table 2.3). Furthermore, the ninety-five percent coverage ranged from 91.8% to 96.2% and from 94.4% to 97.6% in the strongly-associated and weakly-associated studies respectively (Table 2.2 and Table 2.3), suggesting that the original coefficients could be reproduced by the linear regression model, as the linear generating model. After Bonferroni correction, V2 was always statistically significant in the strongly-associated continuous condition. In the weakly-associated condition V2 was not always significant, with $p$-value < 0.0001 between 117 to 211 times in all cases (Table 2.4), indicating that the association effect was not strong from the linear generating models.

<table>
<thead>
<tr>
<th>SAC</th>
<th>$r=0.80$</th>
<th>$r=0.40$</th>
<th>$r=0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>BIAS</td>
<td>COV%</td>
<td>BIAS</td>
</tr>
<tr>
<td>5</td>
<td>0.00012</td>
<td>94.8</td>
<td>0.00008</td>
</tr>
<tr>
<td>20</td>
<td>0.000001</td>
<td>91.8</td>
<td>0.000022</td>
</tr>
<tr>
<td>40</td>
<td>0.00007</td>
<td>93.8</td>
<td>0.000068</td>
</tr>
</tbody>
</table>

Table 2.2. Bias and coverage of V2 (associated variable) in the strongly associated study (SAC).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>WAC</th>
<th>r=0.80</th>
<th>r=0.40</th>
<th>r=0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>BIAS</td>
<td>COV%</td>
<td>BIAS</td>
</tr>
<tr>
<td>5</td>
<td>0.00110</td>
<td>95.6</td>
<td>0.00122</td>
</tr>
<tr>
<td>20</td>
<td>0.00102</td>
<td>94.8</td>
<td>-0.00139</td>
</tr>
<tr>
<td>40</td>
<td>-0.00125</td>
<td>94.8</td>
<td>-0.00131</td>
</tr>
</tbody>
</table>

Table 2.3. Bias and coverage of V2 (associated variable) in the weakly associated study (WAC).

<table>
<thead>
<tr>
<th>p-value</th>
<th>r=0.80</th>
<th>r=0.40</th>
<th>r=0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>WAC</td>
<td>183</td>
<td>211</td>
<td>171</td>
</tr>
</tbody>
</table>

Table 2.4. Number of p-values less than 0.0001. WAC mean weakly associated continuous studies.

2.3.1.2. Interaction effect association

As I did with the single association simulations, I extracted the bias and the ninety-five percent coverage for the interactions in the interaction models and in both strongly and weakly association studies (Table 2.5 and Table 2.6). In the strongly-associated study the bias was ranged from -0.00009 to 0.000009, and the ninety-five percent coverage between 93.4% and 95.0%. In the weakly association study, as expected the bias was centred to 0, between -0.00139 and 0.0014, and the coverage around 95%, between 92.6 and 95.8%. This shows that the regression models could reproduce the generating models.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Since studying interaction effects in real data leads to perform many more tests than in single association studies, and in order to find significant contributions, the p-value threshold used should be lower to take account of multiple testing, in Table 2.7 the number of p-values less than the Bonferroni threshold (1x10^{-5}) is shown for each correlation condition. The regression model could detect the effect from the model between 452 and 464 times in the weakly-association study, and in the strongly-associated study always passed Bonferroni correction (LRT tests).

<table>
<thead>
<tr>
<th>SAC</th>
<th>(r=0.80)</th>
<th>(r=0.40)</th>
<th>(r=0.10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>BIAS</td>
<td>COV%</td>
<td>BIAS</td>
</tr>
<tr>
<td>5</td>
<td>-0.00002</td>
<td>94.4</td>
<td>0.00005</td>
</tr>
<tr>
<td>20</td>
<td>-0.00004</td>
<td>94.6</td>
<td>-0.00003</td>
</tr>
<tr>
<td>40</td>
<td>-0.00009</td>
<td>93.4</td>
<td>-0.00007</td>
</tr>
</tbody>
</table>

Table 2.5. Bias and coverage for the interactions on the strongly-associated interaction model.

<table>
<thead>
<tr>
<th>WAC</th>
<th>(r=0.80)</th>
<th>(r=0.40)</th>
<th>(r=0.10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>BIAS</td>
<td>COV%</td>
<td>BIAS</td>
</tr>
<tr>
<td>5</td>
<td>0.00062</td>
<td>92.6</td>
<td>0.00144</td>
</tr>
<tr>
<td>20</td>
<td>-0.00139</td>
<td>93.6</td>
<td>0.00045</td>
</tr>
<tr>
<td>40</td>
<td>0.00043</td>
<td>94.8</td>
<td>-0.00110</td>
</tr>
</tbody>
</table>

Table 2.6. Bias and coverage for the interactions on the weakly-associated interaction model.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

### 2.3.2. Distributions under H₀

Considering a significance threshold of $\alpha = 0.05$, the VIMs cut-offs of rejecting the null hypothesis when it is true were extracted for all the 500 databases under $H₀$ in each correlation condition (these cut-offs were used to test the power in the next section). In order to see if the cut-offs were well-determined at the significance level 5% in all correlation conditions, in each null output the number of VIM and minimal depth scores greater than or equal to all cut-offs were added and divided by the total 500. Then, the average across all 500 null outputs was calculated. Indeed, the values were always around 5%, between 4.85% and 6.29% (Table A.4, Appendix A). With the exception of VIM$_{\text{rawperm-CF}}$ when $r=0.10$ and $N=5$, which was due to the fact that the VIM was always zero.

To determine whether RF based on the different VIMs and minimal depth and, VIM$_{\text{AUC}}$ and VIM$_{\text{party}}$ based on CIF are biased under correlation conditions in no association situations, I examined the different measures under the null hypothesis. If any predictor has considerable more VIM or less minimal depth than others, it would be a bias that has to be considered due to non-associated (noise) predictors can be influential only for the fact of having correlated predictors in the database. In this study, I show that correlation between predictors had an impact on the different VIMs or minimal depth under the null, having different behaviours on different VIMs or

### Performance of variable importance measures in Random Forest under correlation and application in PGC2

<table>
<thead>
<tr>
<th>p-value</th>
<th>r=0.80</th>
<th>r=0.40</th>
<th>r=0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>WAC</td>
<td>459</td>
<td>452</td>
<td>457</td>
</tr>
</tbody>
</table>

Table 2.7. Number of p-values less than p-value threshold $1\times10^{-5}$ on the weakly-associated interaction model study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

minimal depth. In this section, three conditions \((r = 0.10 \text{ and } N = 5, \ r = 0.40 \text{ and } N = 20, \ r = 0.80 \text{ and } N = 40)\) are illustrated in the Figure 2.4, Figure 2.5 and Figure 2.6 (see Appendix A for other correlation conditions; Figure A.1 – A.6). Furthermore, see Appendix A for the median importance scores of the different VIMs and minimal depth under \(H_0\) for correlated and uncorrelated predictors (Table A.5 and Table A.6).

To give some sense to the minimal depth, the median of the depth thresholds were extracted under \(H_0\), which was around 9.9 under all correlation conditions (See Appendix Table A.27.).

![Figure 2.4](image-url)

**Figure 2.4.** RF VIMs, minimal depth, VIMAUC and VIMparty under \(H_0\) for \(V_2\), two variable correlated \(V_3\) and \(V_6\), and two independent variables \(V_{42}\) and \(V_{90}\) when \(r = 0.10\) and \(N = 5\).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 2.5. RF VIMs, minimal depth, VIMAUC and VIMparty under H0 for V2, two variable correlated V3 and V6, and two independent variables V42 and V90 when r = 0.40 and N = 20.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Performance of variable importance measures in Random Forest under correlation and application in PGC2

Figure 2.6. RF VIMs, minimal depth, VIMAUC and VIMparty under H0 for V2, two variable correlated V3 and V6, and two independent variables V42 and V90 when r = 0.80 and N = 40.

As seen in the figures, under predictor correlation the VIMs and minimal depth median scores showed differences between correlated and uncorrelated variables, p-values from Wilcoxon test were extracted to test that difference formally (under predictor correlation conditions, all showed to be less than 0.05). VIM\textsubscript{Gini} and scaled PVIMs were biased under predictor correlation under H0, showing more distance between VIMs for correlated and for uncorrelated predictors as more correlation between predictors, but they behaved in an opposite way (Figure 2.4, Figure 2.5 and Figure
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

2.6). VIM\textsubscript{Gini} showed more inflation for the uncorrelated predictors (Median VIMs difference between correlated and uncorrelated was equal to 0.81 in the extreme condition), as shown previously (Nicodemus and Malley 2009); (Nicodemus 2011). However, the scale PVIMs inflated more the VIMs of the correlated variables with a difference of 5.22 and 0.27 in VIM\textsubscript{Breiman-RF} and VIM\textsubscript{Liawman-RF} respectively in the extreme condition. All VIM medians for correlated predictors were positive, which lead to the inflation of these VIMs for the correlated in comparison to the uncorrelated predictors, as seen by Nicodemus et al. (2010c). In addition, minimal depth showed more inflation for the uncorrelated variables than for the correlated ones with both mtry numbers (Figure 2.4, Figure 2.5 and Figure 2.6; difference between medians 1.04 with mtry =39 and 1.18 mtry =29). Note that this inflation was in terms of negative minimal depth values to compare them to the VIMs, as lower values in minimal depth means a greater association (opposite to the VIMs that larger VIMs means a greater association), using negative minimal depth scores allows to correspond larger values to more association. The inflation was always higher when the number of randomly selected variables in the RF was larger, for example when r=0.10 and N=5, minimal depth with mtry=39 the difference between medians was 0.014 (0.003 with mtry=27). Previously, under H\textsubscript{0}, minimal depth was shown to increase (in terms of VIM, this would be a decrease) with a larger mtry (Ishwaran et al. 2011). This was explained because with a larger mtry, the chance of splitting a noisy variable increases.

The unscaled PVIMs, VIM\textsubscript{rawperm-RF}, VIM\textsubscript{AUC} and VIM\textsubscript{party}, showed the same pattern as the scaled PVIMs, greater correlation leading to greater inflation in VIMs for the correlated variables (Figure 2.4, Figure 2.5 and Figure 2.6), but it was observed that they were unbiased under predictor correlation, as has been previously shown (Nicodemus and Malley 2009); (Nicodemus et al. 2010c); (Nicodemus 2011). Among the unscaled PVIMs, VIM\textsubscript{rawperm-RF} was the one with the largest difference between the VIMs of the correlated and uncorrelated predictor, showing a slight difference even in the extreme correlation condition (0.002). VIM\textsubscript{rawperm-CF} was observed to be unbiased, but with more variability for uncorrelated than for the correlated predictors. This small inflation on the VIM\textsubscript{rawperm-RF} was also reported by Nicodemus (2010); the authors also
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

showed more inflation in $\text{VIM}_{\text{rawperm-RF}}$ than the unconditional unscaled PVIM from CIF and the conditional PVIM under $H_0$ (they used CIF rather than RF). In this study, $\text{VIM}_{\text{rawperm-CF}}$ showed more variability in the scores for the uncorrelated predictors, as seen previously (Nicodemus et al. 2010c).

Therefore, RF based on different VIMs and minimal depth was dependent on the predictor correlation under the null and that the different VIMs and minimal depth do not show the same behaviour under correlation conditions.

2.3.3. Power detecting the true signal

2.3.3.1. Single effects association

To examine the power of the different VIMs and minimal depth in the single association models, I checked whether RF rejects the null hypothesis when it is not true. When $V_2$ is strongly-associated with the outcome, all unconditional VIMs and minimal depth rejected the null hypothesis all 500 times under all correlation conditions (Table 2.8). However, unexpectedly, this study suggests different behaviour of the $\text{VIM}_{\text{rawperm-CF}}$ under predictor correlation than the original study (Strobl et al. 2008). $\text{VIM}_{\text{rawperm-CF}}$ was not able to detect the strong signal of $V_2$ and, therefore, accepted the null hypothesis when the strength of correlation was low ($r = 0.10$) (Table 2.8). The number of variables correlated affected the behaviour of $\text{VIM}_{\text{rawperm-CF}}$: when $N = 20$, it had a better performance with a high correlation ($r = 0.80$), rejecting the null hypothesis around half the times; when $N = 40$, the percentage of the PVIM was worse, showing the highest power at 5.03%. The behaviour of the PVIM suggests that permuting the variable in a grid where more variables are correlated leads to lower prediction accuracy, even though the $\text{VIM}_{\text{rawperm-CF}}$ is applied in each tree (see section 2.3.5. for the explanation of the behaviour).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

As expected, the power of RF based on the different VIMs and minimal depth in detecting the true signal was lower in the weakly-associated model (Table 2.9). VIM_{Gini} and permutation VIMs showed different behaviours under predictor correlation. VIM_{Gini} lost power with more strength of correlation as well as when more variables were correlated under medium-high correlation conditions (Table 2.9). This could be related to the fact that VIM_{Gini} showed higher values for uncorrelated variables as seen in other studies (Nicodemus and Malley 2009); (Nicodemus 2011), which is investigated in the next subsection. Nevertheless, unconditional PVIMs showed less power than VIM_{Gini} when the correlation was low (r = 0.10), in this case they showed more power with a larger N. However, when the correlation was high (r = 0.80), they had more power when N = 5. They showed more power detecting the true signal than VIM_{Gini} when the correlation was higher, which could be related to the fact that V2 was a correlated variable and PVIMs gave more importance to correlated variables, as shown in previous studies (Strobl et al. 2008); (Nicodemus and Malley 2009); (Nicodemus et al. 2010c). The correlated variables are preferred at the first

<table>
<thead>
<tr>
<th>SAC</th>
<th>r=0.80</th>
<th>r=0.40</th>
<th>r=0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>GINI</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>LIAW</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>100</td>
<td>53.75</td>
<td>0.034</td>
</tr>
<tr>
<td>Party</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>AUC</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>mindepth 39</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>mindepth 27</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2.8. Power of detecting V2 in the single strongly-associated study (SAC), VIMs, mtry=39 and mtry=27 Mindepth.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

splits because the association is tested between the outcome and one predictor. Because the associated predictor was correlated with other non-associated ones, the other N-1 non-associated correlated predictors correlated more (showed more association) with the outcome than the uncorrelated ones and, therefore, were selected more frequently at the first split (Nicodemus et al. 2010c). When using the PVIMs from CIFs, these previous studies investigated the performance of $\text{VIM}_{\text{party}}$ (unconditional unscaled PVIM from CIF) but not the behavior of $\text{VIM}_{\text{AUC}}$. Furthermore, the observed increase in power of PVIMs when correlation was high and N=5, compared to N=20 and N=40, might have happened because with more variables correlated, the greater the chance of correlated variables to be in the pool for selection in the tree, and more chance to compete with one another to split the tree. Among the unconditional PVIMs, it was observed that $\text{VIM}_{\text{AUC}}$ had the highest power when the correlation was medium-low, however when the correlation was $r = 0.80$, the unconditional unscaled PVIM ($\text{VIM}_{\text{rawperm-RF}}$) showed the best ability to detect the true weak signal (Table 2.9).

Minimal depth with both different numbers of mtry did not show much difference in power among the different correlation conditions, but did under correlation conditions, showing slight higher power when the mtry was smaller. With both mtry values, minimal depth showed a considerable decrement in power when the correlation was high ($r = 0.80$) as well as when the correlation was medium ($r = 0.40$) and N medium-high. When correlation was high ($r = 0.80$), minimal depth rejected $H_0$ around half the time when N=5, but when N increased only rejected it less than 15% of times, having very low power when N = 40 (mtry =27 4.45%, mtry =39 5.10%) (Table 2.9). $\text{VIM}_{\text{rawperm-CF}}$, as in the strongly-associated study, showed no power in detecting the true signal when the correlation was low ($r = 0.10$) as well as when N = 20 and N = 40. Furthermore, it did not have greater than 45.5% power in any condition. In general, the conditional PVIM from RF was the least powerful, which contradicts the original study (Strobl et al. 2008). Reasons for these differences are discussed in a later section.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Performance of variable importance measures in Random Forest under correlation and application in PGC2

2.3.3.2. Interaction effects association

In this case, I studied the power of capturing both variables involved in the interaction and marginal effects of the full interacting models under correlation conditions. Under the strong association model, all unconditional VIMs and minimal depth performance well and always rejected $H_0$ (Table 2.10 and Table 2.11). $\text{VIM}_{\text{rawperm-CF}}$ was as powerful as the others when detecting the effect of the uncorrelated interacting predictor $V_{90}$. However, it had similar power as in the single association study for detecting the true correlated interacting predictor, with some power when $N = 5$, but detected the true signal less than 32% of the time.

<table>
<thead>
<tr>
<th>WAC</th>
<th>$r=0.80$</th>
<th>$r=0.40$</th>
<th>$r=0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>GINI</td>
<td>34.65</td>
<td>10.10</td>
<td>4.89</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>75.93</td>
<td>74.77</td>
<td>50.40</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>72.81</td>
<td>67.98</td>
<td>39.21</td>
</tr>
<tr>
<td>LIAW</td>
<td>72.80</td>
<td>67.94</td>
<td>39.21</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>13.25</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Party</td>
<td>65.43</td>
<td>59.75</td>
<td>45.75</td>
</tr>
<tr>
<td>AUC</td>
<td>65.47</td>
<td>60.89</td>
<td>45.41</td>
</tr>
<tr>
<td>mindepth 39</td>
<td>49.55</td>
<td>12.89</td>
<td>4.45</td>
</tr>
<tr>
<td>mindepth 27</td>
<td>51.52</td>
<td>14.08</td>
<td>5.10</td>
</tr>
</tbody>
</table>

Table 2.9. Power of detecting V2 in the single weakly-associated study (WAC), VIMs, mtry=39 and mtry=27 Mindepth.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>SAC V₂</th>
<th>r=0.80</th>
<th>r=0.40</th>
<th>r=0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5 20 40</td>
<td>5 20 40</td>
<td>5 20 40</td>
</tr>
<tr>
<td>GINI</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>LIAW</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>100 2.56 0.43</td>
<td>100 20.61 2.07</td>
<td>0 0 0</td>
</tr>
<tr>
<td>Party</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>AUC</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>mindepth 39</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>mindepth 27</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
</tbody>
</table>

Table 2.10. Power of detecting V₂ in the strong interaction study (SAC), VIMs, mtry=39 and mindepth.  

<table>
<thead>
<tr>
<th>SAC V₉₀</th>
<th>r=0.80</th>
<th>r=0.40</th>
<th>r=0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5 20 40</td>
<td>5 20 40</td>
<td>5 20 40</td>
</tr>
<tr>
<td>GINI</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>LIAW</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 0 0 0</td>
</tr>
<tr>
<td>Party</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>AUC</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>mindepth 39</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
<tr>
<td>mindepth 27</td>
<td>100 100 100</td>
<td>100 100 100</td>
<td>100 100 100</td>
</tr>
</tbody>
</table>

Table 2.11. Power of detecting V₉₀ in the strong interaction study (SAC), VIMs, mtry=39 and mindepth.
In the weakly-associated study, VIM\text{Gini}, unconditional PVIMs and minimal depth showed, in general, a similar power when detecting \( V_2 \) (correlated interacting variable under the same correlation conditions) as in the single-associated study (Table 2.12), but with slightly higher power in the single-associated study. This similar behavior may be because \( V_2 \) was still a correlated variable under the same correlated conditions, but the decrement in power may be due to it now having to compete with the other associated variable to be selected to split the tree.

<table>
<thead>
<tr>
<th>( WAC \ \text{V}_2 )</th>
<th>( r=0.80 )</th>
<th>( r=0.40 )</th>
<th>( r=0.10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>( \text{GINI} )</td>
<td>27.66</td>
<td>4.83</td>
<td>1.24</td>
</tr>
<tr>
<td>( \text{rawpermRF} )</td>
<td>64.59</td>
<td>55.66</td>
<td>36.19</td>
</tr>
<tr>
<td>( \text{BREIMAN} )</td>
<td>66.48</td>
<td>51.89</td>
<td>30.99</td>
</tr>
<tr>
<td>( \text{LIAW} )</td>
<td>66.46</td>
<td>51.87</td>
<td>30.94</td>
</tr>
<tr>
<td>( \text{rawpermCF} )</td>
<td>8.65</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>( \text{Party} )</td>
<td>67.01</td>
<td>53.29</td>
<td>43.62</td>
</tr>
<tr>
<td>( \text{AUC} )</td>
<td>65.40</td>
<td>52.89</td>
<td>43.86</td>
</tr>
<tr>
<td>( \text{mindepth } 39 )</td>
<td>28.01</td>
<td>4.88</td>
<td>0.45</td>
</tr>
<tr>
<td>( \text{mindepth } 27 )</td>
<td>30.01</td>
<td>4.59</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Table 2.12. Power of detecting \( V_2 \) in the weak interaction study (\( WAC \)), VIMs, \( mtry=39 \) and \( mtry=27 \) Mindepth.

When detecting \( V_{90} \) (uncorrelated interacting variable), VIM\text{Gini}, the unconditional unscaled PVIM, the PVIMs from CIF and minimal depth showed an increase in power mainly under high correlation conditions, and mostly with a larger number of variables correlated (Table 2.13). This increase in power of the unconditional PVIMs from RF and CIF might be because with a higher \( N \), there is a higher probability of having correlated predictors in the pool of predictors to be selected to split the tree associated predictors. Then, as previously shown (Nicodemus and Malley 2009); (Nicodemus et
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

al. 2010c), the uncorrelated predictors had a higher selection frequency because the correlated predictors were competing with each other; this competition was stronger with higher correlation. Among VIM$_{Gini}$, the unconditional PVIMs and minimal depth, the highest difference in power between detecting the uncorrelated and the correlated interacting predictors was observed in VIM$_{Gini}$ and minimal depth with both mtry values. The fact that VIM$_{Gini}$ showed that large increase was due to V$_{90}$ being an uncorrelated predictor and, as shown previously (Nicodemus and Malley 2009); (Nicodemus 2011), VIM$_{Gini}$ gives larger values to uncorrelated predictors. The observed increase in minimal depth for V$_{90}$ might be related to the fact that it might inflate (decrease positive values of minimal depth) the values for uncorrelated predictors, which is studied in the next section. Moreover, the fact that selection frequencies for uncorrelated variables are higher with more correlation and with more variables correlated could also be one of the reasons. There was not much difference in power of minimal depth between both different values of mtry. However, the scaled PVIMs showed similar power detecting V$_{90}$ and V$_2$, with the exception of the extreme correlation condition when it is more capable of capturing the effect of V$_{90}$.

Interestingly, VIM$_{rawperm-CF}$ showed a completely different behaviour to the single-associated study when only a correlated variable was influential. As the correlation was higher, so also the power in detecting the true signal of the uncorrelated interacting predictor (V$_{90}$) was higher as well as with a higher value of N. In fact rejected the null hypothesis more than 59% of the times, with the exception of low correlation (r = 0.10) when VIM$_{rawperm-CF}$ showed low power (Table 2.13). VIM$_{rawperm-CF}$ was dramatically more powerful in detecting the signal from the uncorrelated interacting predictor than from the correlated one. This suggests that VIM$_{rawperm-CF}$ is able to detect uncorrelated associated variables and that correlation improves its ability to capture the true signal, although it has poor performance when the variable is correlated. VIM$_{rawperm-CF}$ was previously shown to give higher scores in median and more variability for the uncorrelated predictors (Nicodemus et al. 2010c), so this fact might be the reason of the higher power in detecting the uncorrelated true predictor.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>WAC V₉₀</th>
<th>r=0.80</th>
<th>r=0.40</th>
<th>r=0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>GINI</td>
<td>65.73</td>
<td>75.98</td>
<td>86.53</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>65.78</td>
<td>61.93</td>
<td>50.90</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>63.91</td>
<td>51.54</td>
<td>41.44</td>
</tr>
<tr>
<td>LIAW</td>
<td>63.91</td>
<td>51.54</td>
<td>41.47</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>67.05</td>
<td>81.34</td>
<td>89.74</td>
</tr>
<tr>
<td>Party</td>
<td>82.59</td>
<td>91.77</td>
<td>94.71</td>
</tr>
<tr>
<td>AUC</td>
<td>82.58</td>
<td>91.82</td>
<td>94.84</td>
</tr>
<tr>
<td>mindepth 39</td>
<td>67.06</td>
<td>79.65</td>
<td>88.57</td>
</tr>
<tr>
<td>mindepth 27</td>
<td>65.87</td>
<td>79.94</td>
<td>89.46</td>
</tr>
</tbody>
</table>

Table 2.13. Power of detecting $V_{90}$ in the weak interaction study (WAC), VIMs, mtry=39 and mtry=27 Mindepth.

Overall, the power of VIM$_{\text{Gini}}$, VIM$_{\text{rawperm-CF}}$ and minimal depth was greater in detecting the uncorrelated interacting predictor than the correlated one when the correlation was medium-high ($r = 0.80$ and $r = 0.40$); while the three different measures were weak detecting the correlated associated predictor, mostly in the extreme correlation situation ($r = 0.80$ and $N = 40$). The unconditional unscaled PVIMs from CIF showed more difference in power detecting the correlated and the uncorrelated interacting predictors than the unscaled PVIM from RF. The power of the scaled PVIMs was similar with respect to both interacting predictors except for the situation of extreme correlation.

It is important to say that the power of the different VIMs and minimal depth was also calculated considering the null distributions after permuting the outcome of each database under $H_A$. In this case, all VIMs and minimal depth had similar power compared to when they were considering null models (defined on the subsection
For this reason, the power results from the null hypothesis when permuting the outcome are illustrated in the Appendix A (Table A.21 - A.26).

2.3.4. Distributions of RF VIMs under HA

2.3.4.1. Single association study

After applying different VIMs and minimal depth from different implementations of RF under HA, the findings of this study showed that the strength of correlation and the number of correlated variables had a dramatic impact on the performance of RF. Previous studies showed that VIMGini and the PVIMs were sensitive to correlation conditions between variables (Díaz-Uriarte and Alvarez de Andrés 2006); (Strobl et al. 2008); (Nicodemus and Malley 2009); (Nicodemus et al. 2010c). Predictor correlation was observed to influence the behaviour of the VIMs and minimal depth mainly when the association between V2 and the outcome is weak. See Appendix A for the VIM median values for V2, for both correlated and uncorrelated predictors (Table A.7, Table A.8 and Table A.9 in the strong single association study; Table A.10, Table A.11 and Table A.12 in the weak single association study). The figures of all VIMs and minimal depth (both mtry values) for all other correlation conditions (different than the ones shown here), under the weakly-associated study, are illustrated in the Appendix A in the Figure A.15 - A.20. To better understand the behavior of minimal depth, the median of the depth threshold for variable selection when applying minimal depth with both values of mtry is reported. It was 8.984 in the strongly-associated single study and 8.971 in the weakly-associated one (Table A.28. Appendix A).

Under HA, RF based on the different VIMs, VIMAUC, VIMparty and minimal depth showed the largest scores for the influential predictor under all correlation conditions when the association was strong (Figure 2.7, as an illustration; see Appendix A Figure A.7 - A.14 for the other conditions), with the exception of VIMrawperm-CF when the correlation was low (r = 0.10) that gave no importance to all variables. In general, it
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

was observed that predictor correlation affected the VIMs and minimal depth for the non-associated predictors in both association studies. Despite is not appreciable in the figures because the scale of the Y-axis on the plots is different, the difference between VIM and minimal depth medians for correlated and uncorrelated predictors was higher when the association with the single predictor was stronger (see Appendix A). The p-values from the Wilcoxon test for all VIMs and minimal depth were all less than 0.05 under all correlation conditions, with the exception of when $r = 0.10$ and $N = 5$, which shows the statistically significant difference between the median scores for correlated predictors and the median scores for uncorrelated non-associated predictors.

![Figure 2.7. RF VIMs under HA for $V_2$, two variable correlated $V_3$ and $V_6$, and two independent variables $V_{42}$ and $V_{90}$ when $r = 0.80$ and $N = 40$ in the strongly-association single study.](image-url)
In spite of the fact that in the strongly-associated condition $\text{VIM}_{\text{Gini}}$ ranked the non-associated correlated predictors higher than uncorrelated ones, $\text{VIM}_{\text{Gini}}$ showed larger scores for the uncorrelated ones when the association between $V_2$ and the outcome was weak and as more correlation among variables and a larger $N$ was present (Figure 2.8, Figure 2.9 and Figure 2.10). In fact, under high correlation conditions ($r = 0.80$) and medium-high numbers of correlated variables, $\text{VIM}_{\text{Gini}}$ resulted in lower values for the influential variable than the uncorrelated predictors. This inflation for the uncorrelated predictors by $\text{VIM}_{\text{Gini}}$ has also been reported previously (Nicodemus and Malley 2009); (Nicodemus 2011). Higher values for uncorrelated variables may be due to the variable importance is based on a decrease in impurity and that uncorrelated variables did not “share” information with others. Furthermore, as the association of $V_2$ was low and $V_2$ was also correlated with other variables led to have higher values of the $\text{VIM}_{\text{Gini}}$ for the uncorrelated predictors. Larger values of $N$ implied more chance of correlated predictors belonging to the pool of predictors available for splitting, these predictors shared information with the true predictor ($V_2$), which made the measure less able to capture the correlated predictors as the uncorrelated predictors did not have any common information with $V_2$ (the VIM is based on a decrease in impurity).

In addition, minimal depth showed a similar performance of $\text{VIM}_{\text{Gini}}$. With both values for mtry, the largest scores were observed for the uncorrelated predictors when correlation was high ($r = 0.80$), even larger than the associated predictor ones when $N = 20$ and $N = 40$. If the association was low, a larger mtry led to a greater difference between the medians for correlated and uncorrelated predictors. If the association was stronger, they showed shorter distance between medians when mtry = 39, but the difference in medians when mtry = 39 showed a slightly different value to when mtry = 27 (see Appendix A Table A.7, Table A.8 and Table A.9 in the strong single association study; Table A.10, Table A.11 and Table A.12 in the weak single association study). This was in accordance with what Ishwaran et al (2011) reported: when the signal was strong, a larger mtry resulted in a good performance of minimal depth, but when the association was weak, a larger mtry might not be optimal.
As more correlation was present among predictors, unconditional PVIMs, both scaled and unscaled, were larger for correlated non-associated predictors than for uncorrelated non-associated ones. This inflation for correlated variables may be because correlated variables are chosen more often in first splits in the tree, as suggested by (Strobl et al. 2008); (Nicodemus et al. 2010c). As explained above, and as shown in Nicodemus et al. (2010c), the inflation for correlated non-associated variables under $H_0$ was due to the correlation between them and the associated variable, which led to more correlation between each correlated non-associated predictor, when testing their association (single or univariate) with the outcome, which therefore resulted in greater association. For instance, in real studies where SNPs can be in LD, if one SNP (let’s called it SNPa) is in LD with other non-associated SNP (SNPb), SNPb may also appear as an associated variant with the outcome. This behaviour was not because of data generation, it relates to how trees are built, as the first split the association is tested between one single variable and the outcome. In addition, a larger number of correlated variables (N) also overestimated the $\text{VIM}_{\text{rawperm-RF}}$ and the scaled PVIMs for correlated predictors and made the difference between the median VIMs higher, although a slight difference in the unscaled PVIMs was observed ($\text{VIM}_{\text{rawperm-RF}}$ shows the largest difference with a value of 0.0027 when $r = 0.80$ and $N = 4$). This related to the finding of a previous study which compared unscaled PVIM with the scaled ones (Nicodemus et al. 2010c). The authors observed that with a larger mtry the values for correlated variables can be inflated. Here, mtry is set up to be equal to 39 under all correlation conditions, so if only 5 variables are correlated over a total of 100, there is more probability of the uncorrelated being randomly selected for the pool of variables used to split the tree. However, with more variables correlated, there is a higher probability for correlated variables to be selected at the first split, and correlated variables being ranked higher than uncorrelated may be because they are correlated with the associated predictor. $\text{VIM}_{\text{party}}$ and $\text{VIM}_{\text{AUC}}$ also resulted in larger scores for the correlated predictors under high correlation conditions, but this difference did not show an increase when there were more correlated variables. The inflation of $\text{VIM}_{\text{party}}$ for non-associated correlated predictors compared to the non-associated uncorrelated ones was also shown by Nicodemus et al. (2010c).
Strobl (2008) showed that the inflation and variability of the conditional PVIM for correlated predictors was lower than the unscaled PVIM under high correlation conditions. The results of VIM_{rawperm-CF} in this study also showed less variability for correlated predictors than VIM_{rawperm-RF}, but the PVIM showed more variability for uncorrelated predictors than for correlated ones, which was in accord to what Nicodemus et al. (2010c) showed, although the medians were similar across all predictors, the influential V₂, the correlated and the uncorrelated ones. It is important to say that both previous studies (Strobl et al. 2008); (Nicodemus et al. 2010c) applied the conditional PVIM using CIF, not RF.

Figure 2.8. RF VIMs under HA for V₂, two variable correlated V₃ and V₆, and two independent variables V₄₂ and V₉₀ when r = 0.10 and N = 5 in the weakly-association single study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 2.9. RF VIMs under HA for V2, for two variable correlated V3 and V6, and for two independent variables V42 and V90 when $r = 0.40$ and $N = 20$ in the weakly-association single study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 2.10. RF VIMs under HA for V2, for two variable correlated V3 and V6, and for two independent variables V42 and V90 when $r = 0.80$ and $N = 40$ in the weakly-association single study.
2.3.4.2. Interaction association study

In this subsection I will present the results for the interaction synthetic data under the alternative hypothesis, the Figure 2.11 illustrates the behaviour of the different measures in the strong association study under the extreme correlation condition. Figure 2.12, Figure 2.13, and Figure 2.14 illustrate the lowest, medium and extreme correlation conditions under the weak association study. See Appendix A for the other correlation condition under both weak and strong association studies (Figures A.21 - A.28 in the strong association study; and Figure A.29 - A.34). In the interaction studies, the median depth threshold for both mtry under all correlation condition was 10.193 and 9.953 in the strong interaction and weak interaction studies respectively (see Table A.28, Appendix A).

All RF VIMs for interaction effects had similar performance to the single effect associated study under a strong association. They clearly ranked the correlated interacting predictor higher even under high correlation conditions (Figure 2.11), with the exception of VIM_{rawperm-CF} for capturing the signal of \( V_2 \), when the correlation is \( r = 0.10 \) (all values were 0) and under medium-high correlation when N was larger than five.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 2.11. RF VIMs under HA in the strongly-associated interaction study for $V_2$ and $V_{90}$ (interacting variables), two variable correlated $V_3$ and $V_6$, and one independent variable $V_{43}$, when $r = 0.80$ and $N = 40$.

However, weak association lead to worse performance, mainly under high predictor correlation (Figure 2.12, Figure 2.13, and Figure 2.14). Higher strength of correlation as well as larger number of variables correlated resulted in less capability of the different measures to capture the signal of the correlated interacting predictor. When $r = 0.10$ and $N = 5$, the VIM scores of the unconditional VIMs and minimal depth were higher for the interacting variables, and in general the distributions of $V_2$ and $V_{90}$ were similar.
When the correlation was higher, \( V_{90} \) (an interacting and uncorrelated predictor) was ranked clearly higher than non-associated predictors by all VIMs and minimal depth, even under correlation conditions; with greater correlation there was better capability to detect the signal (ranked higher than \( V_2 \) too). Nevertheless, their ability to capture the signal of \( V_2 \) (the correlated associated predictor) was not the same.

Under high predictor correlation, and as said above, in the interaction association study all VIMs and minimal depth have a similar performance as when only a single correlated predictor was associated with the outcome. \( \text{VIM}_\text{Gini} \) and minimal depth gave greater importance to uncorrelated variables than to correlated ones, including the interacting correlated one. Minimal depth with \( \text{mtry} = 27 \) ranked \( V_2 \) slightly higher than with \( \text{mtry} = 39 \) when \( r = 0.80 \) (for \( V_{90} \) it was the opposite). Under high correlation, permuted unconditional PVIMs, \( \text{VIM}_{\text{AUC}} \) and \( \text{VIM}_{\text{party}} \) gave the largest importances to both interacting variables and then to the correlated predictors. The unscaled PVIM had larger VIM scores for both interacting predictors under high correlation conditions as well as when more variables were correlated. However, it was observed that the \( \text{VIM}_{\text{AUC}} \) and \( \text{VIM}_{\text{party}} \) median scores for \( V_2 \) decreased under high correlation with a larger \( N \), although they showed the opposite behaviour for \( V_{90} \) larger scores with more correlation and larger \( N \).

The \( \text{VIM}_{\text{rawperm-CF}} \) only suggested association with \( V_{90} \), the uncorrelated interacting variable. The inflation of \( \text{VIM}_{\text{rawperm-CF}} \) and scaled PVIMs for the uncorrelated variables when they were associated was shown by Nicodemus et al. (2010c), but the authors did not study RF for capturing interactions, only main effects. So, this also suggests that when a predictor is interacting and uncorrelated results with a higher importance than an interacting correlated predictor when they are involved in interactions. Furthermore, it might be that \( \text{VIM}_{\text{rawperm-CF}} \) is only able to detect the main effect of the uncorrelated predictor without capturing the signal from the interaction term.

A recent study examined the ability of RF based on different VIMs to detect the signal of interacting predictors from different models including those that involve main
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

effects and the interaction of the predictors with the main effects (Wright, Ziegler and König, 2016). In that study the authors showed that PVIMs and VIM\textsubscript{Gini} resulted in larger scores for the interacting predictors under correlation conditions. In addition, in the results of this study the medians of the PVIMs were higher with high correlation than low correlation for both $V_{90}$ and $V_2$, but were lower for $V_2$ than for $V_{90}$ in high correlation conditions. VIM\textsubscript{Gini} also preferred both interacting predictors with low-medium correlation than non-associated predictors. However, with high correlation ($r = 0.80$), VIM\textsubscript{Gini} showed larger scores for the uncorrelated non-associated predictors than for the interacting correlated predictor or for the correlated non-associated ones, while PVIMs still gave higher scores to both interacting predictors.

Minimal depth showed similar behaviour in the interacting models between both values of mtry, and similar behaviour to VIM\textsubscript{Gini}. Under high correlation conditions the median of the scores for the interacting correlated predictor was lower than the medians for the non-associated predictors (Figure 2.14). Minimal depth showed slightly larger scores for both interacting variables with a large mtry, with the exception of high correlation when minimal depth had a slight larger values for $V_2$ with mtry = 27. This was in accordance with what Wright \textit{et al.} (2016) reported: minimal depth was not able to capture interacting effects under correlation.

In summary, all VIMs and minimal depth showed higher ranks for the uncorrelated interacting variable than for the correlated interacting variable when the correlation was high (0.80). Unconditional RF PVIMs showed the lowest distances between the VIM medians for both interacting predictors compared to the conditional PVIM, the unscaled PVIMs from CIF, and minimal depth. This difference in the median values between the interacting predictors may be due to the interaction between them. As the interaction involved two predictors with also their main effects, the interaction could be correlated with the variables of the main effects, which would transform $V_{90}$ in a correlated predictor but only with the interaction effect, and $V_2$ with all other N-1 correlated predictors and the interaction effect. In this case, $V_{90}$ would only correlate with the interaction, and as the number of correlated variables affected the
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

performance of the VIMs and minimal depth, the larger values for $V_{90}$ may be due to that fact.

Figure 2.12. RF VIMs under HA in the weakly-associated interaction study for $V_2$ and $V_{90}$ (interacting variables), two variable correlated $V_3$ and $V_6$, and one independent variable $V_{42}$, when $r = 0.10$ and $N = 5$. 
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 2.13. RF VIMs under HA in the weakly-associated interaction study for V_2 and V_90 (interacting variables), two variable correlated V_3 and V_6, and one independent variable V_42, when $r = 0.40$ and $N = 20$. 
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Performance of variable importance measures in Random Forest under correlation and application in PGC2

Figure 2.14. RF VIMs under HA in the weakly-associated interaction study for V2 and V90 (interacting variables), two variable correlated V3 and V6, and one independent variable V42 and V90 when $r = 0.80$ and $N = 40$. 
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

2.3.5. Conditional PVIM with different correlation cut-off

VIM_{rawperm-CF} showed no power to detect the true signal when the correlation was low (r = 0.10), in either the single association models or the interaction models in either association study. The reason for this lack of power may be the correlation cut-off, which was fixed at 0.05. As the cut-off was so low (although greater than the estimated correlation for the uncorrelated predictors; around 0), the PVIM might be considering many variables with correlation higher than 0.05, so the grid where the predictors were permuted would become so small that there was more chance for the permuted predictor to be similar to the original one, therefore causing less prediction accuracy. To investigate this hypothesis, multiple thresholds were studied. Under all correlation conditions, the three cut-offs considered before (K = 0.05, K = 0.35 and K = 0.75) were applied.

As expected, in the single association study, when the correlation cut-off was fixed at 0.05, VIM_{rawperm-CF} showed no power (Table 2.14 and Table 2.15) because the scores were zero under both H_0 and H_A. However, when the cut-off was set to 0.35 or to 0.75, the PVIM showed different amounts of power when the correlation was 0.10 or 0.80 for K = 0.35, as well as when the correlation was 0.40 and 0.10 for K = 0.75 (Table 2.14 and Table 2.15). In general, if the cut-off was higher than the correlation between predictors (for example, K = 0.75 when r = 0.10 and r = 0.40), the PVIM showed a similar power to the unconditional unscaled PVIM, which may be because the PVIM did not find any predictor correlated with another (or only a few of them). As VIM_{rawperm-CF} permuted the variable considering all observations, or grids with a lot of observations, the shuffled predictor was different to the original one (non-permuted) and, therefore, its power was similar to the VIM_{rawperm-RF} one. If the cut-off was lower than the correlation among variables (for example, K = 0.35 both r = 0.40 and r = 0.80), the power decreased in general, but more when the number of correlated variables was medium-high (N = 20 and N = 40) than when there was 5 correlated variables, which was seen because the more correlated predictors there were, the smaller the degree of permutation, and the less the difference between the predictor before and after
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

permuting, which suggests that more correlated predictors may lead to a smaller chance of detecting association.

<table>
<thead>
<tr>
<th>V2 weak</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5 20 40</td>
<td>5 20 40</td>
<td>5 20 40</td>
</tr>
<tr>
<td>K=0.05</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>K=0.35</td>
<td>13.99</td>
<td>0.0056</td>
<td>45.21</td>
</tr>
<tr>
<td>K=0.75</td>
<td>13.25</td>
<td>0</td>
<td>74.62</td>
</tr>
</tbody>
</table>

Table 2.14. Power of VIM_{rawperm-CF} in detecting V2 under all correlation conditions with the three different cut-offs in the weak single association study.

<table>
<thead>
<tr>
<th>V2 strong</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5 20 40</td>
<td>5 20 40</td>
<td>5 20 40</td>
</tr>
<tr>
<td>K=0.05</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>K=0.35</td>
<td>100</td>
<td>49.71</td>
<td>100</td>
</tr>
<tr>
<td>K=0.75</td>
<td>100</td>
<td>53.75</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2.15. Power of VIM_{rawperm-CF} in detecting V2 under all correlation conditions with the three different cut-offs in the strong single association study.

Figure 2.15 and Figure 2.16 illustrate the conditional PVIM when the three cut-offs were under the extreme and the medium correlation condition (r = 0.40 and N = 20) respectively, in the weakly-associated single study. The PVIM showed higher rankings for V2 when the correlation was medium, but in the extreme situation still did not show larger scores for V2 because of a large number of correlated variables.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 2.15. VIMrawperm-CF under HA in the weakly-associated single study for $V_2$, two variable correlated $V_3$ and $V_6$, and one independent variable $V_{42}$ and $V_{90}$ when $r = 0.40$ and $N = 20$. 
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Performance of variable importance measures in Random Forest under correlation and application in PGC2

Figure 2.16. VIM$_{rawperm-CF}$ under HA in the weakly-associated single study for $V_2$, two variable correlated $V_3$ and $V_6$, and one independent variable $V_{42}$ and $V_{90}$ when $r = 0.80$ and $N = 40$.

In the interaction study the results suggested the same behavior for the PVIM as in the single association study with the correlated interacting variable $V_2$ (Table 2.16 and Table 2.18; Figure 2.17 and Figure 2.18). For the uncorrelated variable, when the cut-off was very low, VIM$_{rawperm-CF}$ showed no power (Table 2.17 and Table 2.19), because the importance scores were null in all correlation conditions, as well as for the correlated variable (Figure 2.17 and Figure 2.18), which might be due to the variables after and before permutation were similar. When the correlation was low ($K = 0.35$ and $K = 0.75$) but the cut-offs were greater than the correlation condition,
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

VIM_{rawperm-CF} had the same behavior as VIM_{rawperm-RF} because it was permuting all variables across all observations (the same reason as for the correlated variable). This also happened when the cut-off was K = 0.75 and the correlation between N predictors was 0.40 (see bottom plot of Figure 2.17). But when the cut-off was slightly lower than the strength of correlation, the capability of VIM_{rawperm-CF} to capture the effect of the uncorrelated interacting variable with more correlation among predictors and larger N increased (Table 2.17 and Table 2.19). Under high correlation conditions (r = 0.80) VIM_{rawperm-CF} showed almost no difference detecting V_{90} between having K = 0.35 and K = 0.75, (Table 2.17 and Table 2.19); it was slightly more powerful when N = 20 and N = 40 with K = 0.35, and slightly less powerful when N = 5 with K = 0.35.

<table>
<thead>
<tr>
<th>V2 weak</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5 20 40</td>
<td>5 20 40</td>
<td>5 20 40</td>
</tr>
<tr>
<td>K=0.05</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>K=0.35</td>
<td>8.13 0.08</td>
<td>31.20 0.16</td>
<td>56.42 59.39</td>
</tr>
<tr>
<td>K=0.75</td>
<td>13.25 0.02</td>
<td>74.62 66.58</td>
<td>61.93 66.28</td>
</tr>
</tbody>
</table>

Table 2.16. Power of VIM_{rawperm-CF} in detecting V_{90} under all correlation conditions with the three different cut-offs in the weakly-associated interaction study.

<table>
<thead>
<tr>
<th>V90 weak</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5 20 40</td>
<td>5 20 40</td>
<td>5 20 40</td>
</tr>
<tr>
<td>K=0.05</td>
<td>0 0 0</td>
<td>0 0 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>K=0.35</td>
<td>66.86 81.83</td>
<td>90.64 59.75</td>
<td>67.57 77.36</td>
</tr>
<tr>
<td>K=0.75</td>
<td>67.05 81.34</td>
<td>89.74 57.53</td>
<td>59.57 52.64</td>
</tr>
</tbody>
</table>

Table 2.17. Power of VIM_{rawperm-CF} in detecting V_{90} under all correlation conditions with the three different cut-offs in the weak interaction association study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>V2 strong</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5  20  40</td>
<td>5  20  40</td>
<td>5  20  40</td>
</tr>
<tr>
<td>K=0.05</td>
<td>0  0  0  0  0  0  0  0  0</td>
<td>0  0  0  0  0  0  0  0  0</td>
<td></td>
</tr>
<tr>
<td>K=0.35</td>
<td>100  15.66  0.02  100  20.61  2.07  100  100  100</td>
<td>100  100  100  100  100  100  100  100  100</td>
<td></td>
</tr>
<tr>
<td>K=0.75</td>
<td>100  25.64  0.44  100  100  100  100  100  100</td>
<td>100  100  100  100  100  100  100  100  100</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.18. Power of VIM$_{rawperm-CF}$ in detecting V$_2$ under all correlation conditions with the three different cut-offs in the strong interaction association study.

<table>
<thead>
<tr>
<th>V90 weak</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5  20  40</td>
<td>5  20  40</td>
<td>5  20  40</td>
</tr>
<tr>
<td>K=0.05</td>
<td>0  0  0  0  0  0  0  0  0</td>
<td>0  0  0  0  0  0  0  0  0</td>
<td></td>
</tr>
<tr>
<td>K=0.35</td>
<td>100  100  100  100  100  100  100  100  100</td>
<td>100  100  100  100  100  100  100  100  100</td>
<td></td>
</tr>
<tr>
<td>K=0.75</td>
<td>100  100  100  100  100  100  100  100  100</td>
<td>100  100  100  100  100  100  100  100  100</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.19. Power of VIM$_{rawperm-CF}$ in detecting V$_{90}$ under all correlation conditions with the three different cut-offs in the strong interaction association study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 2.17. VIMrawperm-CF under HA in the weakly-associated interaction study for V2, two variable correlated V3 and V6, and one independent variable V42 and V90 when r = 0.40 and N = 20.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 2.18. VIM\textsubscript{rawperm-CF} under HA in the weakly-associated interaction study for \( V_2 \), two variable correlated \( V_3 \) and \( V_6 \), and one independent variable \( V_{42} \) and \( V_{90} \) when \( r = 0.80 \) and \( N = 40 \).
2.4. Discussion

This simulation showed that the strength of correlation and the number of correlated variables affected the performance of VIMs and minimal depth. As most real data will include correlations between predictors, the choice of VIM is crucial to avoid spurious association signals.

RF is well-studied to detect associations with binary outcomes and features coded as 0, 1, 2 as in real genetic studies (classical GWAS). But the present study performed RF to examine its capability to predict associations between continuous outcomes and continuous predictors, interactions between them, since in high-dimensional genetic studies data may need to be transformed because of population stratification (PS) resulting in continuous new variables. Such models are intended to find associations between continuous phenotypes and continuous genotypes (Zhao et al. 2012).

According to the empirical power and to the distributions of VIMs and minimal depth, having a strong association between the predictor and the outcome ensures a good capability of all VIMs and minimal depth to detect the right variables both in single and interaction association studies. An exception is the conditional PVIM when either the number of correlated variables or the correlation is not low. Nevertheless, according to single association effects, in the weakly associated continuous study, which is much more similar to what one might expect from a GWAS (non-classical one, e.g. after PS) as the effect size of each SNP is low, unconditional PVIMs have a better performance than VIM\textsubscript{Gini} under predictor correlation, as has been shown in previous studies (Nicodemus and Malley 2009); (Nicodemus 2011). Under $H_0$, VIM\textsubscript{Gini} showed a bias which may lead to spurious results when is applied under predictor correlation conditions. In addition, this study suggests that the unconditional unscaled PVIM is superior to the scaled ones under correlation conditions, as previously proposed by Díaz-Uriarte and Alvarez de Andrés (2006) and Nicodemus et al. (2010c). These three PVIMs overestimated the importance of correlated predictors under correlation, but the scaled ones showed to be biased under the $H_0$ as the medians
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Performance of variable importance measures in Random Forest under correlation and application in PGC2

of VIM scores for the correlated predictors were considerable greater than the others (under $H_0$ all medians should be around 0).

Under $H_A$, $VIM_{AUC}$ and $VIM_{party}$ (PVIMs from CIF) showed the highest power to detect the signal of $V_2$ under medium-low correlation conditions (power decreased under high correlation conditions). In contrast, unconditional PVIMs from RF displayed the highest levels of power under high correlation when either $N = 5$ or $N = 20$. Under the extreme correlation condition ($r = 0.80$ and $N = 40$), the unconditional unscaled PVIM from RF ($VIM_{rawperm-RF}$) was the most powerful importance measure under study, although slightly higher than the unconditional PVIMs from CIF. All unconditional PVIMs gave larger scores for $V_2$ (the associated predictor) and then for the non-associated correlated predictors. Previous studies also showed an inflation for the correlated non-associated predictors compared to the uncorrelated non-associated predictors from the unconditional unscaled PVIM using RF (Nicodemus et al. 2010c) and CIF (Strobl et al. 2008); (Nicodemus et al. 2010c); $VIM_{AUC}$ was not studied. The similar behaviour between both PVIMs from CIF was due to them both basically measure prediction accuracy, as AUC measures how well one model is able to predict.

However, minimal depth and $VIM_{Gain}$ ranked the uncorrelated non-associated predictors higher than the correlated non-associated predictor. Furthermore, they resulted in larger VIM median scores for the uncorrelated non-associated predictors than for $V_2$ (associated) when under high strength of correlation, and either $N = 20$ or $N = 40$. Very little difference in power was seen between minimal depth when mtry = 27 and mtry = 39, but slightly better performance was observed when mtry was lower under high predictor correlation conditions and when the association was weak. In a previous study, a large value of mtry was suggested and when the association was strong but a large value could be detrimental under weak association (Ishwaran et al. 2011). Moreover, the authors showed that minimal depth performed well under correlation conditions when the association was strong, but did not study its behaviour under weak association conditions in presence of correlation. The present study is in accordance with their results under strong association. However, this study found that minimal depth had a good performance with both values of mtry when the strength of
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

correlation was low, but when the strength was high minimal depth showed misleading results with both mtry numbers under weak association.

Despite the fact that conditional PVIM is unbiased under $H_0$, RF based on $\text{VIM}_{\text{rawperm-CF}}$ (which has been specifically created to deal with situations under correlation) could not detect the true association most of the time under any situation with correlation with the exception of $r=0.40$ and $N=5$ and $r=0.80$ and $N=5$. Because of this behaviour, $\text{VIM}_{\text{rawperm-CF}}$ was investigated, considering the same three correlation cut-offs in all nine correlation conditions. The results suggest that a very small cut-off affects the degree of permutation, this leading to spurious results. In addition, after setting a larger cut-off than the correlation present among predictors, $\text{VIM}_{\text{rawperm-CF}}$ performs the same as $\text{VIM}_{\text{rawperm-RF}}$ because the predictor is shuffled across all observations. When the strength of correlation was median-high and the cut-off was lower (slightly), the number of correlated predictors played an important role in the behaviour of $\text{VIM}_{\text{rawperm-CF}}$. In those cases when $N = 20$ and $N = 40$ this PVIM was underpowered when detecting true positives under weak associations, and under strong associations mainly when $N = 40$. The difference in behaviour from the original study (Strobl et al. 2008) is because the authors considered a higher correlation ($r = 0.90$) and only four variables were correlated. In fact, in this study under conditions with medium-high correlation and five correlated variables, $\text{VIM}_{\text{rawperm-CF}}$ was also able to detect the strong signal from $V_2$. Another difference is that Strobl et al. (2008) applied the PVIM using CIF, and the present study used RF. But the main reason for this contradictory behaviour is the lack of consideration of a larger number of correlated variables in the original study, as a greater number of correlated predictors leads to a smaller chance of detecting an association.

Therefore, as in Nicodemus et al. (2010c), the present study suggests that $\text{VIM}_{\text{rawperm-RF}}$ may be preferable in studies with a larger number of highly correlated predictors when studying single associations such as GWAS where the causal SNP is in a block of LD with other variants, and SNPs in LD with the causal one can serve as a proxy. But $\text{VIM}_{\text{AUC}}$ and $\text{VIM}_{\text{party}}$, which are based on CIF, may be better to apply in
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

smaller dimensional datasets when looking for true signals in a group of correlated predictors and trying to avoid false signals from correlated predictors.

When detecting interaction effects, previous studies have shown the efficiency of RF. This study extended the investigation by looking at different correlation conditions covering high strengths. A recent study (Wright, Ziegler and König, 2016) also considered correlation between predictors, with correlation of 0.14 with a standard deviation of 0.23. The present study compared seven different VIMs as well as the performance of RF based on minimal depth with two different mtry in a study with 500 iterations both under the alternative and the null distribution.

The interaction simulation results suggest that the different VIMs and minimal depth have a similar performance when detecting a single associated variable correlated with other predictors. However, if the variable involved in the interaction is independent of other predictors, $VIM_{\text{Gini}}$, PVIMs from RF, $VIM_{\text{AUC}}$, $VIM_{\text{party}}$ and minimal depth improve their ability to detect the uncorrelated associated predictor if the correlation among other predictors is higher. $VIM_{\text{AUC}}$ and $VIM_{\text{party}}$ showed the highest power for capturing $V_{90}$ in the extreme correlation condition, followed by $VIM_{\text{rawperm-CF}}$.

Wright et al. (2016) previously found that $VIM_{\text{Gini}}$ gives higher importance to both interacting predictors than the unconditional unscaled PVIM. The reason of the contradictory behaviour of $VIM_{\text{Gini}}$, between their results and the ones from the present study, was that the authors considered a correlation of $r = 0.14$ (SD 0.23), which is lower than 0.80 and 0.40. In this study, it was observed that when the correlation is 0.40, $VIM_{\text{Gini}}$ gives greater scores to both interacting variables than to any non-associated variable. But it was also shown that with a high level of correlation ($r = 0.80$, and $N = 20$ and $N = 40$), $VIM_{\text{Gini}}$ resulted in larger scores for the uncorrelated non-associated predictors than for the correlated predictors, including the influential and the non-associated ones, while PVIMs still give more scores to both interacting predictors. One might think, from the present study, that $VIM_{\text{Gini}}$ is more capable of detecting interactions than single associations. However, its behaviour detecting the correlated interacting predictor was similar to when detecting the main effect from the
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

associated predictor (single study), which was correlated. This clear preference for the uncorrelated interacting predictor may be because $VIM_{Gini}$ gives greater scores to uncorrelated predictors than to the correlated ones, and it is more capable of detecting its effect (only the main effect). So, whether $VIM_{Gini}$ captures predictors involved in interactions better than the ones involved only in main effects is a question to be addressed in further studies. This should include at least one main effect from an uncorrelated predictor (single associated studies) under the same correlation conditions, and these results could be compared with the present interaction association study.

In the real world, it is unusual to find situations where variables are strongly associated with an outcome, so it is so important to pay attention to the weakly-associated study results. In psychiatric genetics, most study variables are weakly associated (low effect size) with the phenotypes and correlated with each other (because of LD), for instance, in a non-classical GWAS; or in RNA sequencing (RNA-seq) data analysis. Therefore, the results presented in this study are a good guide to follow up when applying RF in real situations with continuous outcome and continuous predictors. The findings suggest applying RF based on $VIM_{AUC}$, $VIM_{party}$, $VIM_{rawperm-RF}$ under correlation conditions. The time consumed applying $VIM_{rawperm-RF}$ is significantly lower than performing $VIM_{AUC}$ or $VIM_{party}$. Although $VIM_{rawperm-CF}$ was developed to deal under correlation situations, it showed to be inefficient in predicting true associations of correlated variables.

Hence, this simulation study shows that one should be aware about the characteristics of the data such as correlation between predictors; and when using a particular software which is the default VIM in order to choose the right measure (changing it if it is necessary) and avoid spurious results because of correlation. In the chapter four, a real study is investigated using RF based on the $VIM_{rawperm-RF}$; $VIM_{rawperm-RF}$ was chosen due to the computational constraints (mostly timing).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

2.5. **Application: No significant epistasis in a 29 biomarkers pathway in PGC2**

2.5.1. **Data Extraction**

I performed the study to test for risk of schizophrenia, looking for both single effect and interaction effects (epistasis), in the PGC 2 case status data (Schizophrenia Working Group of the Psychiatric Genomics Consortium). I used genotyped information from 39 different European ancestry cohorts. Genotypes were imputed using IMPUTE2/SHAPEIT software (Howie et al. 2011), taking as a reference the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015); (Ripke et al. 2014). For quality control, the following criteria were considered: autosomal heterozygosity deviation between 0.2 and 0.8; before sample removal SNP missingness < 0.05 and subject missingness < 0.02; after sample removal SNP missingness < 0.02, and between cases and controls a difference in SNP missingness < 0.02; Hardy-Weinberg equilibrium (HWE) for the SNPs (p-value > 10^{-10} in cases or p-value > 10^{-6} in controls).

The database consisted of 58,280 observations including 26,476 cases and 31,804 controls. Table 2.20 illustrates the number of people with schizophrenia, healthy individuals and the total people by study.

<table>
<thead>
<tr>
<th>TARGET DATASETS</th>
<th>TARGET SOURCE</th>
<th>NCASES</th>
<th>NCONTROLS</th>
<th>NTOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABER</td>
<td>UK (Aberdeen)</td>
<td>719</td>
<td>697</td>
<td>1416</td>
</tr>
<tr>
<td>AJSZ</td>
<td>Israel (Lencz/Darvasi Sample)</td>
<td>894</td>
<td>1594</td>
<td>2488</td>
</tr>
<tr>
<td>ASRB</td>
<td>Australia</td>
<td>456</td>
<td>287</td>
<td>743</td>
</tr>
<tr>
<td>BULS</td>
<td>Bulgaria (case control)</td>
<td>195</td>
<td>608</td>
<td>803</td>
</tr>
<tr>
<td>BUTR</td>
<td>Bulgaria (trios)</td>
<td>608</td>
<td>613</td>
<td>1221</td>
</tr>
<tr>
<td>CATI</td>
<td>US (CATIE)</td>
<td>397</td>
<td>203</td>
<td>600</td>
</tr>
<tr>
<td>CAWS</td>
<td>UK (Cardiff)</td>
<td>396</td>
<td>284</td>
<td>680</td>
</tr>
</tbody>
</table>
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>CIMS</th>
<th>US (Boston, CIDR)</th>
<th>67</th>
<th>65</th>
<th>132</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLM2</td>
<td>UK (CLOZUK)</td>
<td>3426</td>
<td>4085</td>
<td>7511</td>
</tr>
<tr>
<td>CLO3</td>
<td>UK (CLOZUK)</td>
<td>2105</td>
<td>1975</td>
<td>4080</td>
</tr>
<tr>
<td>COU3</td>
<td>Cardiff, UK (CogUK)</td>
<td>530</td>
<td>678</td>
<td>1208</td>
</tr>
<tr>
<td>DENM</td>
<td>Denmark (Werge Sample)</td>
<td>471</td>
<td>456</td>
<td>927</td>
</tr>
<tr>
<td>DUBL</td>
<td>Ireland (Corvin Sample)</td>
<td>264</td>
<td>839</td>
<td>1103</td>
</tr>
<tr>
<td>EDIN</td>
<td>UK (Edinburgh)</td>
<td>367</td>
<td>284</td>
<td>651</td>
</tr>
<tr>
<td>EGCU</td>
<td>Estonia (EGCUT)</td>
<td>234</td>
<td>1152</td>
<td>1386</td>
</tr>
<tr>
<td>ERSW</td>
<td>Sweden (Hubin)</td>
<td>265</td>
<td>319</td>
<td>584</td>
</tr>
<tr>
<td>GRAS</td>
<td>Germany (GRAS)</td>
<td>1067</td>
<td>1169</td>
<td>2236</td>
</tr>
<tr>
<td>IRWT</td>
<td>Ireland (WTCCC2)</td>
<td>1291</td>
<td>1006</td>
<td>2297</td>
</tr>
<tr>
<td>LACW</td>
<td>Six Countries/WTCCC controls</td>
<td>157</td>
<td>245</td>
<td>402</td>
</tr>
<tr>
<td>LEMU</td>
<td>Six Countries-trios</td>
<td>197</td>
<td>177</td>
<td>374</td>
</tr>
<tr>
<td>LIE2</td>
<td>US (NIMH CBDB)</td>
<td>133</td>
<td>269</td>
<td>402</td>
</tr>
<tr>
<td>LIE5</td>
<td>US (NIMH CBDB)</td>
<td>497</td>
<td>389</td>
<td>886</td>
</tr>
<tr>
<td>MGS2</td>
<td>US, Australia (MGS)</td>
<td>2638</td>
<td>2482</td>
<td>5120</td>
</tr>
<tr>
<td>MSAF</td>
<td>US (New York) and Israel</td>
<td>325</td>
<td>139</td>
<td>464</td>
</tr>
<tr>
<td>MUNC</td>
<td>Germany (Munich)</td>
<td>421</td>
<td>312</td>
<td>733</td>
</tr>
<tr>
<td>PEWB</td>
<td>Seven countries (PEIC, WTCCC2)</td>
<td>574</td>
<td>1812</td>
<td>2386</td>
</tr>
<tr>
<td>PEWS</td>
<td>Spain (PEIC, WTCCC2)</td>
<td>150</td>
<td>236</td>
<td>386</td>
</tr>
<tr>
<td>PORT</td>
<td>Portugal</td>
<td>346</td>
<td>215</td>
<td>561</td>
</tr>
<tr>
<td>S234</td>
<td>Sweden 2,3,4</td>
<td>1980</td>
<td>2274</td>
<td>4254</td>
</tr>
<tr>
<td>SWE1</td>
<td>Sweden 1</td>
<td>215</td>
<td>210</td>
<td>425</td>
</tr>
<tr>
<td>SWE5</td>
<td>Sweden 5</td>
<td>1764</td>
<td>2581</td>
<td>4345</td>
</tr>
<tr>
<td>SWE6</td>
<td>Sweden 6</td>
<td>975</td>
<td>1145</td>
<td>2120</td>
</tr>
<tr>
<td>TOP8</td>
<td>Norway (TOP)</td>
<td>377</td>
<td>403</td>
<td>780</td>
</tr>
<tr>
<td>UCLA</td>
<td>Netherlands (Ophoff)</td>
<td>700</td>
<td>607</td>
<td>1307</td>
</tr>
<tr>
<td>UCLO</td>
<td>UK (London)</td>
<td>509</td>
<td>485</td>
<td>994</td>
</tr>
<tr>
<td>UKTR</td>
<td>UK (trios)</td>
<td>42</td>
<td>38</td>
<td>80</td>
</tr>
</tbody>
</table>

Performance of variable importance measures in Random Forest under correlation and application in PGC2
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>UMEB</th>
<th>Sweden (Umeå)</th>
<th>341</th>
<th>577</th>
<th>918</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMES</td>
<td>Sweden (Umeå)</td>
<td>193</td>
<td>704</td>
<td>897</td>
</tr>
<tr>
<td>ZHH1</td>
<td>US (New York)</td>
<td>190</td>
<td>190</td>
<td>380</td>
</tr>
<tr>
<td>TOTAL (ALL STUDIES)</td>
<td>Total (All studies)</td>
<td>26476</td>
<td>31804</td>
<td>58280</td>
</tr>
</tbody>
</table>

Table 2.20. Sample size for all 39 cohorts and the number of cases and controls

2.5.2. Pathway

The study was based on 29 molecular biomarkers that Chan et al. (2015) have shown to relate to schizophrenia. I took the genes related to the biomarkers (looking for the genes related with the analyte on the genecards website (http://www.genecards.org/)) and I included them in the study, they are shown in the Table 2.21 with biomarker information.

Gene boundaries were defined at the start and the end position of the gene transcript. After extracting the SNPs from the different genes on the pathway (using biomaRt in R) from all 39 studies, I ended up with 180 SNPs. The significance of these SNPs and the interaction between them with schizophrenia was the main aim to be investigated.

<table>
<thead>
<tr>
<th>MOLECULAR FUNCTION</th>
<th>ANALYTE</th>
<th>GENES</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIPID TRANSPORT</td>
<td>Apolipoprotein H</td>
<td>APOH</td>
</tr>
<tr>
<td></td>
<td>Apolipoprotein A1</td>
<td>APOA1</td>
</tr>
<tr>
<td>INFLAMMATORY RESPONSE</td>
<td>Macrophage migration inhibitory factor</td>
<td>MIF</td>
</tr>
<tr>
<td></td>
<td>Carcinoembryonic antigen</td>
<td>CEACAM5</td>
</tr>
<tr>
<td></td>
<td>Tenascin C</td>
<td>TNC</td>
</tr>
<tr>
<td></td>
<td>Interleukin-10</td>
<td>IL10</td>
</tr>
<tr>
<td></td>
<td>Interleukin-1 receptor antagonist</td>
<td>IL1RN</td>
</tr>
<tr>
<td></td>
<td>Receptor for advanced glycosylation end products</td>
<td>AGER</td>
</tr>
<tr>
<td></td>
<td>Interleukin-8</td>
<td>CXCL8</td>
</tr>
</tbody>
</table>
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>IMMUNE SYSTEM</th>
<th>Haptoglobin</th>
<th>HMGCS1</th>
</tr>
</thead>
<tbody>
<tr>
<td>von Willebrand factor</td>
<td>VWF</td>
<td></td>
</tr>
<tr>
<td>Alpha-2 macroglobulin</td>
<td>A2M</td>
<td></td>
</tr>
<tr>
<td>Beta-2 microglobulin</td>
<td>B2M</td>
<td></td>
</tr>
<tr>
<td>Serum glutamic oxaloacetic transaminase</td>
<td>GOT1</td>
<td></td>
</tr>
<tr>
<td>Interleukin-13</td>
<td>IL13</td>
<td></td>
</tr>
<tr>
<td>Immunoglobulin A</td>
<td>CD79A</td>
<td></td>
</tr>
<tr>
<td>HORMONAL SIGNALLING</td>
<td>Pancreatic polypeptide</td>
<td>PPY</td>
</tr>
<tr>
<td></td>
<td>Leptin</td>
<td>LEPTIN</td>
</tr>
<tr>
<td></td>
<td>Testosterone (total)</td>
<td>STAR,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACAT2,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AACS</td>
</tr>
<tr>
<td>GROWTH FACTOR SIGNALLING</td>
<td>Follicle-stimulating hormone</td>
<td>FSHB</td>
</tr>
<tr>
<td></td>
<td>Thyroid-stimulating hormone</td>
<td>TSHB</td>
</tr>
<tr>
<td></td>
<td>Insulin-like growth factor-binding protein 2</td>
<td>IGFBP2</td>
</tr>
<tr>
<td></td>
<td>AXL receptor tyrosine kinase</td>
<td>AXL</td>
</tr>
<tr>
<td></td>
<td>Stem cell factor</td>
<td>KITLG</td>
</tr>
<tr>
<td>CLOTTING CASCADE</td>
<td>Factor VII</td>
<td>F7</td>
</tr>
<tr>
<td></td>
<td>Angiotensin-converting enzyme</td>
<td>ACE</td>
</tr>
<tr>
<td>HORMONAL SIGNALLING</td>
<td>Chromogranin-Aa</td>
<td>CGA</td>
</tr>
<tr>
<td>GROWTH FACTOR SIGNALLING</td>
<td>Vascular cell adhesion molecule-1a</td>
<td>VCAM-1</td>
</tr>
<tr>
<td>INFLAMMATORY RESPONSE</td>
<td>Eotaxina</td>
<td>CCL11</td>
</tr>
</tbody>
</table>

Table 2.21. Molecular function of the 29 biomarkers and the Genes selected for the study

Performance of variable importance measures in Random Forest under correlation and application in PGC2
2.5.3. Population Stratification

Because of genetic variation between the cohorts due to ancestry, I had to correct for PS to avoid spurious results that the genetic variation may cause in the analysis, as in Zhao et al. (2012). Following the original PGC2 analysis (Ripke et al. 2014), I used the following principal components (PC): PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC9, PC15 and PC18. After performing the PC analysis with a collection of 19,551 autosomal SNPs across all 49 European ancestry studies, Ripke et al. (2014) took the first 20 PCs and they tested their association with schizophrenia applying logistic regression including the studies as dummy variables (study indicator) and the PCs as covariates. The optimal set of PCs selected was the one formed by the 10 principal components cited above.

Therefore, using those 10 PCs, I extracted the residuals from general linear regression models where phenotypes and genotypes were regressed out, and the PCs and studies were considered as independent variables (Zhao et al. 2012). These residuals (continuous variables) were the new variables to study in the schizophrenia risk analysis.

2.5.4. Leave-One-Out Cross-Validation Across 39 Studies

The study design consisted of 39 training dependent datasets and one independent dataset for each training set for replication (Figure 2.19). Training sets included all cohorts except for one study, these single cohorts were considered as the independent test datasets. Therefore, I tested for single and interaction effects using RF in 39 training sets (filtering the amount of variables that might be related with schizophrenia in the training sets). Then, I used LRTs of nested models to test for epistasis, and linear regression models to test for significance of single SNPs in the independent test sets.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 2.19. Example of leave-one-out cross validation in a study design with 5 cohorts, instead with 39.

2.5.5. Random Forest

According to the simulation results, conditional permutation VIMs and the scaled VIMs are not appropriate for use with correlated variables. Among the unconditional
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

PVIMs, I chose the unscaled PVIM, VIM_{rawperm-RF}, to perform RF due to the computational time (on the simulation from the above section, it was more 10 times faster than VIM_{AUC} and VIM_{party} for only one dataset using a computational cluster). I set the number of trees to 1000, the mtry equal to 65 (larger than the number of SNPs divided by 3 default in regression), and the percentage of subsampling of observations for tree-growing was fixed to 63.2.

Under these conditions, I ran RF 100 times over the training datasets, each time changing the random number seed, in order to obtain stable estimates of the VIMs, and also over null data, created after permuting the phenotype. Once the empirical p-value was calculated, I took the empirically significant predictors over the 100 iterations and in each of the 39 training samples. Then, the empirically-significant SNPs from each training dataset were considered on the respective test dataset to try to replicate single effects and epistasis between them. As the empirically significant SNPs were not the same in all training sets (but several SNPs were significant in more than one), the tests in the different independent datasets were not all the same.

2.5.6. Likelihood Ratio Tests (LRTs) between nested models

To detect epistasis between the significant SNPs, I applied LRTs between nested models based on linear regression. The nested models were performed as follows:

**Full model:**  \( Y \sim \beta_1 \text{SNPi} + \beta_2 \text{SNPj} + \beta_3 \text{SNPi} * \text{SNPj} \)

**Reduced Model:**  \( Y \sim \beta_1 \text{SNPi} + \beta_2 \text{SNPj} \)

Where

\( \text{SNPi} , \text{SNPj} \) were empirical significant SNPs (the residuals from PS) from all RF iterations in all the 39 training samples.

\( Y \) is the phenotype (residual from PS).
I performed the analyses with *randomjungle* Centos 64 Bit Version (Build 2.0.0) (Schwarz, König and Ziegler, 2010) and in *R* version 3.0.0 and used the package *lmtest* (Zeileis and Hothorn 2002) for calculating the LRT. Also, for calculating the combined *p*-value of the SNPs from the independent test sets considering both the coefficient direction and the sample size, I used *MetaP* (*a program to combine *p*-values*; Whitlock, 2005).

![Figure 2.20. Illustration of the approach taken on the applied study in section 2.5. This illustration shows the methods taken in the study and the studies which were combined to report the final results (only the effects which were tested in all independent datasets were combined).](image-url)
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

2.5.7. Results

After calculating the empirical $p$-values over all RF iterations based on VIM$_{rawperm-RF}$ in each of the 39 training datasets, the single and interactions effects from the empirically significant SNPs (empirical $p$-value < 0.05) in each training dataset were tested in the left out dataset. I applied general linear models to check if their single effect was significant, and nested models to test for epistasis, on the 39 independent test datasets.

After the test for the effects in the independent test, two SNPs were tested in only one independent test set, one was significant before Bonferroni correction, but it did not pass Bonferroni correction. The number of interactions that were involved in only one independent test set was 377, of which 12 had a $p$-value < 0.05, but none passed Bonferroni correction. The Bonferroni correction threshold for each independent dataset was different, as the number of empirically significant SNPs was not the same in all training datasets, the $p$-values ranged from 0.0007 to 0.0012 in the single association study, and from 0.000023 to 0.000067 in the interaction study.

I noticed that nine SNPs were tested (overlapping in the analysis) in all independent datasets because they were empirically significant in each dataset. These SNPs showed a $p$-value < 0.05 in at least one independent test, so these 9 SNPs were considered for combining their $p$-values from left out independent tests. The combined $p$-values of the 9 SNPs were calculated considering the weights (sample size) and the direction of the coefficients using MetaP.

The results suggested seven significant SNPs, which mapped to two different genes, five in $ACAT2$ and two in $TNC$. The five significant SNPs in $ACAT2$ are in LD with $r^2 = 0.977$ and $D' = 1$ between each other and the two SNPs in $TNC$ are completely in LD $r^2 = 1$ and $D' = 1$. Table 2.22 shows the results for the single effects of the most significant independent SNPs.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>SNP</th>
<th>CHR</th>
<th>GENE</th>
<th>P-VALUE</th>
<th>%R^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs3798211</td>
<td>6</td>
<td>ACAT2</td>
<td>2.62 x 10^{-5}</td>
<td>[0.0023%,0.7%]</td>
</tr>
<tr>
<td>rs3789875</td>
<td>9</td>
<td>TNC</td>
<td>0.0004</td>
<td>[0.0001%,2.5%]</td>
</tr>
</tbody>
</table>

Table 2.22. Results for the most independent SNPs. The combined p-value is across all 39 independent datasets, taking into account the effect direction and the sample size. The range of the variance explained in percentage (%R^2) is also across all 39 independent datasets.

The interaction between the two independent significant SNPs (in single effects) was also tested, but the interaction did not show a statistically significant effect (p-value > 0.05). Due to computational constraints, in addition to the reason that this study was intended as an example of how to used RF in real applications, combining p-values for single and interaction effects from SNPs that were tested in less than the 39 independent tests was not possible.

2.5.8. Discussion

There has been previous research looking at genetic risk for schizophrenia. Ripke et al. (2014) found 108 biomarkers, which were related to schizophrenia. The present study suggested two SNPs which were statistically significantly associated with schizophrenia, using RF to filter SNPs (select a subset) in the training datasets.

Acetyl-coenzyme A acetyltransferase 2 (ACAT2) on chromosome 6, is an enzyme involved in lipid biosynthesis. In a combined transcriptomics, proteomics and metabolomics approach studying post mortem samples from people with schizophrenia several altered metabolic pathways were identified, including lipid metabolism and the gene ACAT2 (Prabakaran et al. 2004). Many genes involved in lipid metabolism, especially cholesterol biosynthesis such as ACAT2, are tightly regulated by myelin related genes. These results are consistent with the finding that there are greater perturbations in white matter (which is composed of bundles of
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

myelinated axons) compared to grey matter in schizophrenia patients (Prabakaran et al. 2004).

The novel finding from the present study was at gene Tenascin C (TNC) which has a number of diverse functions which can modulate cell behaviour directly and indirectly (Ghert et al. 2001). It is expressed in a number of tissues and organs and also in a number of malignancies (Breliere and Chiquet-Ehrismann 2012; (Yang et al. 2016). This extracellular matrix protein interacts with integrins, thus modulating adhesion, and in the brain it is expressed throughout the white matter of rostral brain segments, where it acts as a guidance cue to migrating neurons and axons during development and regeneration (Rettig et al. 1989).

One limitation of the study was the computational constraints because one has to import manually the values on MetaP. All single SNPs and interactions tested in less than 39 studies (SNPs which were not empirically statistically significant in all training datasets, so were not tested in all independent test) were not considered to combine the p-values from all studies they were tested taking into account the sample size and the direction of the coefficients for that reason. As an example, SNPs which were empirically statistically significant in 38 training sets, were tested in the 38 left out independent test (single and interaction effects), but the combined p-values were not calculated. This might hide single and interaction effects that influence risk for schizophrenia. In addition, the small number of SNPs involved in the study might be the reason for the absence of more significant single and interaction effects.

Other methods for combining p-values could have been considered such as Fisher’s method (sum of logs method), Edgington’s method (sum of p method), the sum of z method and the Stoffzers weighted. These methods for combining p-values are available in R in the package metap, but none of them considered the direction of the coefficients in each study which is important to ensure significance when combining p-values. For instance, if in two studies the effects from a single SNP or from an interaction have p-values less than 0.05 but the effect in one study is positive and in
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

the other one is negative, these methods would show a combined $p$-value less than 0.05 because they do not take into account the direction of the coefficients, when actually the combined $p$-value should be greater than 0.05 (different direction). Unfortunately, the package does not consider the direction of the coefficients with any of those methods, which was the reason why MetaP was used.

The lack of significant interactions may be due to fact explained on the above paragraph as well as because the LD between SNPs. Also, the pathway of this study was focused on biomarkers that have shown a single effect in the original study, and it was not focused on proteins that have shown biological interactions (Chan et al. 2015).
Bias of Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors

3. Bias of Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors

3.1. Introduction

RF based on the Gini VIM (VIM\textsubscript{Gini}) is one of the most popular RF VIMs. In fact, it is the default VIM when applying RF in popular programming languages such as Python and R. For instance, the function RandomForestClassifier in the sklearn.ensemble library for Python (Pedregosa \textit{et al.} 2011) performs VIM\textsubscript{Gini} by default. Furthermore, the well-known RandomForest package (Liaw and Wiener 2002) in R calculates the importance scores also by defaulting the same VIM.

However, VIM\textsubscript{Gini} has been previously shown to be biased. VIM\textsubscript{Gini} presents several different sources of bias: (1) under predictor correlation, as shown in Chapter 2; (2) under predictors with different number of categories; and (3) under predictors with the same number of categories but with different class size (Strobl \textit{et al.} 2007b); (Nicodemus and Malley 2009); (Nicodemus 2011); (Boulesteix \textit{et al.} 2012a)).

With regard to the second bias, Strobl \textit{et al} (2007b) showed that VIM\textsubscript{Gini} gives larger scores to predictors with more categories and to continuous predictors (but only one continuous predictor was considered for the study). Even when there is no association with the outcome (under H\textsubscript{0}), predictors with more categories are selected more often early in the trees, due to having more chances to yield a good split (for each variable, for every value of each variable, which is a possible threshold, the reduction in impurity is calculated; variables with more categories have more values and, therefore, a higher chance to be selected).

When considering the third bias, Nicodemus (2011) showed that even when categorical variables had the same number of categories, such as in GWAS where all SNPs have three categories (homozygous minor allele, heterozygous or homozygous
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

major allele), "$V_{Gini}$" preferred predictors with large category frequencies (with large MAF). In addition, Boulesteix et al. (2012a) showed that SNPs with large MAF were favoured by "$V_{Gini}$" under H_0. Under H_A, SNPs with larger MAF showed higher "$V_{Gini}$" scores than influential SNPs with lower MAF. Moreover, this preference did not disappear with a larger sample size (the largest sample size considered under study was 10,000).

Boulesteix et al. (2012a) suggested that "$V_{Gini}$" might be preferred in real studies where all predictors are continuous and uncorrelated between each other, as well as when there signal-to-noise ratio is low. However, these suggestions have not been studied in depth. Strobl et al. (2007b) included only one continuous predictor in the study, and both Nicodemus (2011) and Boulesteix et al. (2012a) only considered categorical variables.

### 3.1.1. Aims

First, based on the suggestion proposed by Boulesteix et al. (2012b), and the bias towards predictors with more categories found by Strobl et al. (2007b), I examined the performance of "$V_{Gini}$" when all predictors followed a normal distribution (all continuous) but with different variances, and were also independent of each other. If "$V_{Gini}$" was inflated because of variability of predictors and not because of their actual association, this would be an important fact that researchers should take into account when applying RF in real situations to avoid spurious results. In addition, in this study "$V_{Gini}$" was investigated when all variables followed the same distribution while considering different precision (different number of decimal places, which determines the number of unique values).

Second, one of the options to have a low signal-to-noise ratio happens when it exists larger noise than signal. However, noise should not affect the behaviour of "$V_{Gini}$". So, it may happen that more noise has an impact on the "$V_{Gini}$" and, therefore, inflates the importance scores due to noise rather than association. In real situations, the error
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

(noise) cannot be distinguished or quantified, but it is still important to investigate whether it has an impact on RF based on VIM$_\text{Gini}$, although a large variation in error might not occur in real applications. Thus, to check the presence of an extra source of bias due to error, I created different synthetic datasets, which examined two different variances of the error under both null and alternative hypotheses.

### 3.2. Methods

#### 3.2.1. Data based on normal distributed variables with different variances

To study if VIM$_\text{Gini}$ prefers the predictors because of their variability and not because of their actual signal, I first examined what happens when no predictor is influential (under H$_0$). Second, I checked under association (H$_A$) whether VIM$_\text{Gini}$ gives larger scores for predictors with higher variability when all of them actually have the same effect size. Both under H$_A$ and under H$_0$, I simulated 500 datasets for each condition under study. Moreover, the VIM was studied with two types of outcomes, binary and continuous.

#### 3.2.1.1. Data simulation under H$_0$

##### 3.2.1.1.1. All variables follow a standard normal distribution

##### 3.2.1.1.1.1. Continuous outcome

Under H$_0$, the outcome only depends on the error. I created the data as follows:

Let $X = [x_{ij}] \sim N(0, \sum_1)$ where dim$(X) = 1000 \times 10$, $\sum_1 = I$ (identity matrix) 

$$ \text{corr}(x_j, x_k)_{j \neq k} = 0. $$

$$ y = e_1 $$

where $e_1 \sim N(0,1)$ (annotation $N(\mu,\sigma)$).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3.2.1.1.2. Binary outcome

To generate the databases when the outcome has only two values (1 or 0), it was necessary to transform the continuous variable into a binary variable.

From the linear model \( y = \beta X + e_1 \), let us define \( y \) as follows (\( \beta = 0 \)):

\[
y_{bin} = \begin{cases} 
1 & \text{when } e_1 > 0 \\
0 & \text{Otherwise}
\end{cases}
\]

As the error follows a normal distribution, the outcome variable is generated from a probit model.

3.2.1.1.2. All variables follow a normal distribution with different variances

3.2.1.1.2.1. Continuous outcome

Let \( Z = [z_{ij}] \sim N(0, \Sigma_2) \) where \( \dim(Z) = 1000 \times 10 \), \( \Sigma_2 = \text{diag}(50,45,40,35,30,25,20,15,10,1) \) (\( \text{corr}(z_j, z_k)_{j \neq k} = 0 \)); and let the outcome \( y \) be as follows:

\[ y = e_1 \]

where \( e_1 \sim N(0,1) \) (annotation \( N(\mu,\sigma) \)).

3.2.1.1.2.2. Binary outcome

As in 3.2.1.1.1.2., except that predictor variance matrix was changed (as in the continuous case).

3.2.1.2. Data simulation under \( H_A \)

Under \( H_A \), I created 500 synthetic datasets for each individual association study in \( R \). There were 10 single association studies, in each study only one predictor was influential over 10 predictors in the database, and the influential predictor was different.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

in each study. Then, in each association study 500 databases were considered. All the 10 predictors followed a normal distribution (continuous) in all studies.

As under $H_0$, there were two alternative conditions, one where the variance of the 10 predictors was the same (all followed a standard normal distribution), another when the variance of the predictors was different. The effect size was fixed to be equal in all association studies in order to have the same impact from the different predictors. Therefore, we simulated 500 databases for 20 different studies.

3.2.1.2.1. **All variables follow a standard normal distribution**

3.2.1.2.1.1. **Continuous outcome**

Let $X = [x_{ij}] \sim N(0, \Sigma_1)$ where $\dim(X) = 1000 \times 10$, $\Sigma_1 = I$ (identity matrix) 

$\text{corr}(x_j, x_k)_{j \neq k} = 0$

$$y_j = 0.3 \times x_j + e_1$$

where $e_1 \sim N(0,1)$ (annotation $N(\mu, \sigma)$); $x_j$ is one of the ten predictors, and $j = 1, \ldots, 10$. Thus, there was one association study for each $j$. The ten associations were performed to be consistent across studies, although all predictors are generated following the same distribution. This datasets were also created to compare the performance of VIM Gini when all datasets have standardised predictors with the same precision to when predictors have different variances, or when predictors have different precision, or to when the error has less variance.

3.2.1.2.1.2. **Binary outcome**

The predictor matrix and the error followed the same distribution and pattern as for the continuous outcome. Therefore, the outcome was modelled under $H_A$ as:

$$y_{binj} = \begin{cases} 1 & \text{when } 0.3 \times x_j + e_1 > 0 \\ 0 & \text{Otherwise} \end{cases}$$
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

where there was one outcome for each $x_j$ associated over the ten predictors ($j = 1, \ldots, 10$).

### 3.2.1.2.2. All variables follow a normal distribution with different variances

#### 3.2.1.2.2.1. Continuous outcome

Let $Z = [z_{ij}] \sim N(0, \Sigma_2)$ where $\dim(Z) = 1000 \times 10$, $\Sigma_2 = \text{diag}(50,45,40,35,30,25,20,15,10,1)$ (corr($z_j$, $z_k$)$_{j\neq k} = 0$)

$$y_j = 0.3 \times z_j + e_1$$

where $e_1 \sim N(0,1)$; $z_j$ is one of the ten predictors, and $j = 1, \ldots, 10$. Also in this case, there was one association study for each $j$.

#### 3.2.1.2.2.2. Binary outcome

The outcome was generated as in 3.2.1.2.1.2. but the databases were changed as the predictor matrix was constituted by normal distributed variables, each one with different variance, as in the continuous case.

### 3.2.2. Data based on normal distributed predictors with different cut-points

Based on Strobl et al. (2007b), the present study also investigated the performance of VIM$_{\text{Gini}}$ when all predictors followed a standard normal distribution but with different numbers of cut-points (different precision). Five hundred different datasets were generated considering all variables with a different cut-point for each situation. The first variable ($X_1$) was rounded to one decimal place, the second one ($X_2$) with two decimal places and so on, the last variable ($X_{10}$) was rounded with 10 decimal places. VIM$_{\text{Gini}}$ was applied under both $H_0$ and $H_A$. The results from this subsection were compared to the ones from the subsection 3.2.1.1.1. for both the continuous outcome
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

and the binary outcome, when all variables followed a standard normal with the same number of cut-points and with an error following a standard normal.

3.2.2.1. Data simulation under $H_0$

3.2.2.1.1. Continuous outcome

Under $H_0$, 500 datasets were generated without predictor association:

Let $X = [x_{ij}] \sim N(0, \Sigma_1)$ where $\text{dim}(X) = 1000 \times 10$, $\Sigma_1 = I$ (identity matrix)

$(\text{corr}(x_j, x_k))_{j \neq k} = 0$

$y = e_1$

where $e_1 \sim N(0,1)$, $X_1$ has 1 decimal place, $X_2$ has 2 decimal places, $X_{10}$ has 10 decimal places.

3.2.2.1.2. Binary outcome

As in 3.2.1.1.2.:

$$y_{bin} = \begin{cases} 1 & \text{when } e_1 > 0 \\ 0 & \text{Otherwise} \end{cases}$$

the outcome does not depend on the predictors, which have different number of decimal places.

3.2.2.2. Data simulation under $H_A$

Five hundred synthetic databases were considered for each individual association study, one for each influential predictor. As before (subsection 3.2.1.2.), in each of the ten association studies all predictors have the same conditions as well as the error, and the coefficients of the linear generating models were fixed to be equal.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

### 3.2.2.2.1. All variables follow a standard normal distribution with different cut-points

#### 3.2.2.2.1.1. Continuous outcome

The ten different models were defined as follows:

Let $X = \begin{bmatrix} x_{ij} \end{bmatrix} \sim N(0, \Sigma_1)$ where $\dim(X) = 1000 \times 10, \Sigma_1 = I$ (identity matrix) 
\[
\text{corr}(x_j, x_k)_{j \neq k} = 0
\]

\[ y_j = 0.3 \times x_j + e_1 \]

where $e_1 \sim N(0,1)$; $x_j$ is one of the ten predictors, and $j = 1, ..., 10$. In each association study, $X_1$ was rounded with 1 decimal place, $X_2$ with 2 decimal places, ..., $X_{10}$ with 10 decimal places.

#### 3.2.2.2.1.2. Binary outcome

The different databases for each association study considered 10 predictors following a standard normal distribution, each one rounded with a different number of decimal places, and the outcome generated as follows:

\[ y_{\text{bin}, j} = \begin{cases} 1 & \text{when } 0.3 \times x_j + e_1 > 0 \\ 0 & \text{Otherwise} \end{cases} \]

where $e_1 \sim N(0,1)$. One outcome for each $x_j$ associated variable with $j = 1, ..., 10$.

### 3.2.3. Data based on normal distributed errors with different variance

#### 3.2.3.1. Data simulation under $H_0$

In this subsection, I investigated whether the variance of the error has an impact on the $\text{VIM}_{\text{Gini}}$ when there is no association. In the subsections above, I simulated different
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

databases using models under $H_0$ where the error followed a standard normal distribution. Here, I generated the models, while considering an error following a normal distribution with the same mean but with lower variance (less than 1). In this way I could compare VIM$_{Gini}$ behaviour under $H_0$ to the situation in subsection 3.2.1.1.1.

3.2.3.1.1. All variables follow a standard normal distribution

3.2.3.1.1.1. Continuous outcome

Let $X = [x_{ij}] \sim N(0, \Sigma_1)$ where $\dim(X) = 1000\times10$, $\Sigma_1 = I$ (identity matrix) 
$(\text{corr}(x_j, x_k)_{j\neq k}=0)$

$$y = e_2$$

where $e_2 \sim N(0,0.5)$ (annotation $N(\mu,\sigma)$).

3.2.3.1.1.2. Binary outcome

The predictors followed a standard normal distribution, but the error followed a normal distribution with variance 0.25 (standard deviation 0.5). So, the outcome under the null was then generated as follows:

$$y_{bin} = \begin{cases} 1 & \text{when } e_2 > 0 \\ 0 & \text{Otherwise} \end{cases}$$

3.2.3.2. Data simulation under $H_A$

In this subsection, the models under the alternative hypothesis were illustrated. Five hundred datasets were generated for each individual association study out of a total of 10 as in subsection 3.2.1.2., but in this case with different error variance. The coefficients of the linear generating models were always the same.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3.2.3.2.1. All variables follow a standard normal distribution

3.2.3.2.1.1. Continuous outcome

Let $X = [x_{ij}] \sim N(0, \Sigma_1)$ where $\text{dim}(X) = 1000 \times 10$, $\Sigma_1 = I$ (identity matrix) (corr($x_j$, $x_k$)_{j\neq k}=0)

$$y_j = 0.3 \times x_j + e_2$$

where $e_2 \sim N(0,0.5)$ (annotation N($\mu, \sigma$)); $x_j$ is one of the ten predictors, and $j = 1, \ldots, 10$. The variance of the predictor was different, but the association was the same, and $e_2$ had less variance than in subsection 3.2.1.2.1.1.

3.2.3.2.1.2. Binary outcome

For each of the 10 association studies, the outcome was generated as

$$y_{bin,j} = \begin{cases} 1 & \text{when } 0.3 \times x_j + e_2 > 0 \\ 0 & \text{Otherwise} \end{cases}$$

where $e_2 \sim N(0,0.5)$ (annotation N($\mu, \sigma$)); all predictors $x_j$ and the non-associated ones followed a standard normal distribution.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Bias of Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors.

Figure 3.1. Illustration of the data generation under $H_A$ in all different conditions when the outcome is continuous. The top one corresponds to when all predictors and the error follow a standard normal distribution. The three bottom conditions are, from left to right, when predictors follow a standard normal with different variances, when the predictors have different number of decimal places, and when the error variance is lower. The top one is going to be compared to each of the other three conditions in the approach in order to make conclusions.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>CONTINUOUS AND BINARY OUTCOME</th>
<th>UNDER H₀</th>
<th>UNDER Hₐ</th>
</tr>
</thead>
</table>
| $X_{10} \sim N(0, I), e_1 \sim N(0, 1)$  
All variables standard normal and same precision | 500 datasets in each of the four cases or studies  
Study of reference | 500 datasets in each association model (500 when $X_1$ is associated, ...  
Study of reference |
| $X_{10} \sim N(0, \Sigma_2), e_1 \sim N(0, 1)$  
Variables with different variance | 500 datasets in each of the four cases or studies | 500 datasets in each association model (500 when $X_1$ is associated, ...  
Study of reference |
| $X_{10} \sim N(0, I), e_1 \sim N(0, 1)$  
Variables with different precision | 500 datasets in each of the four cases or studies | 500 datasets in each association model (500 when $X_1$ is associated, ...  
Study of reference |
| $X_{10} \sim N(0, I), e_1 \sim N(0, 0.5)$  
Less error variance | 500 datasets in each of the four cases or studies | 500 datasets in each association model (500 when $X_1$ is associated, ...  
Study of reference |

Table 3.1. Summary of the approach taken under $H₀$ and $Hₐ$ in the four different conditions. The case when all variables follow a standard distribution with same variance and same precision and when the error follow a standard normal is used as the reference to compared to the other three different conditions.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3.2.4. Random Forest based on VIM\textsubscript{Gini} simulation.

For details of RF method and how VIM\textsubscript{Gini} is defined see the Methods sections of Chapter 2. The performance of RF based on VIM\textsubscript{Gini} was investigated under the H\textsubscript{A} in 10 different association studies, each study depending only on one of the ten predictors; and all of them considered the same coefficient (impact signal) in the generating models for both outcomes, continuous and binary. In addition, I examined the performance of VIM\textsubscript{Gini} in the absence of association under the null hypothesis (H\textsubscript{0}) for both outcomes. Under both H\textsubscript{0} and H\textsubscript{A}, predictors with and without different variance, with and without different cut-points, as well as errors with different variance were studied. RF based on VIM\textsubscript{Gini} was applied to the different databases using randomjungle CentOS 64-Bit Version (Build 2.0.0) (Schwarz, König and Ziegler, 2010). To generate the different databases, the R package mtvnorm version 1.0-6 was used (Genz and Bretz, 2009). The package can be used to generate multivariate normal probabilities, quantiles, densities, and random deviates as well as for multivariate Student’s t distribution. Furthermore, to dichotomize the outcome, I used the function ra2ba of the R package bindata (Leisch et al. 2015), which considers the value 0 as the threshold.

To apply RF based in VIM\textsubscript{Gini}, I set up the values of the different RF parameters to be fixed across all implementations. All RF iterations built the Forest using subsampling, the number of trees was fixed to be $n_{\text{tree}} = 1000$, the $m_{\text{try}}$ (number of randomly chosen variables at each split) equal to the default value for both outcomes, continuous (number of total predictors divided by three) and binary (square root of the total number of predictors). The default $m_{\text{try}}$ was chosen since the number of noise variables was not large. Note that the default in randomjungle is the square root of the total number of variables, but as the databases have 10 predictors, both values rounded match ($m_{\text{try}}$ value is an integer).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3.3. Results

3.3.1. Bias, coverage and p-values

To know whether the synthetic databases were generated correctly, the bias, coverage and p-values were extracted from linear regression models when the outcome was continuous, and from logistic regression models with a probit link when the outcome was binary (since the error followed a normal distribution). Under $H_0$, the significance of the generated models was tested with LRTs of nested models: each full model was considered as the association of a single predictor and the reduced model as only the intercept. Under $H_A$ the full model was the truly-associated model, with only the single influential predictor in each of the ten studies.

In the binary studies, the outcome before transformation always followed a normal distribution with expected mean ($\mu$) = 0 (i.e. symmetric around zero). Thus, the number of cases and controls are expected to be similar. In fact, the median difference between the number of cases and controls in all conditions was observed to be 22.

3.3.1.1. Continuous outcome

Under the null distribution the bias was near 0 and the coverage around 95% in all different situations. When all predictors and the error followed a standard normal distribution (Table 3.2) the bias was between -0.0017 and 0.0024 values, the coverage ranged from 93.4% to 96.4%. The number of times the predictor was significant was low, the maximum value being 33 when the p-value considered was 0.05; after Bonferroni correction ($p$-value threshold 0.0001) only two predictors were detected as significant once each over the 500 models. The linear regression models could also reproduce the linear generating models when the predictors had differing variance ($\Sigma=\text{diag}(50,45,40,35,30,25,20,15,10,1)$, where $\Sigma$ was the variance-covariance matrix of the predictors) (Table 3.3). The bias ranged from -0.0029 to 0.0005, the coverage from 93.8% to 97%, and the number of times the full model was significant ranged from 22 to 15 times over the 500 models with a $p$-value < 0.05. No full model reached
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

statistical significance after Bonferroni correction. When the predictors had different cut-points, the bias and the coverage ranged from -0.0019 to 0.0017 and from 91.8% to 95.6% respectively (Table 3.4). After Bonferroni correction, only the single models from two different predictors became significant, but one time over the 500 models; considering a \( p \)-value < 0.05 all predictors were significant less than 36 times. Furthermore, when all predictors followed a standard normal but the error variance was lower, the estimated bias and coverage from the linear regression models were around 0 (from -0.0012 to 0.0017) and around 95% (92.8% from to 96.6%) respectively. From 17 to 36 models with \( p \)-value < 0.05, none with \( p \)-value < 0.0001 (Bonferroni) (Table 3.5).

<table>
<thead>
<tr>
<th>N(0,1)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>0.0023</td>
<td>-0.0017</td>
<td>0.0025</td>
<td>0.0007</td>
<td>0.00001</td>
<td>0.0012</td>
<td>-0.0002</td>
<td>0.0009</td>
<td>-0.0002</td>
<td>-0.0015</td>
</tr>
<tr>
<td>%COVER</td>
<td>93.4</td>
<td>95.6</td>
<td>96</td>
<td>94.6</td>
<td>95</td>
<td>96.4</td>
<td>96</td>
<td>94.4</td>
<td>93.8</td>
<td>94.4</td>
</tr>
<tr>
<td>P&lt;0.05</td>
<td>33</td>
<td>22</td>
<td>20</td>
<td>27</td>
<td>25</td>
<td>18</td>
<td>20</td>
<td>28</td>
<td>31</td>
<td>28</td>
</tr>
<tr>
<td>P&lt;0.0001</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.2. Bias, coverage, number of \( p \)-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under H0, when all predictors and the error followed a standard normal distribution. Continuous outcome.

<table>
<thead>
<tr>
<th>Diff variance</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>-0.0001</td>
<td>0.000003</td>
<td>0.0001</td>
<td>0.0005</td>
<td>-0.00004</td>
<td>-0.0004</td>
<td>-0.0002</td>
<td>0.0006</td>
<td>0.0002</td>
<td>-0.0029</td>
</tr>
<tr>
<td>%COVER</td>
<td>95</td>
<td>94.6</td>
<td>95</td>
<td>95.8</td>
<td>97</td>
<td>96</td>
<td>95.8</td>
<td>94.8</td>
<td>94.4</td>
<td>93.8</td>
</tr>
<tr>
<td>P&lt;0.05</td>
<td>25</td>
<td>27</td>
<td>25</td>
<td>21</td>
<td>15</td>
<td>20</td>
<td>21</td>
<td>26</td>
<td>28</td>
<td>31</td>
</tr>
<tr>
<td>P&lt;0.0001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.3. Bias, coverage, number of \( p \)-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under H0, when all predictors followed a normal distribution but with different amounts of variance. The error followed a standard normal distribution. Continuous outcome.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Table 3.4. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under H₀, when all predictors followed a standard normal distribution but with different number of decimal places. The error followed a standard normal distribution. Continuous outcome.

<table>
<thead>
<tr>
<th>cutpoints</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>0.0017</td>
<td>0.0016</td>
<td>0.0004</td>
<td>-0.0019</td>
<td>-0.0006</td>
<td>-0.001</td>
<td>-0.0014</td>
<td>0.0016</td>
<td>0.0012</td>
<td>0.0017</td>
</tr>
<tr>
<td>%COVER</td>
<td>93.8</td>
<td>94.4</td>
<td>91.8</td>
<td>95.4</td>
<td>93.4</td>
<td>95.6</td>
<td>95.4</td>
<td>92.8</td>
<td>95.2</td>
<td></td>
</tr>
<tr>
<td>p&lt;0.05</td>
<td>31</td>
<td>28</td>
<td>41</td>
<td>23</td>
<td>33</td>
<td>22</td>
<td>22</td>
<td>36</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>p&lt;0.0001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.5. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under H₀, when all predictors followed a standard normal distribution. The error followed a normal distribution with 0.5 standard deviation. Continuous outcome.

<table>
<thead>
<tr>
<th>N(0,0.5)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>-0.0011</td>
<td>-0.0003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>-0.00001</td>
<td>0.0017</td>
<td>0.0005</td>
<td>-0.0012</td>
<td>-0.0001</td>
<td>-0.0005</td>
</tr>
<tr>
<td>%COVER</td>
<td>93.8</td>
<td>95</td>
<td>95.4</td>
<td>95</td>
<td>94.2</td>
<td>95</td>
<td>94.6</td>
<td>96.6</td>
<td>96</td>
<td>92.8</td>
</tr>
<tr>
<td>p&lt;0.05</td>
<td>31</td>
<td>25</td>
<td>23</td>
<td>25</td>
<td>29</td>
<td>25</td>
<td>27</td>
<td>17</td>
<td>20</td>
<td>36</td>
</tr>
<tr>
<td>p&lt;0.0001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In addition, under Hₐ, the linear regression models were shown to reproduce the linear generating models - the bias was always around 0 and the coverage around 95%. All models were always significant (p-value < 0.05 as well as after Bonferroni correction), showing that the single associated models were not generated from a weak association. As Chapter 2 showed that weak association has an impact on the VIM, the models were generated with a strong association, but not as strong as the strongly associated situations of Chapter 2. In this way, if VIM₆ₙₖ has a particular behaviour in some of the conditions, it is due to the particular condition and not due to the strength of the association.

The bias ranged from -0.0054 to 0.001 when predictors and error followed a standard normal distribution (Table 3.6), from -0.0007 to 0.0012 when all predictors had different variance (Table 3.7), from -0.0012 to 0.0029 when all predictors had different cut-points (Table 3.8), and from -0.0016 to 0.0009 when the error variance was 0.25 (Table 3.9). The coverage ranged from 92.4% to 96% (Table 3.6) when predictors and error followed a standard normal, from 93.4% to 96.6% (Table 3.7) when each
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

predictor had different variance, and it was between 93.6% and 97.4% (Table 3.8), and between 93.6% and 96.8% (Table 3.9) when the predictors had different number of decimals (precision) and the variance of the error was 0.25 respectively.

### Table 3.6. Bias and coverage under $H_0$, when all predictors and the error followed a standard normal distribution. Continuous outcome. All models were significant before and after correction.

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>0.001</td>
<td>-0.002</td>
<td>0.0008</td>
<td>0.0001</td>
<td>-0.0021</td>
<td>-0.001</td>
<td>-0.003</td>
<td>-0.0001</td>
<td>-0.0054</td>
<td>-0.0024</td>
</tr>
<tr>
<td>%COVER</td>
<td>95.8</td>
<td>94.6</td>
<td>95.4</td>
<td>96</td>
<td>95.2</td>
<td>95.2</td>
<td>96</td>
<td>95.8</td>
<td>92.4</td>
<td>94.2</td>
</tr>
</tbody>
</table>

### Table 3.7. Bias and coverage under $H_0$, when all predictors followed a normal distribution, each one with different variance. The error followed a standard normal distribution. Continuous outcome. All models were significant before and after correction.

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>-0.0001</td>
<td>-0.0003</td>
<td>0.0001</td>
<td>-0.00001</td>
<td>0.0004</td>
<td>-0.0001</td>
<td>-0.0005</td>
<td>-0.00001</td>
<td>-0.0001</td>
<td>-0.0007</td>
</tr>
<tr>
<td>%COVER</td>
<td>94.8</td>
<td>96.6</td>
<td>95.4</td>
<td>95.4</td>
<td>95.6</td>
<td>93.8</td>
<td>96</td>
<td>93.4</td>
<td>95.8</td>
<td>95.2</td>
</tr>
</tbody>
</table>

### Table 3.8. Bias and coverage under $H_0$, when all predictors followed a standard normal distribution, each one rounded with different number of decimal places. The error followed a standard normal distribution. Continuous outcome. All models were significant before and after correction.

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>0.0014</td>
<td>0.0026</td>
<td>-0.0001</td>
<td>0.0024</td>
<td>-0.0005</td>
<td>-0.0005</td>
<td>0.001</td>
<td>0.0029</td>
<td>-0.0013</td>
<td>0.0026</td>
</tr>
<tr>
<td>%COVER</td>
<td>96.2</td>
<td>94.8</td>
<td>94.6</td>
<td>94.6</td>
<td>93.8</td>
<td>97.4</td>
<td>95</td>
<td>94.2</td>
<td>94.8</td>
<td>93.6</td>
</tr>
</tbody>
</table>

### Table 3.9. Bias and percentage of coverage under $H_0$, when all predictors followed a standard normal distribution. The error followed a normal distribution with standard deviation of 0.5. Continuous outcome. All models were significant before and after correction.

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>0.0002</td>
<td>-0.0016</td>
<td>0.0001</td>
<td>-0.00003</td>
<td>0.0007</td>
<td>-0.0002</td>
<td>-0.0009</td>
<td>0.0003</td>
<td>0.0007</td>
<td>0.0009</td>
</tr>
<tr>
<td>%COVER</td>
<td>94.6</td>
<td>94.6</td>
<td>96</td>
<td>95.6</td>
<td>93.6</td>
<td>93.8</td>
<td>95.4</td>
<td>96.2</td>
<td>96.8</td>
<td>94.6</td>
</tr>
</tbody>
</table>
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3.3.1.2. Binary Outcome

In this case the bias and the coverage were estimated from logistic regression models, using the probit function as a link because the error followed a normal distribution using the expected values of the models illustrated in the Methods section and the observed coefficients. Under both $H_A$ and $H_0$, the models were shown to reproduce the generating models in all different conditions.

Under the null hypothesis, the bias was always around 0 for all predictors, between -0.0027 and 0.0025 when the predictors and the error followed a standard normal distribution (Table 3.10), between -0.0006 and 0.0007 when the variance of each predictor was different ($\Sigma=\text{diag}(50,45,40,35,30,25,20,15,10,1)$, $\Sigma$ was the variance matrix for the predictors) (Table 3.11), between -0.0031 and 0.0039 when the number of decimal places was different for each predictor (Table 3.12), and between -0.0021 and 0.0032 when the variance of the error was 0.25 (Table 3.13). The coverage was always around 95%. When the predictor and the error followed a standard normal, the coverage ranged from 93.4% to 95.2% (Table 3.10), from 93.8% and 97% when the predictors had different variance (Table 3.11), from 92.6% to 95.8% when the predictors had different cut-points (Table 3.12), and from 93.6% and 97% when the standard deviation of the error was 0.5 (Table 3.13). The number of $p$-value < 0.05 was low, in general less than 33 in all different conditions, and maximum one model passed Bonferroni correction for each single predictor in all conditions.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>N(0,1)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>-0.0002</td>
<td>0.0016</td>
<td>-0.0005</td>
<td>0.0004</td>
<td>-0.0028</td>
<td>0.002</td>
<td>0.0025</td>
<td>0.0021</td>
<td>-0.0014</td>
<td>-0.0003</td>
</tr>
<tr>
<td>%COVER</td>
<td>94.4</td>
<td>94.8</td>
<td>94.4</td>
<td>93.6</td>
<td>93.4</td>
<td>93.4</td>
<td>95.2</td>
<td>95.2</td>
<td>93.4</td>
<td>94.4</td>
</tr>
<tr>
<td>P&lt;0.05</td>
<td>28</td>
<td>26</td>
<td>28</td>
<td>32</td>
<td>33</td>
<td>24</td>
<td>24</td>
<td>33</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>P&lt;0.0001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.10. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under H₀, when predictors and the error followed a standard normal distribution. Binary outcome.

<table>
<thead>
<tr>
<th>Diff variance</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>-0.0002</td>
<td>0.0001</td>
<td>0.0002</td>
<td>-0.0004</td>
<td>0.0002</td>
<td>-0.0006</td>
<td>0.0007</td>
<td>0.00001</td>
<td>0.0004</td>
<td>-0.0004</td>
</tr>
<tr>
<td>%COVER</td>
<td>93.8</td>
<td>94.2</td>
<td>95.2</td>
<td>94</td>
<td>95.2</td>
<td>97</td>
<td>95</td>
<td>94.6</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>P&lt;0.05</td>
<td>31</td>
<td>29</td>
<td>24</td>
<td>30</td>
<td>24</td>
<td>15</td>
<td>25</td>
<td>27</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>P&lt;0.0001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.11. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction under H₀, when all predictors are normally distributed with different variances. The error is standard normal distributed. Binary outcome.

<table>
<thead>
<tr>
<th>N(0,0.5)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>-0.0007</td>
<td>-0.0003</td>
<td>-0.0021</td>
<td>0.0006</td>
<td>0.0032</td>
<td>-0.0004</td>
<td>0.0001</td>
<td>0.0008</td>
<td>0.0009</td>
<td>-0.0002</td>
</tr>
<tr>
<td>%COVER</td>
<td>95.4</td>
<td>94.6</td>
<td>94.8</td>
<td>93.8</td>
<td>94.4</td>
<td>95.4</td>
<td>93.6</td>
<td>94.2</td>
<td>97</td>
<td>94.2</td>
</tr>
<tr>
<td>P&lt;0.05</td>
<td>23</td>
<td>27</td>
<td>26</td>
<td>31</td>
<td>28</td>
<td>23</td>
<td>32</td>
<td>29</td>
<td>15</td>
<td>29</td>
</tr>
<tr>
<td>P&lt;0.0001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.12. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under H₀, when all predictors follow a standard normal distribution, each one rounded with different number of decimal places. The error is standard normal distribution. Binary outcome.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Bias of Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors

<table>
<thead>
<tr>
<th>cutpoints</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>-0.0003</td>
<td>-0.0031</td>
<td>0.0025</td>
<td>-0.0005</td>
<td>0.004</td>
<td>-0.001</td>
<td>-0.0007</td>
<td>-0.0023</td>
<td>-0.0013</td>
<td>-0.0016</td>
</tr>
<tr>
<td>%COVER</td>
<td>94.4</td>
<td>94.2</td>
<td>94.6</td>
<td>94.6</td>
<td>95.8</td>
<td>94.2</td>
<td>95.2</td>
<td>95.2</td>
<td>92.6</td>
<td>93.4</td>
</tr>
<tr>
<td>P&lt;0.05</td>
<td>29</td>
<td>29</td>
<td>27</td>
<td>27</td>
<td>21</td>
<td>29</td>
<td>24</td>
<td>24</td>
<td>37</td>
<td>33</td>
</tr>
<tr>
<td>P&lt;0.0001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.13. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under H0. The error followed a normal distribution with 0.5 standard deviation. Binary outcome.

Under H_A, the bias was also around 0, with the largest range from -0.0029 to 0.0062 when the predictors had different number of decimals. The coverage ranged from 94.2% to 96.8% (Table 3.14) when all predictors and error followed a standard normal, from 93.2% to 95.6% (Table 3.15) when the predictors had different variance, from 93.0% to 96.6% (Table 3.16) when the predictors had different decimals, and from 93.4% to 96.0% (Table 3.17) when the variance of the error was 0.25. The effect of the associated predictor in each association study was statistically significant before and after Bonferroni correction (the number of p-values less than 0.05 and 0.0001 was always 500) in all conditions.

<table>
<thead>
<tr>
<th>N(0,1)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>-0.0024</td>
<td>0.0008</td>
<td>-0.0017</td>
<td>0.0002</td>
<td>-0.0008</td>
<td>-0.001</td>
<td>0.0027</td>
<td>0.004</td>
<td>0.0032</td>
<td>0.0012</td>
</tr>
<tr>
<td>%COVER</td>
<td>95</td>
<td>94.8</td>
<td>95.8</td>
<td>96.2</td>
<td>96.8</td>
<td>95</td>
<td>94.2</td>
<td>95.6</td>
<td>94.8</td>
<td>95</td>
</tr>
</tbody>
</table>

Table 3.14. Bias and coverage under H_A, when all predictors and the error followed a standard normal distribution. Binary outcome. All models were significant before and after correction.

<table>
<thead>
<tr>
<th>Diff Variance</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>0.0007</td>
<td>0.0022</td>
<td>0.0015</td>
<td>0.0014</td>
<td>0.0012</td>
<td>0.0016</td>
<td>0.0025</td>
<td>-0.0009</td>
<td>0.0013</td>
<td>0.004</td>
</tr>
<tr>
<td>%COVER</td>
<td>95</td>
<td>95.6</td>
<td>95.4</td>
<td>94.6</td>
<td>94.8</td>
<td>95.2</td>
<td>93.2</td>
<td>93.6</td>
<td>94</td>
<td>95.2</td>
</tr>
</tbody>
</table>

Table 3.15. Bias and coverage under H_A, when all predictors followed a normal distribution, each one with different variance. The error followed a standard normal distribution. Binary outcome. All models were significant before and after correction.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>cutpoints</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>0.0006</td>
<td>0.0062</td>
<td>0.0024</td>
<td>0.0002</td>
<td>-0.0024</td>
<td>-0.0007</td>
<td>0.001</td>
<td>0.0014</td>
<td>0.0027</td>
<td>-0.0029</td>
</tr>
<tr>
<td>%COVER</td>
<td>94.6</td>
<td>94.2</td>
<td>94.2</td>
<td>96.4</td>
<td>96.6</td>
<td>94</td>
<td>94.6</td>
<td>93</td>
<td>95.4</td>
<td>96.4</td>
</tr>
</tbody>
</table>

Table 3.16. Bias and coverage under $H_\alpha$, when all predictors followed a standard normal distribution, each one rounded different number of decimal places. The error followed a standard normal distribution. Binary outcome. All models were significant before and after correction.

<table>
<thead>
<tr>
<th>N(0,0.5)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>-0.0007</td>
<td>0.0002</td>
<td>0.0021</td>
<td>0.0015</td>
<td>0.0004</td>
<td>0.0015</td>
<td>-0.0018</td>
<td>0.0025</td>
<td>0.001</td>
<td>0.0017</td>
</tr>
<tr>
<td>%COVER</td>
<td>96</td>
<td>95.2</td>
<td>93.4</td>
<td>94.6</td>
<td>96</td>
<td>94.2</td>
<td>94.4</td>
<td>93.6</td>
<td>95</td>
<td>95</td>
</tr>
</tbody>
</table>

Table 3.17. Bias and percentage of coverage under $H_\alpha$, when all predictors followed a standard normal distribution. The error followed a normal distribution with a standard deviation of 0.5. Binary outcome. All models were significant before and after correction.

Once the data were shown to be well-generated, VIM$_{Gini}$ was applied in all different conditions to examine its behavior and make conclusions about its performance in real situations when continuous variables are used to model either a continuous outcome or a binary outcome.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3.3.2. VIM\textsubscript{Gini} for normal distributed variables with and without the same variance

3.3.2.1. Continuous outcome

3.3.2.1.1. Under the null hypothesis

The first approach in the study was to investigate the VIM\textsubscript{Gini} behavior under \( H_0 \). If there was a systematic bias under \( H_0 \), it was expected that the VIM would have a similar performance under \( H_A \).

When influential predictors did not exist, VIM\textsubscript{Gini} scores were similar for all predictors when all of them followed a standard normal distribution as well as when all of them followed normal distributions with different variance (\( \Sigma = \text{diag}(50,45,40,35,30,25,20,15,10,1) \), \( \Sigma \) was the variance matrix of uncorrelated predictors). In addition, VIM\textsubscript{Gini} did not show inflation when the variances of the predictor’s distributions were different compared to when the variance = 1 (Figure 3.2). Surprisingly, the VIM\textsubscript{Gini} medians for all ten predictors were larger than 60, which was higher than expected, given that there was no association. This suggested that something was affecting the VIM which was neither the variability of the predictors, nor the association of the predictors. This is discussed in section 3.3.4.1.1. See Appendix B for the VIM median for all predictors under \( H_0 \) (Table B.1).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

**Bias of Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors**

![Graph](image)

Figure 3.2. VIM$_{Gini}$ under $H_0$. The top plot illustrates the VIM when all predictors follow a standard normal distribution. The bottom plot shows the VIM when all predictors follow a normal distribution, but each one with a different variance. Continuous outcome.

### 3.3.2.1.2. Under the alternative hypothesis

Even though a systematic bias was not found under the null hypothesis, it could be that variability of the predictors has an impact on VIM$_{Gini}$ under association conditions. First, single association models were studied where all predictors followed a standard normal distribution. In these models only one predictor is associated from a total of ten and the association is the same for all true predictors (in all 10 models the coefficient was 0.3). When all predictors followed the same distribution, VIM$_{Gini}$ showed the same pattern across all models. VIM$_{Gini}$ for the associated predictor was around the same value in each case as well as for the non-influential predictors (Figure 3.3). Median VIMs for all predictors are shown in the Table B.3 of Appendix B.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 3.3. VIM\textsubscript{Gini} under H\textsubscript{A}. The figure illustrates VIM\textsubscript{Gini} in the ten different single association models, depending on which variable is associated, when all predictors follow a standard normal distribution. Continuous outcome. Each number i of the X axis corresponds to the subscript of the variable X\textsubscript{i}. 

Bias of Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

However, if all predictors had different variance ($\sum = \text{diag}(50, 45, 40, 35, 30, 25, 20, 15, 10, 1)$), $\sum$ was the variance matrix of the uncorrelated predictors, VIM$_{Gini}$ resulted in larger scores for the influential predictors with higher variability, although all of them had the same impact on the outcome (Figure 3.4) (see Table B.7 of Appendix B for the median VIMs of all predictors). This behaviour suggests that VIM$_{Gini}$ was inflating the scores of the influential variables only because of their variability, preferring those with more variability instead of showing similar scores for all influential variables (Figure 3.4). VIM$_{Gini}$ ranked the predictors also by variability and not only by association. Thus, X$_1$ had the largest variance (50) and it received the largest VIM$_{Gini}$ (VIM median = 2478.8), followed by X$_2$ (VIM median = 2228.02, variance of X$_2$ = 45) and so on, X$_{10}$ had the lowest VIM$_{Gini}$ (VIM median = 118.5) as well as having the least variability (variance of the X$_{10}$ = 1). All predictors had the same number of cut-points (the number of unique values of all predictors was 1000), so the inflation for predictors with higher variability must have some other explanation. In fact, this inflation was observed because larger values of a predictor with higher variance (when this was associated) multiplied by the same effect size (coefficient was the same in all association studies) leads to a larger value of the outcome (e.g. when a predictor followed a standard normal). Therefore, in terms of variance, greater variability of the predictor leads to greater variability in the outcome, and so higher VIM$_{Gini}$ scores for the predictors with greater variability.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 3.4. $\text{VIM}_{\text{Gini}}$ under $H_A$. The figure illustrates $\text{VIM}_{\text{Gini}}$ in the ten different single models, depending on which variable is associated, when all predictors follow a normal distribution, but each one with different variance. Continuous outcome. Each number $i$ of the X axis corresponds to the subscript of the variable $X_i$. 

Bias of Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3.3.2.2. Binary outcome

3.3.2.2.1. Under the null hypothesis

When the outcome was binary, VIM\textsubscript{Gini} also did not find a bias resulting from predictor variance under the null hypothesis. All median VIMs were approximately equal (31.5) among the ten predictors with different variance (Figure 3.5, bottom plot), and when all followed the same distribution (Figure 3.5). Median VIMs are shown in the Table B.2 of the Appendix B.

![Figure 3.5. VIM\textsubscript{Gini} under H\textsubscript{0}. The top plot illustrates the VIM when all predictors follow a standard normal distribution. The bottom plot shows the VIM when all predictors follow a normal distribution, but each one with a different variance. Binary outcome.](image-url)
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

### 3.3.2.2.2. Under the alternative hypothesis

As when the outcome was continuous, $VIM_{\text{Gini}}$ showed larger scores for influential predictors with higher variance under $H_A$. When all predictors followed a standard normal distribution (Figure 3.6), the median $VIM_{\text{Gini}}$ score was approximately around 46 for the influential predictors, and about 29 for the non-influential predictors in all single association studies (see Table B.4 in the Appendix B). However, when the variance of each predictor was different ($\Sigma=\text{diag}(50,45,40,35,30,25,20,15,10,1)$, $\Sigma$ is the variance matrix of the predictors), $VIM_{\text{Gini}}$ gave the largest scores to the predictor with the highest variance (median $VIM = 192.1$, variance 50) when this was influential, and the lowest median scores to the one with the smallest variance (median $VIM = 46.4$, variance 1) (Figure 3.7). See Appendix B for the median VIMs (Table B.8).

When the outcome was binary, all predictors also had the same number of unique values, and therefore same number of cut-points. But predictors with more variance are more widely separated and they may have more chance to yield in a better split, as they may have more chance to split the observations clearly between the left and the right branches of the split. This might be the reason for the observed $VIM_{\text{Gini}}$ inflation (as the measure is based on the Gini index).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 3.6. VIM\textsubscript{Gini} under H\textsubscript{A}. The figure illustrates VIM\textsubscript{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a standard normal distribution. Binary outcome. Each number i of the X axis corresponds to the subscript of the variable X\textsubscript{i}.  

Bias of Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 3.7. VIM\textsubscript{Gini} under H\textsubscript{A}. The figure illustrates VIM\textsubscript{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a normal distribution, but with different variances (\(\Sigma = \text{diag}(50,45,40,35,30,25,20,15,10,1)\)), \(\Sigma\) is the variance matrix of the predictors). Binary outcome. Each number \(i\) of the X axis corresponds to the subscript of the variable \(X_i\).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3.3.3. **VIM\textsubscript{Gini} normal distributed predictors with the same variance but with rounded to a different number of decimal places**

Working with continuous predictors in real situations, it might happened that predictors have different precision or that at least one predictor have different precision than others. For instance, in one study one might include the age (continuous variable without decimals), cognitive or environmental variables that are usually continuous and without decimal places, and also gene expression from different genes. Furthermore, data might come from different type of datasets or different sources such as in gene expression studies, for example Petralia et al. (2015) considered in their study gene expression data, protein-protein interactions, time-series gene expression and knockout data. Also, Banf and Rhee (2017) considered three types of datasets: conserved non-coding sequences and conserved non-coding promoter sequences; DNA binding predictions for transcription factors and experimental DNA binding motifs of other transcription factors; and expression atlas involving RNA samples from several tissues and developmental stages. As another example, in GWAS when taken the residuals because of PS, new variables are continuous but it might happen that the new variables have different precision.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3.3.3.1. Continuous outcome

3.3.3.1.1. Under the null hypothesis

VIM\textsubscript{Gini} behaviour when all predictors follow a standard normal was compared to the situations when predictors varied in precision, which was related to their scale of measurement, as in real studies continuous predictors may be measured with different numbers of decimal places. Since the predictors had different precision, there were a different number of unique values they each could take and, therefore, they have different number of cut-points. The following table (Table 3.18) illustrates the number of cut-points that each variable had under H\textsubscript{0} in median (500 datasets) when the variables have different precision. Table 3.19 shows the number of cut-points when all variables have the same precision.

<table>
<thead>
<tr>
<th>cutpoints</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>58</td>
<td>373</td>
<td>873</td>
<td>986</td>
<td>999</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Table 3.18. Median of the number of unique values of the variable X\textsubscript{i} under H\textsubscript{0}. Each variable X\textsubscript{i} has i number of decimal places. Continuous outcome.

<table>
<thead>
<tr>
<th>N(0,1)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Table 3.19. Median of the number of unique values of the variable X\textsubscript{i} under H\textsubscript{0}. All predictors follow a standard normal distribution with the same number of decimal places. Continuous outcome.

As showed in the top plot of Figure 3.2, when all predictors followed the same distribution and their scale of measurement was the same, VIM\textsubscript{Gini} did not prefer one predictor over another. However, VIM\textsubscript{Gini} behave differently if the number of decimal places for each predictor was different. VIM\textsubscript{Gini} showed lower scores for predictors with fewer unique values (less cut-points), but showed similar VIM medians for predictors with more than 3 decimal places (Figure 3.8) which had more cut-points (predictors with more than 3 decimal places had similar or same number of unique values). Therefore, under no association, VIM\textsubscript{Gini} showed an inflation for predictors.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

with more cut-points. See Table B.1 in Appendix B for the VIM median when all predictors had different precision.

![VIM Gini under H0](image)

**Figure 3.8. VIM\textsubscript{Gini} under H\textsubscript{0}.** The top plot illustrates the VIM when all predictors follow a standard normal distribution. The bottom plot shows the VIM when all predictors follow a normal distribution, but each one with different number of decimal places. X\textsubscript{1} has one decimal place, X\textsubscript{2} has two decimal places, ..., X\textsubscript{10} has ten decimal places. Continuous outcome.

This bias towards continuous predictors with more cut-points is related to the bias towards variables with more categories found by Strobl \textit{et al.} (2007b), due to the bias of the variable selection in each individual tree because of the Gini index criterion (Boulesteix 2006; Strobl \textit{et al.} 2007a). Because this index is calculated within the range of the predictors for all cut-points, the one with the largest Gini index score overall, is the predictor selected for the split (in its best cut-point). As in a multiple
Bias of Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors

Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

testing case, the predictors with more precision (more tests) are more likely to have a good Gini index value by chance, so the \( VIM_{\text{Gini}} \) showed inflation for those variables. For instance, the number of Gini index values that have to be computed for the predictor \( X_1 \) (median 58 unique values) is fewer than for \( X_4 \), and the largest Gini index value from \( X_1 \) has to be compared with the largest Gini index value from the values in all cut-points of \( X_4 \) (median 986 unique values). Therefore, the value from \( X_4 \) would usually be preferred.

### 3.3.3.1.2. Under the alternative hypothesis

Each association study had all predictors with different precision: in all studies the \( X_i \) variable had \( i \) decimal places. The datasets for the association studies were simulated in this way to be consistent with the case when all predictors had different variance, instead of considering each dataset with all predictors rounded with the same number of places (one dataset with predictors with one decimal place, other with predictors with two decimal places, and so on). Under the \( H_A \), it was expected that \( VIM_{\text{Gini}} \) would show larger median values for continuous predictors with more decimal places than those with fewer decimal places. When all variables followed a standard normal with the same number of cut-points (Table 3.20), \( VIM_{\text{Gini}} \) did not show a preference for any of the variables, and all \( VIM_{\text{Gini}} \) scores for the influential variables were around 117 in all association studies under \( H_A \) (Figure 3.3).

In this situation, when predictors had the same variance but different number of decimal places, \( VIM_{\text{Gini}} \) did not give the same scores for the influential predictors of each of the models (Figure 3.9). When \( X_1 \) was influential and had only one decimal place, it showed a lower \( VIM_{\text{Gini}} \) value compared to \( VIM_{\text{Gini}} \) for the influential predictors in the other models, even though the effect size was the same in all association studies. When \( X_1 \) was not associated, \( VIM_{\text{Gini}} \) also showed the lowest values for that variable. As under \( H_0 \), the medians for influential predictors which had more than 3 decimal places, were about the same. Furthermore, the value of \( VIM_{\text{Gini}} \) for influential predictors with more than 3 decimal places was approximately the same.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

as when the influential predictors had the same number of decimal places, because they had the same or a similar number of cut-points (Table 3.20 and Table 3.21). See Table B.5 in Appendix B for the VIM median values.

<table>
<thead>
<tr>
<th>N(0,1)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Table 3.20. Median of the number of unique values of the variable Xᵢ under Hₓ. All predictors follow a standard normal distribution with the same number of decimal places. Continuous outcome.

<table>
<thead>
<tr>
<th>cutpoints</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>58</td>
<td>373</td>
<td>871.5</td>
<td>986</td>
<td>999</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Table 3.21. Median of the number of unique values of the variable Xᵢ under Hₓ. Each variable Xᵢ has i number of decimal places. Continuous outcome.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Bias of Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

The fact that \( \text{VIM}_{\text{Gini}} \) showed lower scores for predictors with less than 3 decimal places was because \( \text{VIM}_{\text{Gini}} \) is based in the Gini index, and as in multiple testing situations, predictors with more cut-points had more chance of being selected for the next split, as explained before under \( H_0 \). The fact that for predictors with more 3 decimal places \( \text{VIM}_{\text{Gini}} \) had a similar behavior was because the number of unique values was so similar or equal.

3.3.3.2. Binary outcome

3.3.3.2.1. Under the null hypothesis

As when the outcome was continuous, \( \text{VIM}_{\text{Gini}} \) showed lower scores for predictors with fewer than 3 decimal places under \( H_0 \), being the lowest for \( X_1 \) that was rounded with only one decimal place (Figure 3.10 bottom plot; see Table B.2 in Appendix B for the VIM medians). This inflation for predictors with more than 3 decimal places suggests a systematic bias when there is no predictor associated with the outcome, which should be considered in real situations to avoid spurious results. The reason for this behaviour is the same as when the outcome was continuous: because \( \text{VIM}_{\text{Gini}} \) is based on the Gini index. The median number of unique values of the predictors when all variables had the same precision are shown in Table 3.22, and when all predictors have different precision are shown in Table 3.23.

<table>
<thead>
<tr>
<th>N(0,1)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Table 3.22. Median of the number of unique values of the variable \( X_i \) under \( H_0 \). All predictors follow a standard normal distribution with the same number of decimal places. Binary outcome.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>cutpoints</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>58</td>
<td>373</td>
<td>874</td>
<td>987</td>
<td>999</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Table 3.23. Median of the number of unique values of the variable Xᵢ under H₀. Each variable Xᵢ has i number of decimal places. Binary outcome.

Figure 3.10. VIMGF under H₀. The top plot illustrates the VIM when all predictors follow a standard normal distribution. The bottom plot shows the VIM when all predictors follow a normal distribution, but each one with different number of decimal places. X₁ has one decimal place, X₂ has two decimal places, ..., X₁₀ has ten decimal places. Binary outcome.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3.3.3.2.2. Under the alternative hypothesis

Under $H_A$, it was expected that $VIM_{Gini}$ would give similar results to those under the null hypothesis. In fact, $VIM_{Gini}$ also inflated the scores for predictors with more decimal places than 3 when these predictors were associated with the outcome, and also when the predictors were not associated compared to the values for other non-influential ones (Figure 3.11; see Table B.6 in Appendix B for the VIM medians). The reason for this inflation was related to the fact that the Gini index is more likely to select variables with more precision (cut-points), and $VIM_{Gini}$ is based on this index, as explained before in subsection 3.3.3.1.1. Therefore, Gini is also biased when predictors have different number of decimal places under both $H_0$ and $H_A$. Table 3.23. and Table 3.24. show the median of unique values of each predictor when all predictors have the same precision and different precision, respectively.

<table>
<thead>
<tr>
<th>N(0,1)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Table 3.24. Median of the number of unique values of the variable $X_i$ under $H_A$. All predictors follow a standard normal distribution with the same number of decimal places. Binary outcome.

<table>
<thead>
<tr>
<th>cutpoints</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>58</td>
<td>372.5</td>
<td>873</td>
<td>986</td>
<td>999</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Table 3.25. Median of the number of unique values of the variable $X_i$ under $H_A$. Each variable $X_i$ has $i$ number of decimal places. Binary outcome.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Bias of Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors.

Figure 3.11. VIM\textsubscript{Gini} under H\textsubscript{A}. The figure illustrates VIM\textsubscript{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a standard normal distribution, but each one has different number of decimal places. Binary outcome. Each number i of the X axis corresponds to the subscript of the variable X\textsubscript{i}.
3.3.4. VIM\textsubscript{Gini} for error with different variances

3.3.4.1. Continuous outcome

3.3.4.1.1. Under the null hypothesis

Taking into account the null results when the error followed a standard normal distribution and all variables followed a standard normal, in this subsection I will compare VIM\textsubscript{Gini} under those conditions to the case when all variables follow a normal distribution but with error variance 0.25 under no association. In both cases, only noise exists, and VIM\textsubscript{Gini} should not give more importance to any of the predictors. The VIM should also be around the same value in both situations, otherwise it is measuring the variability of the error. It is important to say that the median of the unique values of each predictor was always 1000 as well as the medians of the outcome in both situations, when the error had the two different variances. Also, all predictors had the same variability in both situations.

The results from the above subsection 3.3.2.1.1 showed that the median value of VIM\textsubscript{Gini} was higher than 60. With decreased variance of the error, VIM\textsubscript{Gini} showed lower scores for all predictors (approximately 15.5) (Figure 3.2) (see Appendix B Table B.1 for the VIM medians). Therefore, VIM\textsubscript{Gini} was inflating the scores of all predictors when more error variance was present in the model. This inflation was not a real association, which suggests another bias of VIM\textsubscript{Gini}, in this case, towards more error variance. It is difficult to know the reason for this inflation, as VIM\textsubscript{Gini} is based on the decrease of impurity and is supposed to measure how likely it is that one variable has an impact on the outcome, and not the variance of the error.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

![Figure 3.12. VIM\textsubscript{Gini} under H\textsubscript{0}. The top plot illustrates the VIM when the error follows a standard normal distribution. The bottom plot shows the VIM when error has a variance of 0.25. Continuous outcome.](image)

However, this bias related to how our models were generated - a linear regression model, where the outcome was actually the error under the null hypothesis. As the variability of the error was lower, the variability of the outcome was also lower (variance = 0.25, as expected). This is related to the situation when the predictors with higher variance led to higher variance of the outcome under the alternative (section 3.3.2.1.2), and therefore, inflation of \(\text{VIM}_{\text{Gini}}\). So, the decrease of the variability of the outcome because of lower error variance may be the reason for the decrease on the \(\text{VIM}_{\text{Gini}}\) scores. Here, any predictor had higher scores than any other as their variability was the same, but in general the \(\text{VIM}_{\text{Gini}}\) was lower. This bias was unexpected as \(\text{VIM}_{\text{Gini}}\) was supposed to be checking the variables and it should be blind to the noise.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3.3.4.1.2. Under the alternative hypothesis

As VIM_{Gini} is biased under \( H_0 \), it was expected that VIM_{Gini} would perform similarly under \( H_A \). The median VIM_{Gini} was about 117 for the predictors which were truly associated in each single model when the error followed a standard normal distribution. The non-influential predictors showed the same value in all association studies when the error had a variance of one, as under \( H_0 \) (subsection 3.3.2.1.2, Figure 3.3). Under \( H_A \), the median number of unique values of each predictor and the outcome was also 1000 in both cases (datasets generated when the error had a variance of 1 and a variance of 0.25). Furthermore, the variability of all predictors was the same in both situations.

Here, when the error followed a normal distribution with a variance 0.25 (standard deviation of 0.5) - lower than before - VIM_{Gini} showed the same scores for all influential variables across all individual association studies, and the same median values among the non-associated ones (Figure 3.3). However, VIM_{Gini} showed a decline on the scores for all predictors, in every study compared to when the error had higher variance (median VIMs for influential predictors was around 67; see Appendix table B.9 for the VIM medians), as under \( H_0 \). The observed variability of the outcome was 0.34 when the error had a variance of 0.25, as expected (variance expected of \( y \) from the generating model is \( 0.3^2 \times 1 + 0.25 (\beta^2 \times \mu + \sigma^2) \)), and was observed to be 1.09 when the error had variance one (as expected: \( 0.3^2 \times 1 + 1(\beta^2 \times \mu + \sigma^2) \)). The decline in VIM_{Gini} with a lower error variance was due to the same reason (variability decreased of the outcome because of lower error variance) as under \( H_0 \).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure 3.13. VIM\textsubscript{Gini} under H\textsubscript{A}. The figure illustrates VIM\textsubscript{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a standard normal distribution, but the error \sim N(0,0.5). Continuous outcome. Each number \(i\) of the X axis corresponds to the subscript of the variable \(X_i\).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Therefore, the overall inflation of VIM\textsubscript{Gini} due to more error variance was shown under both hypotheses, which suggested that the variance of the error had an impact on the VIM with or without predictor association. Related to what Boulesteix \textit{et al.} (2012b) suggested, VIM\textsubscript{Gini} may be preferred when the signal-to-noise ratio is low, but more variance in the noise may lower the ratio. However, greater error variance actually inflated the predictor VIM\textsubscript{Gini} scores. So, under association (signal not equal to 0), VIM\textsubscript{Gini} might be seen to better detect the correct signal as it would show larger values for the influential predictors, but the inflation would be due to more variance in the noise, not because of true association. This bias towards more error variance is an important fact to take into account in real studies, as one wants to avoid noise, although it cannot be usually removed. Therefore, the use of another VIM is suggested, such as the unconditional unscaled permutation VIM.

3.3.4.2. Binary outcome

3.3.4.2.1. Under the null hypothesis

The behaviour of VIM\textsubscript{Gini} under H\textsubscript{0} when the error had smaller variance and the outcome was binary was different to when the outcome was continuous. When the outcome was binary, the distribution of VIM\textsubscript{Gini} scores was similar for all 10 predictors. The medians for all predictors when the error had two different variances were approximately the same value (31.5) (Figure 3.4; see Table B.2 in Appendix B for the VIM medians). So, when the outcome was binary, there was no inflation of the VIM\textsubscript{Gini} and, therefore, no bias towards greater error variance when predictors were not influential. This fact was due to the predictors having the same number of cut-points, and the observed variance of the outcome was 0.25 in both cases – when the datasets were generated with both error variances (1 and 0.25).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3.3.4.2.2. Under the alternative hypothesis

Under $H_A$, when the outcome was binary and the error had 0.25 variance, VIM$_{Gini}$ showed the opposite behaviour to when the outcome was continuous. Having the same coefficient for each influential predictors in all association studies and less error variance led to larger VIM$_{Gini}$ scores for the influential predictors than when the variance of the error was higher (Figure 3.5; Table B.10. in Appendix B for the VIM medians). Here, the number of unique values of the predictors was also 1000 as in the continuous case and as under $H_0$ in both studies with two different error variance. Moreover, as under $H_0$, the variance of the outcome was 0.25 when the error had two different variances.

Figure 3.14. VIM$_{Gini}$ under $H_0$. The top plot illustrates the VIM when the error followed a standard normal distribution. The bottom plot shows the VIM when error $\sim N(0,0.5)$. Binary outcome.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Bias of Random Forest variable importance measures based on the Gini importance based on the error variance and the variability of the predictors.

Figure 3.15. VIM\textsubscript{Gini} under H\textsubscript{A}. The figure illustrates VIM\textsubscript{Gini} in the ten different single models, depending on which variable is associated, when all predictors followed a standard normal distribution, but the error ~ N(0,0.5). Binary outcome. Each number i of the X axis corresponds to the subscript of the variable X\textsubscript{i}.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

If the coefficients of the predictors were the same but the error variance was lower (with the same mean), the association between the predictors and the outcome became stronger, which made the method more capable of detecting the true signals. This may be the reason for the $\text{VIM}_{\text{Gini}}$ inflation observed when the error variance was lower. Therefore, $\text{VIM}_{\text{Gini}}$ was shown not to be biased under either $H_0$ or $H_A$.

Table 3.26 and Table 3.27 show a summary of the results under the null hypothesis and under the alternative hypothesis. It is important to say that the three different cases or studies where compared to the case when all variables and the error followed a standard normal distribution.

<table>
<thead>
<tr>
<th>UNDER $H_0$</th>
<th>CONTINUOUS OUTCOME</th>
<th>BINARY OUTCOME</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{10} \sim N(0, \Sigma_2), e_1 \sim N(0,1)$ Variables with different variance</td>
<td>Unbiased</td>
<td>Unbiased</td>
</tr>
<tr>
<td>$X_{10} \sim N(0,I), e_1 \sim N(0,1)$ Variables with different precision</td>
<td>Biased. Inflates the scores for more precise variables</td>
<td>Biased. Inflates the scores for more precise variables</td>
</tr>
<tr>
<td>$X_{10} \sim N(0,I), e_1 \sim N(0,0.5)$ Less error variance</td>
<td>Biased. Inflates the scores when the error variance is higher</td>
<td>Unbiased</td>
</tr>
</tbody>
</table>

Table 3.26. Summary of $\text{VIM}_{\text{Gini}}$ behaviour on the three different studies compared to when all variables and error followed a standard normal distribution under $H_0$.

<table>
<thead>
<tr>
<th>UNDER $H_A$</th>
<th>CONTINUOUS OUTCOME</th>
<th>BINARY OUTCOME</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{10} \sim N(0, \Sigma_2), e_1 \sim N(0,1)$ Variables with different variance</td>
<td>Inflates the scores for the predictors with more variance</td>
<td>Inflates the scores for the predictors with more variance</td>
</tr>
<tr>
<td>$X_{10} \sim N(0,I), e_1 \sim N(0,1)$ Variables with different precision</td>
<td>Inflates the scores for more precise variables</td>
<td>Inflates the scores for more precise variables</td>
</tr>
<tr>
<td>$X_{10} \sim N(0,I), e_1 \sim N(0,0.5)$ Less error variance</td>
<td>Inflates the scores when the error variance is higher</td>
<td>Inflates the scores when the error variance is lower</td>
</tr>
</tbody>
</table>

Table 3.27. Summary of $\text{VIM}_{\text{Gini}}$ behaviour on the three different studies compared to when all variables and error followed a standard normal distribution under $H_A$.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

### 3.4. Discussion

In this study, $\text{VIM}_{\text{Gini}}$ was applied to different simulated datasets comprising uncorrelated and continuous predictors (normally distributed) and two types of outcomes, a continuous outcome modelled by linear regression models, and a binary outcome modelled by logistic regression models using the probit link. In previous studies, $\text{VIM}_{\text{Gini}}$ was shown to be biased under predictor correlation (Nicodemus and Malley 2009); (Nicodemus 2011) towards categorical predictors with more categories (Strobl et al. 2007b), and towards SNPs with higher minor allele frequency (Nicodemus 2011); (Boulesteix et al. 2012a). Thus, $\text{VIM}_{\text{Gini}}$ was suggested for use when all predictors were continuous and uncorrelated, and also when the signal-to-noise ratio was low (Boulesteix et al. 2012b).

Therefore I performed a simulation study to examine the behaviour of $\text{VIM}_{\text{Gini}}$ in three different situations: (1) when all predictors follow a standard normal with different variances; (2) when the predictors have the same variance but are rounded to different numbers of decimal places; and (3) when the error follows a standard normal but with different variances. In all the three conditions, $\text{VIM}_{\text{Gini}}$ was compared to the case when all predictors and the error follow a standard normal distribution (all predictors and error have mean 0 and same variance 1, and also same precision).

In these three situations under $H_0$, $\text{VIM}_{\text{Gini}}$ was biased by the scale of measurement of continuous variables, showing lower scores for the predictors with one or two decimal places than for the ones with more decimal places. This bias occurred when the outcomes were both continuous and binary. In addition, when the outcome was continuous, $\text{VIM}_{\text{Gini}}$ showed a bias towards error with more variance when the independent variable (outcome) was continuous, inflating the scores for all predictors when the error variance was higher (1 compared to when the variance was 0.25).

Under $H_A$, $\text{VIM}_{\text{Gini}}$ inflated the scores for the predictors with more variance, for predictors with more cut-points (more than three decimal places), and when the error variance was higher with a continuous outcome. When the outcome was binary,
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

VIM\textsubscript{Gini} showed larger scores both for variables with more variability and for variables with more cut-points. With the binary outcome, VIM\textsubscript{Gini} inflation also happened when error variance was smaller, in this case because the signal from the true associated predictors was more capable or easy to detect.

This is an important fact to consider in real studies to avoid spurious results. In real studies, researchers might use different types of data - where predictors have more variability, where continuous variables have different number of decimal places, or where different types of data have different noise sources - when studying a particular phenotype (continuous or binary). For instance, RF was used in pathway analysis to investigate groups of genes at once, such as in gene expression studies (Pang \textit{et al.} 2006); (Pang and Zhao 2008), sometimes using data from different sources KEGG, BioCarta, and manually as Pang \textit{et al.} (2006). Pang and Zhao (2008) applied RF based on VIM\textsubscript{Gini}, as they argued that it was possible that the measure was not biased, because the gene expression data were normalized and because they did not use categorical predictors.

More recent studies have also applied RF based on the decrease in impurity (VIM\textsubscript{Gini}) in gene expression data. Huynh-Thu \textit{et al.} (2010) used VIM\textsubscript{Gini} to rank regulatory links of association between genes in microarray gene expression data (variables and outcome were continuous). The authors performed simulations to test their proposed genetic regulatory network model using VIM\textsubscript{Gini}. In the simulations they added random noise to the expression data measurement as well as in the dynamics of the networks. Furthermore, Petralia \textit{et al.} (2015) also proposed a model to build genetic regulatory networks using VIM\textsubscript{Gini}, taking into account information from different types of data, such as gene expression data, time-series experiments, protein-protein interactions and knockout experiments. The authors also tested for the association between genes to build the networks. They considered the gene expression datasets as the main input (variables and outcome were continuous), and they used the other types of data to calculate weights to include prior information when sampling the data to be part of the pool of variables selected to split the tree. The regulatory links were ranked using the decrease in impurity (VIM\textsubscript{Gini} is based on the decrease in impurity).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

The results from the present study show that if the genes in some groups have more variability of the expression than in other groups, and the genes are associated, VIM\textsubscript{Gini} will prefer the ones with more expression variability even though the association with the outcome (e.g. phenotype or other gene) is the same as the genes in other groups with less variability. Furthermore, VIM\textsubscript{Gini} would not be helpful in real studies when continuous predictors vary in precision, and their association with a continuous phenotype or a binary phenotype is under study. The VIM would show more importance for genes that have more than 3 decimal places, even though the impact of other genes with less than 3 decimal places is the same, and also when there is no real evidence of influence from any gene in any set. For the same reason, when one set of genes has different precision from another, this would lead to spurious results.

The study suggests normalizing the predictors to avoid inflation because of more variability, as well as rounding continuous predictors to the same number of decimal places. However, this study is the first to find a VIM\textsubscript{Gini} bias towards more error variance with continuous outcomes. This would have an important impact, causing misleading results, when sets of genes (pathways) present more noise than others, which may happen in real studies even though noise is not visible. Therefore, when applying RF in real situations, the use of other VIMs should be considered in order to avoid inflation due to the noise, rather than real associations.
4. Detecting significantly associated interactions with schizophrenia and cognition in abnormal behaviour and pathways from the Mouse Genotype Informatics (MGI) database

4.1. Introduction

Psychosis is a syndrome characterised by hallucinations and delusions and is considered a psychiatric human syndrome rather than a disorder in itself. The major psychiatric disorder that features psychosis is schizophrenia, but it can also be observed in individuals with BP and MDD. Patients with schizophrenia, BP and MDD also show abnormal cognitive function which is, in some cases, detected from childhood (Johnson, 2005); (Kahn and Keefe, 2013); (The National Academies Collection, 2015).

Substantial progress has been made in identifying common risk variants (SNPs) contributing to susceptibility to the major psychoses. Individual SNPs have small effects but the aggregate role of many SNPs, as measured by the PRS, can make a significant contribution to risk as demonstrated in schizophrenia (Ripke et al. 2014). There is also a growing appreciation of the genetic overlap between the psychoses, depression (where psychosis is a less common symptom) and other psychiatric disorders (Huang et al. 2010). Yet across all of these psychotic disorders the majority of genetic variance is yet to be explained. Therefore, I focused my study on testing for epistasis (gene-gene interactions) to see if that might explain more variation in psychosis and cognition.

The RDoC project aims to make a direct connection between observed phenotypes from the cellular level through behaviour and genetics and has attracted much attention from researchers in recent years (Insel 2014). In other words, the RDoC project seeks to study positive and negative valence systems, cognition, social processes and arousal & regulatory systems, as well as the relation of these domains with genomic,
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

molecular, cellular, circuit, physiological and behavioural factors. RDoC is a research framework for new ways of studying mental disorders trying to identify a spectrum of intermediate phenotypes which overlap between disorders, such as cognitive abnormalities, rather than to study one particular ‘diagnosis’ category. In the spirit of the RDoC initiative, we studied a case sample that included individuals with psychosis and DSM-IV diagnoses including schizophrenia, schizoaffective disorder, BP and MDD with psychosis, looking for genetic factors associated with a common phenotype across disorders.

Animal models attempt to imitate a human condition such as the psychopathology of psychotic disorders in humans in order to study psychosis in animals. As psychoses are human illnesses, it is difficult to reproduce them in animal models, therefore, it is important to model the psychosis, as positive symptoms, instead to imitate a particular psychotic disease (Schobel et al. 2013);(Moran et al. 2014); (Papaleo et al. 2014); (Dachtler et al. 2016).

As said in previous chapters, over the last decade ML algorithms have been increasingly used in genetics and neuroscience. In genetics, with the introduction of increasingly larger GWAS, these techniques have become necessary due to the high dimensionality of data. The challenge of managing “Big Data” with more variables than observations makes ML, which can efficiently handle the “\( p >> n \)” problem, attractive to researchers. RF is a non-linear, non-parametric supervised ML algorithm which has shown excellent performance in high dimensional data analysis. One of RF’s main characteristics is that it returns measures of variable importance, which is a measure of the strength of the association between a predictor and the outcome in the context of all other predictors.

The main aim of our study was to use RF based on the unscaled PVIM, named VIMrawperm-RF in Chapter 2, to test for epistasis between genes in both case-control and cognitive outcomes, IQ and verbal IQ. As patients with schizophrenia, schizoaffective disorder, BP and MDD with psychosis show abnormal behaviours and cognitive
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Impairment, we used pathways based on The Mouse Genome Informatics database (Blake et al. 2017) for abnormal behaviour: abnormal emotion/affect behaviour [MP:0002572], including four pathways: abnormal aggression-related behaviour [MP:0002061], abnormal depression-related behaviour [MP:0003360], abnormal fear/anxiety-related behaviour [MP:0002065], and abnormal response to novelty [MP:0003107]. These phenotypes were selected based on behaviour that may be impaired in psychosis and may effect cognition, and genes selected were from various types of mouse models that affect or disrupt the genes involved in each pathway.

4.2. Methods

4.2.1. Data and analysis

The study was performed to test for risk of psychosis, along with full-scale IQ and verbal IQ in cases, in a previously described Irish case-control psychosis cohort (Hargreaves et al., 2014); (Irish Schizophrenia Genomics Consortium and the Wellcome Trust Case Control Consortium 2, 2012). To reduce multiple testing and LD between SNPs, I used functional SNPs only (synonymous, missense, splice region variants, and 3’ and 5’ UTR (Table 4.1.)). I extracted the SNPs inside of the gene range with plink6 (Purcell et al. 2007), and I included SNPs with MAF 0.01 or greater as well as SNPs that passed the Hardy-Weinberg test in controls (p-value threshold 0.05).

<table>
<thead>
<tr>
<th>MGI Phenotype</th>
<th>N Human Genes</th>
<th>N Functional SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggression</td>
<td>79</td>
<td>440</td>
</tr>
<tr>
<td>Depression</td>
<td>86</td>
<td>440</td>
</tr>
<tr>
<td>Fear/Anxiety</td>
<td>272</td>
<td>1446</td>
</tr>
<tr>
<td>Novelty</td>
<td>189</td>
<td>1067</td>
</tr>
</tbody>
</table>

Table 4.1. Number of genes and SNPs by pathway.
The case/control study involved 2,049 cases with a DSM-IV diagnosis (major psychotic disorder including schizophrenia and schizoaffective disorder, individuals with BP and MDD with psychosis) and 1,794 controls. Controls were blood donors and were not screened for psychosis, although this is unlikely to lead to significant misclassification as Irish blood donors are not financially remunerated and psychotic disorders are rare in the general population (approximately 1-2%). The cognitive subset included individuals with IQ > 70: 306 narrow psychosis (schizophrenia or schizoaffective disorder); and 71 broad psychosis, including BP or MDD with psychosis. IQ was calculated by the Wechsler Adult Intelligence Scale. As RF is sensitive to class imbalance (Boulesteix et al. 2012b), the study design for case status was to use a balanced training set with 80% of control observations (the group with smaller sample size) and the same number of cases, and independent test sample with the remaining cases and the remaining 20% controls. For the cognitive variables I was able to use the 100% of patients’ observations.

4.2.2. Random Forest

Based on the results of the second chapter, I chose the unscaled PVIM to perform RF. As explained in the introduction chapter, RF is a ML technique able to detect interactions from its natural architecture in recursive trees, which provides certain dependency in a hierarchical way through the forest (García-Magariños et al. 2009). In fact, RF has efficiently detected SNP-SNP interaction effects in previous studies (Lunetta et al., 2004); (Bureau et al., 2005); (Nicodemus, Callicott, et al., 2010); (Nicodemus, Law, et al., 2010); (Schwarz, König and Ziegler, 2010). Although RF is able to detect interactions with or without main effects, its results are difficult to interpret; the ranking based on RF VIMs does not tell us much about which variable has a significant single or interacting contribution, therefore I tested for interactions in the independent test datasets using LRTs between nested generalised linear models to validate the interaction effects between them.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

I set the number of trees to 1000 and the \textit{mtry}, which is the size of randomly chosen variable sets, equal to the square root of the number of SNPs which was different in every pathway taking into account the different number of SNPs per pathway. The percentage of subsampling of observations for tree-growing was fixed to 63.2, which is the average percentage of unique values under replacement. Using these parameters, I ran RF 500 times on the original data, each time changing the random number seed, in order to obtain stable estimates of the VIMs by taking the median of the 500 values for each SNP; and also over null data, created after permuting the phenotype, in order to calculate a null distribution of the VIMs and to calculate an empirical \( p \)-value to avoid false positives. In this way, our \( p \)-value threshold for detecting significant results is \( p = 0.05 \). Indeed I used RF like a filter to reduce the multiple testing; if I had used regression models I would have had to test at least \( 90,122,025 \) tests with a Bonferroni \( p \)-value threshold of \( 5.55 \times 10^{-10} \) (in the pathway with the smallest number of SNPs). Once we calculated the empirical \( p \)-value, we took the 30 most empirically-significant predictors from the training data. RF analyses were conducted using Random Jungle (Schwarz, König and Ziegler, 2010).

\subsection*{4.2.3. Likelihood Ratio Tests (LRTs) between nested models}

The 30 most significant SNPs were taken forward for follow up with LRTs of nested models in our independent test sample to test for epistasis associated with psychosis case status as well as to test for significant interactions on IQ and verbal IQ. The nested models were performed as follows:

\textbf{2 way SNP interactions}

\begin{align*}
\text{Full model: } & Y \sim \beta_1 \text{SNP}_1 + \beta_2 \text{SNP}_2 + \beta_3 \text{SNP}_1 \ast \text{SNP}_2 \\
\text{Reduced Model: } & Y \sim \beta_1 \text{SNP}_1 + \beta_2 \text{SNP}_2
\end{align*}
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

3 way SNP interactions

*Full model:* \( Y \sim \beta_1 \text{SNP}_i + \beta_2 \text{SNP}_j + \beta_3 \text{SNP}_k + \beta_4 \text{SNP}_i \times \text{SNP}_j + \beta_5 \text{SNP}_i \times \text{SNP}_k + \beta_6 \text{SNP}_j \times \text{SNP}_k + \beta_7 \text{SNP}_i \times \text{SNP}_j \times \text{SNP}_k \)

*Reduced Model:* \( Y \sim \beta_1 \text{SNP}_i + \beta_2 \text{SNP}_j + \beta_3 \text{SNP}_k + \beta_4 \text{SNP}_i \times \text{SNP}_j + \beta_5 \text{SNP}_i \times \text{SNP}_k + \beta_6 \text{SNP}_j \times \text{SNP}_k \)

Where

\( \text{SNP}_i, \text{SNP}_j \) and \( \text{SNP}_k \) are within the most significant 30 SNPs across all RF iterations. \( Y \) is the phenotype. In the case/control study it is a binary trait, and in the IQ and verbal IQ studies it is a continuous trait. Finally, I used Nagelkerke’s pseudo-$R^2$ for logistic regression models and standard $R^2$ for linear regression models to determine the amount of variation explained in outcomes. I performed the analyses with R version 3.0.0 and used the packages fmsb and lmtest for calculating the LRT and the Nagelkerke’s pseudo-$R^2$ respectively.

First, I performed LRTs to detect 2-way interactions of the Top30 SNPs. If any statistically significant 2-way interactions were observed, I tested 3-way interactions between those two SNPs compared to the remaining 28 SNPs. Figure 4.1 illustrates a diagram of the study design of this study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Detecting significantly associated interactions with schizophrenia and cognition in abnormal behaviour and pathways from the Mouse Genotype Informatics (MGI) database
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

In general, when applying ML techniques in real studies, there is no better technique than other, to know which ones are suitable for being applied under certain conditions that real datasets may present, a simulation study (under those conditions) should be performed. In Chapter 2, I performed a simulation considering different correlation patterns between variables to know which VIM is suitable to be applied for detecting interactions under correlation conditions. Although, this thesis was not focused on comparing RF with other ML techniques, it is important to say that there exist other ML techniques which can be applied for detecting of interactions. For instance, multifactor dimensionality reduction (MDR), SVM or LASSO. When the number of variables is greater than several hundreds, MDR is so slow, in fact, induces an extreme computational burden (Niel et al. 2015). So, examining all 2-way combinations of SNPs can be a computational challenging, which becomes more challenging when examining higher order interactions such as 3-way interactions (Bush and Moore 2012). SVM does not return variable importance which would make difficult to filter the variables and select the top variables which might interact. In addition, in order to apply LASSO, interactions have to be explicitly incorporated on the model, in other words, all order interaction need to be fixed up manually when programming. However, RF incorporates all SNPs into the model and gives VIMs taking into account main and interaction effects.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

4.3. Results

The overlap between genes in the four pathways considered was modest, with the largest number of overlaps between two pathways (the fear/anxiety and the depression ones). Thus the pathways were largely independent from one another. After extracting the most empirically-significant 30 SNPs over all RF iterations based on RF PVIM on the training dataset; and after applying the nested models on the independent test data in our case/control study, several 2 and 3-way interactions with a p-value < 0.05 were found from LRTs in different MGI-based pathways, but without passing Bonferroni correction. As we were testing all possible 2-way interactions between the top 30 SNPs, we ended up validating in our independent dataset (30*29)/2 = 435 possible interactions. Secondly, we observed that several SNP interactions which suggested risk for psychosis (p-value 0.05) also had a p-value lower than 0.05 in cognition (without Bonferroni correction).

4.3.1. MGI: Aggression-related behaviour phenotype pathway

Using logistic regression on the top 30 empirically-significant SNPs from RF analysis of our training data in an independent test dataset, we found twenty 2-way interactions and thirty-five 3-way interactions with a p-value < 0.05 when studying the association with psychosis. These 2-way and 3-way interactions were then tested for association with IQ and verbal IQ: two gene-gene interactions in case-control analysis showed a p-value < 0.05 in cognition, one with IQ and one with verbal IQ. One gene-gene interaction explained 1.04% of the variance in IQ, and the another accounted for 1.14% of variance in the verbal IQ (Table 4.2, Table 4.3.).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>STUDY</th>
<th>GENES</th>
<th>SNPs</th>
<th>P-VALUE</th>
<th>R²%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case/Control</td>
<td>LAMA2* CYP19A1</td>
<td>rs3749878 * rs934633</td>
<td>0.0091</td>
<td>0.95</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td>0.046</td>
<td>1.04</td>
</tr>
</tbody>
</table>

Table 4.2. 2-way interactions with p-value < 0.05 in psychosis and IQ in aggression pathway.

<table>
<thead>
<tr>
<th>STUDY</th>
<th>GENES</th>
<th>SNPs</th>
<th>P-VALUE</th>
<th>R²%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case/Control</td>
<td>MYO5C* ESR2</td>
<td>rs10163109*rs8006145</td>
<td>0.0024</td>
<td>1.27</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td></td>
<td></td>
<td>0.036</td>
<td>1.14</td>
</tr>
</tbody>
</table>

Table 4.3. 2-way interaction with p-value < 0.05 in psychosis and verbal IQ in aggression pathway.

From the 35 statistically significant 3-way interactions before correcting for multiple testing in the psychosis case-control study, four were found to influence both IQ and verbal IQ (Table 4.4.). However, these 3-way interactions did not involve the significant 2-way interactions observed in cognition and also did not remain statistically significant after Bonferroni correction in cognition.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>STUDY</th>
<th>GENES</th>
<th>SNPs</th>
<th>P-VALUE</th>
<th>R²%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case/control</td>
<td>GRIN1<em>PPT1</em>HYDIN</td>
<td>rs1126442<em>rs3131661</em>rs1798532</td>
<td>0.035</td>
<td>0.61</td>
</tr>
<tr>
<td>IQ</td>
<td>NTRK2<em>KIRREL3</em>HYDIN</td>
<td>rs1047896<em>rs3802815</em>rs1798532</td>
<td>0.007</td>
<td>1.82</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td></td>
<td></td>
<td>0.021</td>
<td>1.36</td>
</tr>
<tr>
<td>Case/control</td>
<td>NTRK2<em>ESR2</em>HYDIN</td>
<td>rs1047896<em>rs8006145</em>rs1798532</td>
<td>0.007</td>
<td>0.99</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td>0.019</td>
<td>1.73</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td></td>
<td></td>
<td>0.019</td>
<td>1.40</td>
</tr>
<tr>
<td>Case/control</td>
<td>MC5R<em>CYP19A1</em>CACNA1B</td>
<td>rs1541276<em>rs934633</em>rs11137342</td>
<td>0.019</td>
<td>0.76</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td>0.019</td>
<td>1.44</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td></td>
<td></td>
<td>0.034</td>
<td>1.17</td>
</tr>
</tbody>
</table>

Table 4.4. 3-way interaction with p-value < 0.05 in psychosis, IQ and verbal IQ in aggression pathway.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

4.3.2. MGI: Depression-related behaviour phenotype pathway

Nineteen significant 2-way interactions from the 26 empirically significant SNPs were found to be associated with psychosis before correcting by multiple testing. However, no significant results remained after the correction. Here I performed fewer tests in the independent test dataset: \((26 \times 25)/2 = 325\). Only one interaction showed a \(p\)-value < 0.05 with IQ and none of interactions had a significant impact on verbal IQ (Table 4.5.). We found thirty 3-way interactions with a \(p\)-value < 0.05 in psychosis, but again only one was significant in cognition before multiple testing, losing its impact on IQ after Bonferroni correction (Table 4.6.). The suggested epistatic effect from the 3-way interaction and the 2-way interaction before correcting by multiple testing come from different genes.

<table>
<thead>
<tr>
<th>STUDY</th>
<th>GENES</th>
<th>SNPs</th>
<th>(P)-VALUE</th>
<th>(R^2)%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case/control</td>
<td>GAD2*GRIN2A</td>
<td>rs2839677* rs9806806</td>
<td>0.023</td>
<td>0.71</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td>0.015</td>
<td>1.53</td>
</tr>
</tbody>
</table>

Table 4.5. 2-way interaction with \(p\)-value < 0.05 in psychosis and IQ in depression pathway.

<table>
<thead>
<tr>
<th>STUDY</th>
<th>GENES</th>
<th>SNPs</th>
<th>(P)-VALUE</th>
<th>(R^2)%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case/control</td>
<td>IPCEF1<em>SLITRK1</em>UBA6</td>
<td>rs2236259<em>rs9593836</em>rs4860853</td>
<td>0.044</td>
<td>0.56</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td>0.040</td>
<td>1.08</td>
</tr>
</tbody>
</table>

Table 4.6. 3-way interaction with \(p\)-value < 0.05 in psychosis and IQ in depression pathway.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

4.3.3. MGI: Fear/anxiety-related behaviour phenotype pathway

Twenty-five 2-way interactions were found to have some impact in psychosis, however without any significant evidence after correcting for multiple testing. Five of these interactions also had a \( p \)-value < 0.05, in both IQ and verbal IQ (Table 4.7.). Three interactions were between SNPs in the same genes, \( CRHR1 \) and \( ESR1 \). One SNP in \( ESR1 \) was statistically significantly involved in the three epistatic effects after multiple testing, and three different SNPs in \( CRHR1 \) were implicated in the interactions. These three SNPs in CRHR1 are in strong LD between each other \( (r^2=1 \) and \( D'=1 \) between rs16940665 and rs16940674; \( r^2=1 \) and \( D'=1 \) between rs16940665 and rs4640231; \( r^2=1 \) and \( D'=1 \) between rs16940674 and rs4640231) (The 1000 Genomes Project Consortium, 2015; using the British from England and Scotland population).

<table>
<thead>
<tr>
<th>STUDY</th>
<th>GENES</th>
<th>SNPs</th>
<th>P-VALUE</th>
<th>R²%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case/control</td>
<td>CRHR1*ESR1</td>
<td>rs16940665*rs2077647</td>
<td>0.003</td>
<td>1.17</td>
</tr>
<tr>
<td>IQ</td>
<td>CRHR1*ESR1</td>
<td>rs16940674*rs2077647</td>
<td>0.006</td>
<td>1.03</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>CRHR1*ESR1</td>
<td>rs4640231*rs2077647</td>
<td>0.003</td>
<td>1.17</td>
</tr>
<tr>
<td>Case/control</td>
<td>CRHR1*ESR1</td>
<td>rs16940665*rs2077647</td>
<td>0.043</td>
<td>1.06</td>
</tr>
<tr>
<td>IQ</td>
<td>CRHR1*ESR1</td>
<td>rs16940674*rs2077647</td>
<td>0.032</td>
<td>1.21</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>CRHR1*ESR1</td>
<td>rs4640231*rs2077647</td>
<td>0.033</td>
<td>1.19</td>
</tr>
</tbody>
</table>

Detecting significantly associated interactions with schizophrenia and cognition in abnormal behaviour and pathways from the Mouse Genotype Informatics (MGI) database
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>Verbal IQ</th>
<th>Case/control</th>
<th>MAPT*ESR1</th>
<th>rs62063776*rs2077647</th>
<th>0.032</th>
<th>1.21</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td></td>
<td>0.003</td>
<td>1.20</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td></td>
<td></td>
<td></td>
<td>0.021</td>
<td>1.39</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td>CTNS*ABCA2</td>
<td>rs2873624*rs7048567</td>
<td>0.004</td>
<td>1.12</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td></td>
<td></td>
<td></td>
<td>0.014</td>
<td>1.57</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td></td>
<td>0.007</td>
<td>1.88</td>
</tr>
</tbody>
</table>

Table 4.7. 2-way interactions with p-value < 0.05 in psychosis, IQ and verbal IQ in fear/anxiety pathway.

In addition, we found 3-way interactions with p-values < 0.05 in psychosis as well as in cognition in patients with psychosis. Seven of the fifty-three statistically significant interactions in psychosis were also linked with IQ and verbal IQ before correcting for multiple testing: one 3-way interaction was significant only in IQ (Table 4.8.), another only in verbal IQ (Table 4.9.) and five in both IQ and verbal IQ (Table 4.10.). Three of the 2-way interactions together with a third SNP also interacted (3-way interaction) showing a significant impact in IQ and verbal IQ from three interactions, two of which were between the same genes. Two interactions were observed among SNPs in CRHR1/ESR1/TOM1L2 from different two LD SNPs in CRHR1 ($r^2=1$ and $D^*=1$). The minor allele frequencies were higher than 20% in the 6 SNPs involved in these three interactions: being 21.43% and 21.44% for the SNPs in CRHR1; and 48.14%, 21.19% and 34.05% for the SNPs in ESR1, MAPT and TOM1L2, respectively.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>STUDY</th>
<th>GENES</th>
<th>SNPs</th>
<th>P-VALUE</th>
<th>R²%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case/control</td>
<td>TOM1L2<em>IDUA</em>UBA6</td>
<td>rs1108648<em>rs4690221</em>rs10794537</td>
<td>0.019</td>
<td>0.75</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td>0.031</td>
<td>1.19</td>
</tr>
</tbody>
</table>

Table 4.8. 3-way interaction with p-value < 0.05 in psychosis and IQ in fear/anxiety pathway.

<table>
<thead>
<tr>
<th>STUDY</th>
<th>GENES</th>
<th>SNPs</th>
<th>P-VALUE</th>
<th>R²%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case/control</td>
<td>ALS2<em>GM2A</em>CTNS</td>
<td>rs3219153<em>rs61740602</em>rs2873624</td>
<td>0.011</td>
<td>1.68</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td></td>
<td></td>
<td>0.031</td>
<td>1.19</td>
</tr>
</tbody>
</table>

Table 4.9. 3-way interaction with p-value < 0.05 in psychosis and verbal IQ in fear/anxiety pathway.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>STUDY</th>
<th>GENES</th>
<th>SNPs</th>
<th>P-VALUE</th>
<th>R²%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case/control</td>
<td>CRHR1<em>ESR1</em>TOM1L2</td>
<td>rs16940665<em>rs2077647</em>rs1108648</td>
<td>0.007</td>
<td>0.99</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td>0.032</td>
<td>1.18</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td></td>
<td></td>
<td>0.048</td>
<td>1.01</td>
</tr>
<tr>
<td>Case/control</td>
<td>CRHR1<em>ESR1</em>TOM1L2</td>
<td>rs4640231<em>rs2077647</em>rs1108648</td>
<td>0.007</td>
<td>0.99</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td>0.031</td>
<td>1.18</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td></td>
<td></td>
<td>0.047</td>
<td>1.01</td>
</tr>
<tr>
<td>Case/control</td>
<td>MAPT<em>ESR1</em>TOM1L2</td>
<td>rs62063776<em>rs2077647</em>rs1108648</td>
<td>0.010</td>
<td>0.91</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td>0.020</td>
<td>1.38</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td></td>
<td></td>
<td>0.036</td>
<td>1.13</td>
</tr>
<tr>
<td>Case/control</td>
<td>ALS2<em>GM2A</em>CTNS</td>
<td>rs3219153<em>rs61740602</em>rs222754</td>
<td>0.001</td>
<td>1.37</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td>0.006</td>
<td>1.92</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td></td>
<td></td>
<td>0.003</td>
<td>2.27</td>
</tr>
<tr>
<td>Case/control</td>
<td>NOS1<em>CHD6</em>ADCY1</td>
<td>rs3741475<em>rs3746543</em>rs2471267</td>
<td>0.036</td>
<td>0.61</td>
</tr>
<tr>
<td>IQ</td>
<td></td>
<td></td>
<td>0.031</td>
<td>1.18</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td></td>
<td></td>
<td>0.009</td>
<td>1.73</td>
</tr>
</tbody>
</table>

Table 4.10. 3-way interaction with p-value < 0.05 in psychosis, IQ verbal IQ in fear/anxiety pathway.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

4.3.4. MGI: Response to novel object phenotype pathway

In the response to novel object pathway there were twenty-four 2-way and thirty 3-way interactions that had an uncorrected $p$-value $< 0.05$ (but Bonferroni corrected $p$-value $> 0.05$). However, no statistically significant interactions after Bonferroni correction were found to be related with case/control status or either IQ or verbal IQ in 2- or 3-way interactions.

4.4. Discussion

As previous studies have shown, the aetiology of psychosis is complex. While it appears that genetic factors play an important role, only a small fraction of these factors have been identified. The additive effect from PRS has not been able to explain a large amount of variation in psychosis case status. In this study, I tested genetic interactions that could contribute to our understanding of the molecular pathology of psychosis. However, I performed many tests in each pathway and none of the interactions passed Bonferroni correction for multiple testing.

The fact that I did not find significant evidence for interactions influencing risk for psychosis or cognition might result from the limitations of the study. The sample size is small, especially for analyses of cognition. Due to the weak associations between SNPs that are involved in psychotic disorders, relevant interactions might be hidden and therefore, might not be detected because of the small sample sizes. Hence, when finding an independent dataset with larger sample size (mainly in the cognition database) to replicate the results of the present study, it would be crucial to assure whether these interactions are relevant for the risk of psychosis and influence cognition in cases.

In addition, the study design required a balanced training dataset because RF is biased when the data are unbalanced, tending to prefer the category with the higher sample
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

size. But the other way to design the study would be to take 80% of the cases and controls and leave the remaining 20% for test samples. The associations from SNPs are weak (single effects) in complex disorders, which could lead in a different subset of SNPs being in the top 30 empirically significant SNPs, some of which could be in LD with the ones involved in the resulting interactions. These new SNPs might show statistical significance after Bonferroni correction in our independent dataset at least in one of the phenotypes, psychosis or cognition.

In this study the number of random selected variables within RF in each pathway was chosen as the default value for classification studies - the square root of the total number of variables. It has been shown that RF did not classify well using small values of mtry in high dimensional data such as in GWAS (Wu et al. 2012). Furthermore, when working with correlated predictors, applying RF based on unconditional PVIMs with large values for mtry can inflate the VIM of the predictors which are correlated with the true predictor (Nicodemus et al. 2010c). In this study correlation between SNPs was taken into account since SNPs were not pruned. Therefore, in further studies the optimal value of mtry could be estimated by cross-validation (CV) rather than using the default value, although this would be time consuming.

To try to reduce the dimensionality of the data at the start, I included only exonic SNPs - both missense and synonymous, 3 prime and 5 prime. However, the largest study to date found 108 variants associated with schizophrenia, most being non-exonic variants (Ripke et al. 2014). Therefore, further research should consider intron SNPs and reduce the dimensionality only using RF in the training dataset; these SNPs might interact with others, thus increasing the risk of psychosis and perhaps explaining variance in cognition.

In genetics, it is very common to use Bonferroni correction for determining statistical significance despite multiple testing, but this is rather conservative and might hide true effects. Instead, false discovery rate (FDR) could have been used in order to determine significance as it is less conservative and tries to capture the most amount of true
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

positives with the cost of increased numbers of false positives (Benjamini and Hochberg 1995). Also, Bonferroni correction does not take into account dependency between variables, so Benjamini & Yekutieli could have been considered in the study as this method is good to be used when there is correlation between variables (Benjamini and Yekutieli 2001).

The number of ‘top’ empirically-significant variables is arbitrary: I decided to take the top 30 to minimise multiple testing in the test data. Taking a smaller number of top SNPs would further reduce multiple testing and any observed interaction might pass Bonferroni correction. But using fewer top SNPs limits the possible interactions between SNPs which can be examined.

This study tested all possible 2-way interactions but only the 3-way interactions including the 2 SNPs involved in statistically significant 2-way interactions (p-value < 0.05) in order to minimise the number of tests. In the cognition study only interactions which were statistically significant (p-value < 0.05) in the psychosis study were investigated. To minimise multiple testing in future real studies, one solution could be to test just the SNPs which are involved in 3-way interactions in the psychosis study with p-values less than 0.05 for 2-way interactions in the cognition study. However, this might miss relevant true 2-way interaction effects. In addition, there could be effects from higher order interactions between SNPs.

The only pathway that demonstrated both significant 2-way and 3-way interactions (p-value < 0.05) before correcting for multiple testing in verbal IQ and IQ was fear/anxiety. If these interactions had been significant after correcting for multiple testing, it would have lead us to a connection between psychosis and cognition. Information from the genes involved on the interactions (p-value < 0.05) was investigated as an example of what would be have done if the interactions had been significant after multiple testing.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

These interactions involved SNPs in *Corticotropin Releasing Hormone Receptor 1 (CRHR1)*, *Estrogen Receptor 1 (ESR1)* and *Target Of Myb1 Like 2 Membrane Trafficking Protein (TOM1L2)* genes, and in *microtubule-associated protein tau (MAPT)*, *ESR1* and *TOM1L2*. Interactions were found between *CRHR1* and *ESR1*, and between *MAPT* and *ESR1*, and among *CRHR1*, *ESR1* and *TOM1L2*, and *MAPT*, *ESR1* and *TOM1L2*. Even though these findings did not pass multiple testing correction, the genes involved in these interactions have been previously associated with cognition and psychosis.

*CRHR1* codes for receptor of corticotrophin releasing hormone (CRH) locus located at the chromosome 17. It has shown to contribute to risk for depression and psychosis as well as being related to response to antidepressants (Schatzberg *et al.* 2014). It has also been associated with the excitement dimension of BP which is related to mania (Leszczyńska-Rodziewicz *et al.* 2013). CRH through its receptor *CRHR1* is a neurotransmitter, having an impact on the hypothalamic-pituitary-adreanal (HPA) axis, and is related to response to stress, both cognitive and behaviour (Funk *et al.*, 2006); (Polanczyk *et al.*, 2009). The HPA axis shows higher activity in people with MDD with psychosis than healthy people (Keller *et al.* 2006); (Lembke *et al.* 2013). Moreover, it is also involved with cognitive and mood disorders among others (da Silva *et al.* 2016); (Grimm *et al.* 2017).

*ESR1* codes for the α receptor of the Estrogen hormone (Greene *et al.* 1986) on chromosome 6 and has function in brain areas related to emotion (Amin *et al.* 2005) and cognition (Berman *et al.*, 1997); (Osterlund *et al.*, 2000); (Osterlund and Hurd, 2001). SNPs in *ESR1* have been associated with osteoporosis (Ioannidis *et al.* 2004), cancer (Cai *et al.* 2003), cognitive decline (such as episodic memory) and dementia (Ma *et al.* 2014). Other studies have shown an association between *ESR1* mRNA levels and schizophrenia as well as with MDD (Perlman *et al.*, 2004; (Perlman *et al.*, 2005). Moreover, SNPs in the *ESR1* gene have been shown to contribute to risk for schizophrenia and MDD (Ryan *et al.* 2012; Ryan and Ancelin 2012). Różycka *et al.*
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics. (2016) suggested epistasis that was involved with the risk of depression, in particular, interactions between SNPs in COMT and ERS1 (Różycka et al. 2016).

Although there is not much knowledge about the TOMIL2 gene’s function, my findings showed significant interactions including the gene. The TOMIL2 gene is located on chromosome 17 and is expressed primarily in the brain and heart. It has expression in a region deleted in the vast majority of patients with Smith-Magenis syndrome (SMS) (Bi et al. 2002). Individuals with SMS show neurocognitive impairment such as verbal delay, consistent with our finding that this interaction is associated not only with IQ, but also with verbal IQ in the WTCCC2. In mice, the gene has been related to cellular trafficking and immune response, as well as being involved in tumor suppression (Girirajan et al. 2008).

The gene MAPT is also located on chromosome 17 and is expressed in the nervous system including in neurons (Neve et al. 1986). Epistatic or single effects involving this gene have been associated with neurodegenerative disorders such as Parkinson disease (Elbaz et al. 2011); (Yu et al. 2014), Alzheimer's disease (Kwok et al. 2008); (Zhang et al. 2011) and Frontotemporal Dementia (Verpillat et al. 2002). All these disorders are characterized by cognitive impairment. CRHR1, the other gene interacting with ERS1 and TOMIL2, is the 5-prime to MAPT on the genome (Poorkaj et al. 2001).

Although molecular interaction between CRHR1/ERS1, MAPT/ESRI, CRHR1/ERS1/TOMIL2 and CRHR1/ESRI/TOMIL2 has not been reported previously, I intend to study this further in the future. The amount of variation in psychosis case status and cognition within psychosis cases could enable us to conclude that epistasis might contribute more to the genetic architecture of psychosis than PRS.

Animal models are an extremely useful tool for studying the genetic effects of psychosis candidate genes. The advantage of animal models in genetic studies of psychosis is that one can isolate the effects of single genes by using either partial or
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

full gene-knock-out (i.e. removing one or two copies of the gene), knock-in (i.e. addition of an extra copy of the gene), or transgenic models, where copies of human genes are inserted into animal genomes. After genetic mutants are created, they can be deeply phenotyped for a range of cognitive and affective behavioural measures. For instance, studies using knock-out mouse have found an association of CRH1 and ERS1 with abnormal anxiety-related response (Timpl et al. 1998); (Müller et al. 2003); (Refojo et al. 2011) and MAPT has also been associated with anxiety behaviour in mice (Sennvik et al. 2007).

The use of any curated database - such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000) or the MGI database used here - will be limited by current knowledge. However, the use of relevant phenotypes derived from mutant mouse models is a step forward in understanding how these genes may interact and could provide evidence for assessing the phenotypes of double mutants in future studies.

In conclusion, I was unable to find significant evidence for interaction between functional SNPs in the MGI pathways examined. In order to find replication in psychosis, IQ and verbal IQ, future work will need to test the three phenotypes again in a dataset with a larger sample size, mostly in cognition. The use of ML to detect replicated epistasis is an attractive addition to the psychiatric genomics toolkit (Nicodemus et al. 2010a); (Nicodemus et al. 2010b). The present study represents a computational approach that is amenable to investigation in model organisms to understand the underlying biology.
5. Conclusions and future directions

5.1. Summary of thesis

In psychiatric genetics, GWAS has been useful for the discovery of loci playing an important role in a number of diseases: PGC schizophrenia (Ripke et al. 2014), PGC bipolar (Hou et al. 2016b), PGC MDD (Power et al. 2017). The dimensionality of genome-wide data is high, and it becomes much higher when looking for interactions, the subject of this thesis. Searching for interactions in genetic data poses challenging statistical issues including the characteristic of “n<<p” (more variables than observations). ML techniques, such as the RF algorithm, have been used to overcome these hurdles. This technique has been attractive both when studying single gene associations (Goldstein et al. 2010) and epistasis (Lunetta et al. 2004); (Nicodemus et al. 2010a); (Nicodemus et al. 2010b); (Winham et al. 2012). In this section the aims of my PhD and the progress toward these aims are summarised.

5.1.1. Aim 1

The first aim of the thesis was to compare the behaviour of different VIMs and the related measure minimal depth, in order to detect single and interaction effects under predictor correlation conditions. This allowed me to examine which VIM is more suitable when both predictors and outcome are continuous. This aim was addressed in Chapter 2, where I report a simulation study in which different synthetic datasets were generated under nine different correlation conditions in two association models, strong and weak, the latter one being what one might expect in complex disorders. Two VIMs were derived from CIF, and six VIMs and minimal depth were derived from RF. There has been previous research studying the performance of some VIMs and minimal depth to capture single effects under predictor association (Strobl et al. 2008); (Nicodemus and Malley 2009); (Nicodemus et al. 2010c) as well as interaction effects (Wright et al. 2016). However, this thesis describes the first study to include different numbers
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

of predictors with different level of inter-predictor correlation with continuous predictors and continuous outcome.

The simulation study suggested that the different VIMs and minimal depth perform differently depending on the correlation of predictors. Furthermore, correlation between predictors, and the number of correlated variables had an impact on all VIMs to some degree when detecting single and interactions effects. Indeed, some of the VIMs were shown to be biased even when predictors were uncorrelated. However, the unconditional unscaled PVIM from RF and the two from CIF were found to be unbiased, and were able to capture both single and interaction effects even under conditions of predictor correlation. However, the two unconditional permutation VIMs examined from CIF were computationally intractable, for instance around twenty iterations were possible per day for these two PVIMs compared to over one hundred per day for the unconditional unscaled PVIM.

The knowledge gained from the simulation was applied in a case-control study of schizophrenia, using 39 different cohorts from the PGC2 database, to study single and interaction effects of SNPs. Because of PS, the SNPs and the phenotype became continuous variables (Price et al. 2006) as was the case in the simulation study. Two single SNPs showed evidence for association with schizophrenia, one of which was a novel finding. One SNP was in the ACAT2 gene and the other in the TNC gene. ACAT2 is a gene involved in the cholesterol biosynthesis, and significant pathways associated with schizophrenia pathways included this gene (Prabakaran et al. 2004). TNC has not been previously related to schizophrenia, but it is expressed in the brain and involved in neuronal migration, and so it might be associated to psychiatric disorders or to cognition.

This first approach has provided insight into the extent to which VIMs are useful in real situations when using RF to avoid spurious results. Moreover, a real study was performed applying RF to find single and epistatic effects, which found SNPs that may
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

have an important role in schizophrenia. This could help understanding the mechanisms underlying that complex disease.

5.1.2. Aim 2

The second aim of my PhD was to examine the VIM\textsubscript{Gini} when predictors were all continuous and independent of each other. This analysis was based upon suggestions from a previous study: that VIM\textsubscript{Gini} may be preferred with this type of predictor under those conditions (Boulesteix \textit{et al.} 2012b). It was also suggested in that study that it may be preferable to apply VIM\textsubscript{Gini} when the signal-to-noise ratio is low. As these suggestions have not been examined before, the present study did do, as well as investigating the behaviour of VIM\textsubscript{Gini} when the error had two different variances.

In pursuit of this aim, VIM\textsubscript{Gini} was performed in a simulation study in four different cases: (1) when all predictors and error followed a standard normal distribution; (2) when all predictors were normally distributed with different variances and the error followed a standard normal; (3) when all predictors and error had standard normal distributions but predictors keep varying precisions; and (4) when predictors were standard normally distributed but the error had less variance (0.25). A comparative analysis between (1) and the other cases was performed to verify the suggestions made by Boulesteix \textit{et al.}, (2012b) in Chapter 3, considering both continuous and binary outcomes.

The results of this study showed that VIM\textsubscript{Gini} is biased towards predictors with higher precision, with either continuous or binary outcomes, even under conditions of no association (H\textsubscript{0}), since they had a greater number of cut-points. Lower precision leads to lower importance scores because there are fewer unique values. This finding was related to the fact that VIM\textsubscript{Gini} is based on the Gini index, which is most likely to split the variable with the most cut-points. Moreover, when the outcome was continuous, VIM\textsubscript{Gini} was shown to be biased towards error with greater variance, both with and without association between predictor and the outcome. When predictors were
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics. Associated, VIM\textsubscript{Gini} inflated the importance scores of predictors with greater variance, even though the effect size was the same, when the outcome was either continuous or binary.

This thesis has found two additional sources of bias to those reported in the literature, (1) when predictors have been measured with variable precision regardless of whether the outcome is continuous or binary; and (2) when the error variance is higher for continuous outcomes. To minimize the risk of spurious results when using VIM\textsubscript{Gini}, it would be sensible to standardise variables and to use the same number of decimal places for all predictors. However, as VIM\textsubscript{Gini} is biased when the error has greater variance in the case of continuous outcomes, the use of an alternative VIM is suggested by these data, such as any of the unscaled permutation VIMs.

The results from this study are important in order to avoid misleading results in real studies. It also highlights that researchers should be aware of which VIM is used by default in the software they are using, as some R and Python packages use VIM\textsubscript{Gini} as the default.

5.1.3. Aim 3

Based on the results of Chapter 2 and Chapter 3, the third aim of my PhD thesis was to apply RF, based on the unconditional unscaled permutation VIM, to the study of epistasis (2-way and 3-way interactions) in both psychosis and two cognitive phenotypes (IQ and verbal IQ). SNPs were selected from genes belonging to MGI pathways that were previously implicated in behavioural phenotypes in mice (aggression, depression, fear/anxiety and novelty). Genotype data for these SNPs from the WTCCC2 Irish cohort was analyzed in a case-control study. Using RF in a training dataset to prioritize the top 30 empirically significant SNPs reduced the number of SNPs for follow-up analysis in the independent test dataset, although the amount of multiple testing in the training set was still large. In an independent dataset, LRTs were applied to test for interaction between SNPs in each behavioural pathway. The SNPs

Conclusions and future directions 200
Conclusions and future directions

Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

involved in statistically significant 2-way interactions with psychosis before Bonferroni correction were then tested for involvement in 3-way interactions with psychosis. The 2-way and 3-way interactions that showed p-values less than 0.05 (uncorrected) were tested for interaction in a cognition outcome (IQ and verbal IQ) in cases. No evidence was found for 2-way or 3-way interactions for either psychosis or cognition, after correcting for multiple testing.

5.2. Strengths of the study

The work presented in this thesis has several limitations (see section 5.2), although it also has some strengths. In the first simulation study, the different VIMs and minimal depth were applied to synthetic data with a large number of iterations, and 100 variables because of correlation between predictors. This analysis replicated previously findings, i.e. that the strength of correlation had an impact on different VIMs or minimal depth when capturing single and interaction effects (Strobl et al. 2008); (Nicodemus and Malley 2009); (Nicodemus et al. 2010c); (Nicodemus 2011); (Wright et al. 2016). In addition, the number of correlated variables had an important role in determining the behavior of the different VIMs and minimal depth, especially under high correlation conditions. These findings replicate those of Nicodemus (2009), but also extend them as Nicodemus (2009) did not study all the VIMs included in this thesis nor minimal depth. The work presented here can be used to decide which VIM should be used when applying RF in real studies, where several predictors are correlated to a given degree (low, medium or high). In Chapter 2, the results of the simulation study were applied when studying single and interaction effects in a case-control schizophrenia study with a large sample size. Two single SNPs were shown to have an impact with schizophrenia (one of them was a novel finding), after combining all independent 39 test datasets. The two SNPs were tested in each independent dataset because they had been significant in each training dataset. The results from all tests performed in the independent tests were combined taking into account the p-value, the direction of the coefficients and the sample size.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

The results from the second simulation study are essential to consider when working with real data, when predictors are continuous and uncorrelated, to avoid spurious results. The VIM\textsubscript{Gini} biases reported for the first time in Chapter 3 suggest that researchers should check which VIM is being used by default when RF is applied using different packages. When the outcome is binary, one should standardise the predictors and round them to the same number of decimal places and use of other VIMs when the outcome is continuous.

In terms of strength of the study design, both the real study described in Chapter 4 and the application in Chapter 2 included genotyping quality control. Furthermore, HWE was tested in controls to detect genotyping errors in the studies. In Chapter 4, if significant interactions had been found, the use of gene pathways, which have been shown to be relevant to behavioural phenotypes in mice, would have allowed us to also link these genes to a group of different phenotypes (psychosis, IQ and verbal IQ).

The use of RF in the training sets in both studies helped to filter out SNPs that were not empirically significant. Moreover, as RF provided the importance of each SNP, I could order them by importance scores and take a subset of them (top 30). Therefore, the number of tests in the independent test dataset was lower than in the situation when all interactions between empirically significant SNPs were considered, despite the fact that a large amount of tests were still performed.

### 5.3. Limitations of the study

Despite the several strengths of this study, my project also has several limitations, mainly in study design. Although the number of random variables selected (mtry) in RF has been shown to have an impact on the VIMs (Nicodemus \textit{et al.} 2010c); (Wu \textit{et al.} 2012), and it has been suggested that this should be assessed empirically (Nicodemus \textit{et al.} 2010c). In both simulation studies and in both real performed here in the study, a fixed value of mtry was considered (except for minimal depth). This
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

less than ideal procedure was followed as the time consuming nature of the simulations made it intractable to optimise the value of mtry.

In the simulation study performed in Chapter 2, when studying weakly and strongly associations from interactions, only one type of model was considered which included two main effects from independent variables and the interaction between them. One of the limitations of this study is the lack of other types of interaction models such as models that include only interaction between SNPs without main effects (between correlated and uncorrelated SNPs) and models with main and interaction effects but with interaction between variables that are not the one with main effects, as the detection of the interaction could have been masked by the main effects. In fact, the inclusion of models with only interactions would be helpful to know if the VIMs were detecting the interaction because of the actual interaction effect and not because of the main effects, and in this way ensure the detection of interaction, as previously done by Wright et al. (2016) who included different type of models.

In the applied study described in Chapter 2, each gene had a limited number of SNPs, which reduces the ability to capture causal variants within the genes. Also, the small number of SNPs may have led to a failure to detect epistasis. To determine the significance of either single SNPs or the interaction between them, the SNPs that were empirically significant in each training dataset were tested in the corresponding independent test. A different subset of the total collection of empirically significant SNPs was identified in each training dataset. Significant SNPs were reported by a 3-step process because of computational constraints; those which were tested in all test datasets and showing significance after combining all results. This means that some SNPs that perhaps should have been defined as significant might not have been because significance was defined in that way. Instead, it might have been worthwhile to select those SNPs that were found to be significant in at least one training dataset and in the independent database has evidence for association with schizophrenia. This should be done in future research.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Other limitation of this thesis is the lack of a real applied study in Chapter 3. A real application of VIM\textsubscript{Gini} using datasets from different sources which include variables with different precision and different variability should be considered on future research. This application study would help to ensure that those situations can be found in real studies and that the consideration of variables with different precision was not artificial. Also, the real application would show how to deal in real situations when applying VIM\textsubscript{Gini} showing that variables should be rounded with the same number of decimal places as well as normalised, although other VIM would be recommended anyway when the outcome was continuous because of the bias toward error with more variance.

The real study presented in the Chapter 4 was designed to include a balanced training dataset while also considering balanced samples with RF. However, given the RF step was balanced, it was not necessary to balance the training dataset. It might have been better to select 80% of cases and controls for the training dataset and leave the 20% of both cases and controls for the independent test dataset, to have the same distribution of cases and controls in both datasets, as the sample was balanced within RF. Furthermore, the data were not LD pruned, so the study was performed using correlated predictors. The VIM applied in the study (unconditional unscaled permutation VIM) showed, in the simulation study, to have a power between 36.19% and 64.59%, depending on the number of variables that were correlated, to detect correlated interacting true predictors under high correlation conditions. Moreover, the power to detect the uncorrelated interacting associated predictor ranged from 50.90% to 65.78%, depending on the number of correlated predictors, under high correlation conditions. Thus, the sample size of the study might be small to capture additional interacting variants. The lack of a replication dataset with which to test the significant results before correcting for multiple testing is one of the main limitations of the present study.

In addition, the arbitrary selection of the top 30 SNPs reduced the number of empirically significant SNPs and, therefore the number of 2-way and 3-way
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

interactions tested between them in the independent test dataset, except for the depression pathway. The consideration of a subset of SNPs with higher importance scores within the group of empirically significant SNPs helps to reduce the multiple testing in the test dataset. However, not all possible interactions are validated in the independent dataset, which may have led to missing significant 2-way or 3-way interactions.

5.4. Future directions

The use of the *ranger* package in R to optimise the value of mtry by cross validation, under the nine different correlation conditions considered in the first simulation study, would be a sensible next step. Furthermore, under the same correlation conditions, the ability of the different VIMs and minimal depth to capture interaction effects should be investigated in other type of interacting models, such as when no main effects are involved in the model.

Considering the results from the applied study in Chapter 2, future work should take into account the interactions and single effects that were significant in at least one training dataset and also influential in the independent dataset, rather than only the models that were tested in all independent test datasets. The study from which the pathway was taken was focus on biomarkers that have not proved biological interactions (epistasis) (Chan et al. 2015). Thus, future research should select other gene pathways that may be involved in schizophrenia, for testing for epistasis in the PGC2 schizophrenia case-control data.

The number of random variables selected (mtry) in RF in the real study of Chapter 4 should also be assessed empirically using ranger. Further research should be focused on finding a replication dataset, which includes phenotypic data for both psychosis and the endophenotypes (IQ and verbal IQ) to replicate the results of this study. Finding an independent database that corroborates the significant effect before multiple testing

Conclusions and future directions 205
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics. might be more helpful than applying Bonferroni correction, which is overly conservative.

In addition, research into the role of other abnormal behaviour pathways in Generation Scotland should also be considered for study, testing for gene-gene interactions that may be associated with psychotic disorders. Generation Scotland has the genotype information from healthy people and from patients with psychotic disorders (more than 20,000 individuals in total), it also has the information from several cognitive variables such as IQ, verbal IQ and social impairment (between others). People with psychotic disorders usually present social dysfunctions, therefore one possible future study to do is to study the association of single and interaction effects between genes of the abnormal behaviour pathways (from Chapter 4) with social impairment including individuals with psychotic disorders and healthy individuals in Generation Scotland. In this way, a link between cognition, disease and genetics could be found.

5.5. Conclusions

When trying to find single gene and gene-gene interactions that influence risk for psychotic disorders, ML techniques such as RF are useful. Investigating the performance of the different VIMs and minimal depth under correlation conditions that may be present in real studies is one of the main contributions of my thesis. The simulation study results are useful for researchers who are analysing genetic interactions and single associations in presence of correlation; the results may be used as a guideline. In addition, the results of the second simulation should be considered, and researchers should be aware of the issues associated with the use of this VIM_{Gini} in real studies.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

References

Abi-Dargham A, Rodenhiser J, Printz D, Zea-Ponce Y, Gil R, Kegeles LS, Weiss R, Cooper TB, Mann JJ, Van Heertum RL, Gorman JM, Laruelle M (2000) Increased baseline occupancy of D2 receptors by dopamine in schizophrenia. Proceedings of the National Academy of Sciences of the United States of America 97:8104–9


American Psychiatric Association (2013) Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


References 208
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

10.1111/j.1582-4934.2011.01360.x


References 210
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


References 212
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

10.1016/S0140-6736(03)12379-3


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Irish Schizophrenia Genomics Consortium and the Wellcome Trust Case Control Consortium 2 (2012) Genome-wide association study implicates HLA-C*01:02 as a risk factor at the major histocompatibility complex locus in schizophrenia. Biological Psychiatry 72:620–8. doi: 10.1016/j.biopsych.2012.05.035


References
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

10.1016/S2215-0366(15)00577-5

Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

machine learning and data mining methods to detect associations of rare and common variants with complex traits. Genetic Epidemiology 38 Suppl 1:S81-5 . doi: 10.1002/gepi.21830


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Michie D, Spiegelhalter DJ, Taylor CC (1994) Machine Learning, Neural and Statistical Classification


MQ T mental health (2015) MQ lanscape analysis - UK Mental Health Research Funding

Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


References 226
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


References 229
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81:559–75. doi: 10.1086/519795


References 230
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


References
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

doi: 10.1017/thg.2012.7


Studying the ability of finding single and interaction effects with Random Forest, and
its application in Psychiatric genetics.

implementation of Random Forests for high-dimensional data. *Bioinformatics*

Scott LJ, Muglia P, Kong XQ, Guan W, Flickinger M, Upmanyu R, Tozzi F, Li JZ,
Burmeister M, Absher D, Thompson RC, Francks C, Meng F, Antoniades A,
Southwick AM, Schatzberg AF, Bunney WE, Barchas JD, Jones EG, Day R,
Matthews K, McGuffin P, Strauss JS, Kennedy JL, Middleton L, Roses AD,
Watson SJ, Vincent JB, Myers RM, Farmer AE, Akil H, Burns DK, Boehnke M
(2009) Genome-wide association and meta-analysis of bipolar disorder in
individuals of European ancestry. *Proceedings of the National Academy of
Sciences of the United States of America* 106:7501–6. doi:
10.1073/pnas.0813386106

*CBMB Working Paper*

Tomiyama T, Mori H, Krugers H, Joels M, Ramakers GJA, Lucassen PJ, Van
Leuven F (2007) Tau-4R suppresses proliferation and promotes neuronal
differentiation in the hippocampus of tau knockin/knockout mice. *The FASEB

Serretti A, Mandelli L (2008) The genetics of bipolar disorder: genome “hot regions,”
genes, new potential candidates and future directions. *Molecular Psychiatry*
13:742–771. doi: 10.1038/mp.2008.29

Sham PC, Lin MW, Zhao JH, Curtis D (2000) Power comparison of parametric and
nonparametric linkage tests in small pedigrees. *American Journal of Human
Genetics* 66:1661–8. doi: 10.1086/302888

and adoption studies for a genetic contribution to adult psychiatric disorders.
*International Review of Psychiatry* 16:260–83. doi:
10.1080/09540260400014401

Simmons JM, Quinn KJ (2014) The NIMH Research Domain Criteria (RDoC) Project:
implications for genetics research. *Mammalian Genome* 25:23–31. doi:
10.1007/s00335-013-9476-9

Sklar P, Ripke S, Scott LJ, Andreassen OA, Cichon S, Craddock N, Edenberg HJ,
Nurnberger JI, Rietschel M, Blackwood D, Corvin A, Flickinger M, Guan W,
Mattingsdal M, McQuillin A, Kwan P, Wienker TF, Daly M, Dudbridge F,
Holmans PA, Lin D, Burmeister M, Greenwood TA, Hamshere ML, Muglia P,
Smith EN, Zandi PP, NIEVERGELT CM, McKinney R, Shilling PD, Schork NJ,
Bloss CS, Foroud T, Koller DL, Gershon ES, Liu C, Badner JA, Scheftner WA,
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

References


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

References

10.1038/nature08186


Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


References 238
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


References 239
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


References
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.


(2016) The MQ manifesto for young people’s mental health

(1946) WHO | World Health Organization. In: WHO
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Appendix A

Tables and Figures chapter 2

<table>
<thead>
<tr>
<th>WAC</th>
<th>r=0.80</th>
<th>r=0.40</th>
<th>R=0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>BIAS</td>
<td>COV%</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-0.0007</td>
<td>94.4</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.00005</td>
<td>95.6</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.00107</td>
<td>93.2</td>
</tr>
</tbody>
</table>

Table A.1. Bias and coverage of $V_2$ under the null hypothesis.

<table>
<thead>
<tr>
<th>Uncorrelated</th>
<th>r=0.80</th>
<th>r=0.40</th>
<th>r=0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>SAC single</td>
<td>0.00019</td>
<td>-0.0001</td>
<td>0.00012</td>
</tr>
<tr>
<td>WAC single</td>
<td>-0.00001</td>
<td>-0.00004</td>
<td>-0.00007</td>
</tr>
<tr>
<td>SAC interaction</td>
<td>0.00005</td>
<td>0.00006</td>
<td>-0.0001</td>
</tr>
<tr>
<td>WAC interaction</td>
<td>-0.0002</td>
<td>0.00015</td>
<td>0.00031</td>
</tr>
</tbody>
</table>

Table A.2. Medians of the observed correlation between the correlated predictors.

<table>
<thead>
<tr>
<th>Correlated</th>
<th>r=0.80</th>
<th>r=0.40</th>
<th>r=0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>SAC single</td>
<td>0.795</td>
<td>0.800</td>
<td>0.799</td>
</tr>
<tr>
<td>WAC single</td>
<td>0.795</td>
<td>0.799</td>
<td>0.800</td>
</tr>
<tr>
<td>SAC interaction</td>
<td>0.794</td>
<td>0.799</td>
<td>0.800</td>
</tr>
<tr>
<td>WAC interaction</td>
<td>0.795</td>
<td>0.799</td>
<td>0.800</td>
</tr>
</tbody>
</table>

Table A.3. Medians of the observed correlation between the uncorrelated predictors.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

### Table A.4. Percentage of importance scores greater than or equal to the cut-off across all 500 null VIMs or minimal depth.

<table>
<thead>
<tr>
<th>N</th>
<th>5</th>
<th>20</th>
<th>40</th>
<th>5</th>
<th>20</th>
<th>40</th>
<th>5</th>
<th>20</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>gini</td>
<td>5.00</td>
<td>5.15</td>
<td>5.06</td>
<td>5.11</td>
<td>5.13</td>
<td>5.08</td>
<td>5.14</td>
<td>5.08</td>
<td></td>
</tr>
<tr>
<td>rawpermRF</td>
<td>4.95</td>
<td>5.82</td>
<td>6.29</td>
<td>4.92</td>
<td>5.09</td>
<td>5.24</td>
<td>4.97</td>
<td>5.05</td>
<td>4.99</td>
</tr>
<tr>
<td>Breiman</td>
<td>5.05</td>
<td>5.58</td>
<td>5.99</td>
<td>4.99</td>
<td>5.29</td>
<td>5.42</td>
<td>5.07</td>
<td>5.13</td>
<td>5.10</td>
</tr>
<tr>
<td>Liaw</td>
<td>5.05</td>
<td>5.58</td>
<td>5.99</td>
<td>4.99</td>
<td>5.29</td>
<td>5.42</td>
<td>5.07</td>
<td>5.13</td>
<td>5.10</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>4.89</td>
<td>4.92</td>
<td>4.94</td>
<td>4.87</td>
<td>4.89</td>
<td>4.97</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Party</td>
<td>4.96</td>
<td>5.00</td>
<td>5.05</td>
<td>4.85</td>
<td>4.98</td>
<td>4.95</td>
<td>4.88</td>
<td>5.01</td>
<td>4.98</td>
</tr>
<tr>
<td>AUC</td>
<td>4.96</td>
<td>5.00</td>
<td>5.05</td>
<td>4.85</td>
<td>4.99</td>
<td>4.96</td>
<td>4.88</td>
<td>5.00</td>
<td>4.98</td>
</tr>
<tr>
<td>mindepth 39</td>
<td>4.96</td>
<td>5.00</td>
<td>5.05</td>
<td>4.85</td>
<td>4.99</td>
<td>4.95</td>
<td>4.89</td>
<td>5.00</td>
<td>4.98</td>
</tr>
<tr>
<td>mindepth 27</td>
<td>4.98</td>
<td>5.11</td>
<td>5.11</td>
<td>4.93</td>
<td>5.03</td>
<td>5.13</td>
<td>4.97</td>
<td>4.97</td>
<td>4.98</td>
</tr>
</tbody>
</table>

### Table A.5. Median of VIM and minimal depth medians for the correlated variables under H0.

<table>
<thead>
<tr>
<th>N</th>
<th>5</th>
<th>20</th>
<th>40</th>
<th>5</th>
<th>20</th>
<th>40</th>
<th>5</th>
<th>20</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>gini</td>
<td>1.25</td>
<td>1.07</td>
<td>1.44</td>
<td>1.37</td>
<td>1.37</td>
<td>1.5</td>
<td>1.48</td>
<td>1.48</td>
<td></td>
</tr>
<tr>
<td>rawpermRF</td>
<td>0.0003</td>
<td>0.0012</td>
<td>0.002</td>
<td>0.0006</td>
<td>0.00036</td>
<td>0.00068</td>
<td>-0.00007</td>
<td>-0.00006</td>
<td>-0.00002</td>
</tr>
<tr>
<td>Breiman</td>
<td>1.13</td>
<td>3.57</td>
<td>4.93</td>
<td>0.18</td>
<td>1.12</td>
<td>1.98</td>
<td>-0.24</td>
<td>-0.18</td>
<td>-0.06</td>
</tr>
<tr>
<td>Liaw</td>
<td>0.059</td>
<td>0.182</td>
<td>0.254</td>
<td>0.09</td>
<td>0.058</td>
<td>0.103</td>
<td>-0.012</td>
<td>-0.009</td>
<td>-0.003</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>-0.00002</td>
<td>0.000000</td>
<td>0.000000</td>
<td>-0.00003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Party</td>
<td>-0.00003</td>
<td>0.000000</td>
<td>0.000003</td>
<td>-0.000005</td>
<td>-0.00003</td>
<td>-0.00002</td>
<td>-0.00007</td>
<td>-0.00006</td>
<td>-0.000006</td>
</tr>
<tr>
<td>AUC</td>
<td>-0.00003</td>
<td>0.000001</td>
<td>0.000003</td>
<td>-0.000006</td>
<td>-0.00003</td>
<td>-0.00002</td>
<td>-0.00007</td>
<td>-0.00007</td>
<td>-0.000007</td>
</tr>
<tr>
<td>mindepth 39</td>
<td>6.88</td>
<td>7.3</td>
<td>7.35</td>
<td>6.67</td>
<td>6.77</td>
<td>6.79</td>
<td>6.63</td>
<td>6.67</td>
<td>6.63</td>
</tr>
</tbody>
</table>
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

### Table A.6. Median of VIM medians and minimal depth medians for the uncorrelated variables under H0.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>r=0.80</th>
<th>r=0.40</th>
<th>r=0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>gini</td>
<td>1.51</td>
<td>1.63</td>
<td>1.84</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>-0.00008</td>
<td>-0.00009</td>
<td>-0.00009</td>
</tr>
<tr>
<td>Breiman</td>
<td>-0.25</td>
<td>-0.27</td>
<td>-0.29</td>
</tr>
<tr>
<td>Liaw</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.02</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>-0.00008</td>
<td>-0.00009</td>
<td>-0.00009</td>
</tr>
<tr>
<td>Party</td>
<td>-0.00007</td>
<td>-0.00009</td>
<td>-0.00009</td>
</tr>
<tr>
<td>AUC</td>
<td>-0.00007</td>
<td>-0.00008</td>
<td>-0.00008</td>
</tr>
<tr>
<td>mindepth39</td>
<td>6.6</td>
<td>6.49</td>
<td>6.32</td>
</tr>
<tr>
<td>mindepth27</td>
<td>6.6</td>
<td>6.48</td>
<td>6.27</td>
</tr>
</tbody>
</table>

### Table A.7. Median of VIM and minimal for V2 (associated variable) under H1. Strong single study.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>GINI</td>
<td>349.05</td>
<td>326.55</td>
<td>320.78</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>0.450</td>
<td>0.412</td>
<td>0.405</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>91.26</td>
<td>81.73</td>
<td>79.65</td>
</tr>
<tr>
<td>Liaw</td>
<td>1.56</td>
<td>1.54</td>
<td>1.53</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>0.0807</td>
<td>0.0010</td>
<td>0.0002</td>
</tr>
<tr>
<td>Party</td>
<td>0.961</td>
<td>0.870</td>
<td>0.853</td>
</tr>
<tr>
<td>AUC</td>
<td>0.961</td>
<td>0.871</td>
<td>0.854</td>
</tr>
<tr>
<td>mindepth39</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>mindepth27</td>
<td>1.25</td>
<td>1.24</td>
<td>1.24</td>
</tr>
</tbody>
</table>

### Table A.8. Median of VIM medians and minimal depth medians for the correlated variables under H1. Strong single study.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>GINI</td>
<td>58.74</td>
<td>11.16</td>
<td>4.61</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>0.0160</td>
<td>0.0024</td>
<td>0.0010</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>17.82</td>
<td>9.42</td>
<td>6.86</td>
</tr>
<tr>
<td>Liaw</td>
<td>0.81</td>
<td>0.47</td>
<td>0.35</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>0.00010</td>
<td>0.000002</td>
<td>0.0000004</td>
</tr>
<tr>
<td>Party</td>
<td>0.0507</td>
<td>0.0067</td>
<td>0.0024</td>
</tr>
<tr>
<td>AUC</td>
<td>0.0507</td>
<td>0.0067</td>
<td>0.0024</td>
</tr>
<tr>
<td>mindepth39</td>
<td>2.39</td>
<td>4.89</td>
<td>5.79</td>
</tr>
<tr>
<td>mindepth27</td>
<td>2.28</td>
<td>4.49</td>
<td>5.50</td>
</tr>
</tbody>
</table>
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

### Table A.9. Median of VIM medians and minimal depth medians for the uncorrelated variables under Hₐ. Strong single study.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5  20  40</td>
<td>5  20  40</td>
<td>5  20  40</td>
</tr>
<tr>
<td>GINI</td>
<td>0.35  0.26  0.24</td>
<td>0.74  0.63  0.60</td>
<td>0.60  0.84  0.84</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>0.0000  0.0000  0.0000</td>
<td>0.0000  0.0000  0.0000</td>
<td>0.0000  0.0000  0.0000</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>-0.05  -0.05  -0.05</td>
<td>-0.04  -0.05  -0.03</td>
<td>-0.03  -0.07  -0.07</td>
</tr>
<tr>
<td>Liaw</td>
<td>0.00  0.00  0.00</td>
<td>0.00  0.00  0.00</td>
<td>0.00  0.00  0.00</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>-0.00001  -0.00001 -0.000007</td>
<td>-0.000003 -0.000001 -0.000001</td>
<td>0  0  0</td>
</tr>
<tr>
<td>Party</td>
<td>-0.000009  -0.000004 -0.000003</td>
<td>-0.00002 -0.00001 -0.00001</td>
<td>-0.00001 -0.000003 -0.000003</td>
</tr>
<tr>
<td>AUC</td>
<td>-0.000009  -0.000004 -0.000003</td>
<td>-0.000003 -0.000001 -0.000001</td>
<td>-0.000001 -0.000003 -0.000003</td>
</tr>
<tr>
<td>mindepth39</td>
<td>6.70  6.74  6.73</td>
<td>6.66  6.68  6.68</td>
<td>6.68  6.70  6.70</td>
</tr>
<tr>
<td>mindepth27</td>
<td>6.68  6.72  6.73</td>
<td>6.64  6.65  6.65</td>
<td>6.65  6.67  6.67</td>
</tr>
</tbody>
</table>

### Table A.10. Median of VIM and minimal depth for V₂ (associated variable) under Hₐ. Weak single study.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5  20  40</td>
<td>5  20  40</td>
<td>5  20  40</td>
</tr>
<tr>
<td>GINI</td>
<td>1.92  1.41  1.25</td>
<td>2.78  2.24  2.12</td>
<td>2.81  2.91  2.89</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>0.0022  0.0034  0.0037</td>
<td>0.0025  0.0022  0.0026</td>
<td>0.0018  0.0023  0.0025</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>3.75  5.63  6.06</td>
<td>3.55  3.70  4.42</td>
<td>2.97  3.29  3.45</td>
</tr>
<tr>
<td>Liaw</td>
<td>0.19  0.29  0.31</td>
<td>0.18  0.19  0.23</td>
<td>0.14  0.17  0.18</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>0.0001  -0.000006 -0.000002</td>
<td>0.0008  0  0</td>
<td>0  0  0</td>
</tr>
<tr>
<td>Party</td>
<td>0.0009  0.0007  0.0004</td>
<td>0.0018  0.0012  0.0011</td>
<td>0.0017  0.0019  0.0021</td>
</tr>
<tr>
<td>AUC</td>
<td>0.0010  0.0007  0.0004</td>
<td>0.0019  0.0012  0.0011</td>
<td>0.0017  0.0020  0.0020</td>
</tr>
<tr>
<td>mindepth39</td>
<td>5.92  6.69  6.97</td>
<td>5.05  5.61  5.83</td>
<td>5.08  4.94  4.97</td>
</tr>
<tr>
<td>mindepth27</td>
<td>5.92  6.62  6.90</td>
<td>5.17  5.72  5.82</td>
<td>5.19  5.06  5.14</td>
</tr>
</tbody>
</table>

### Table A.11. Median of VIM medians and minimal depth medians for the correlated non-associated variables under Hₐ. Weak single study.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5  20  40</td>
<td>5  20  40</td>
<td>5  20  40</td>
</tr>
<tr>
<td>GINI</td>
<td>1.53  1.15  1.12</td>
<td>1.52  1.40  1.39</td>
<td>1.50  1.49  1.49</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>0.0014  0.0021  0.0027</td>
<td>0.0002  0.0005  0.0007</td>
<td>-0.00004 -0.00004 -0.00004</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>3.05  4.55  5.41</td>
<td>0.65  1.35  2.05</td>
<td>-0.13 -0.13 -0.13</td>
</tr>
<tr>
<td>Liaw</td>
<td>0.16  0.23  0.28</td>
<td>0.03  0.07  0.11</td>
<td>-0.01 -0.01 -0.01</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>-0.0003  0.00  0.00</td>
<td>0.00  0.00  0.00</td>
<td>0.00  0.00  0.00</td>
</tr>
<tr>
<td>Party</td>
<td>0.0002  0.0001  0.00004</td>
<td>-0.00003 -0.00002 -0.00002</td>
<td>-0.00007 -0.00006 -0.00006</td>
</tr>
<tr>
<td>AUC</td>
<td>0.0002  0.0001  0.00005</td>
<td>-0.00003 -0.00002 -0.00001</td>
<td>-0.00001 -0.00001 -0.00001</td>
</tr>
<tr>
<td>mindepth39</td>
<td>6.47  7.08  7.20</td>
<td>6.56  6.72  6.75</td>
<td>6.60  6.60  6.60</td>
</tr>
<tr>
<td>mindepth27</td>
<td>6.42  7.00  7.11</td>
<td>6.56  6.70  6.73</td>
<td>6.61  6.61  6.61</td>
</tr>
</tbody>
</table>
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Table A.12. Median of VIM medians and minimal depth medians for the uncorrelated variables under Hₐ. Weak single study.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5 20 40</td>
<td>5 20 40</td>
<td>5 20 40</td>
</tr>
<tr>
<td>GINI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rawpermRF</td>
<td>1.51 1.61 1.78 1.50 1.53 1.59 1.50 1.50 1.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BREIMAN</td>
<td>-0.24 -0.26 -0.27 -0.22 -0.22 -0.22 -0.22 -0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liaw</td>
<td>-0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rawpermCF</td>
<td>-0.00008 -0.000088 -0.0000919 -0.00007</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>Party</td>
<td>-0.0001 -0.0001 -0.00009 -0.00007 -0.00007 -0.00007 -0.00007 -0.00007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUC</td>
<td>-0.0001 -0.0001 -0.00009 -0.00007 -0.00007 -0.00007 -0.00007 -0.00007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mindepth27</td>
<td>6.59 6.48 6.33 6.60 6.57 6.52 6.61 6.60 6.59</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.13. Median of VIM and minimal depth for V₂ (associated correlated variable) under Hₐ. Strong interaction study.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5 20 40</td>
<td>5 20 40</td>
<td>5 20 40</td>
</tr>
<tr>
<td>GINI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rawpermRF</td>
<td>48393.30 43774.15 42254.90 65046.25 61749.85 60544.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BREIMAN</td>
<td>24.41 21.46 20.97 33.02 31.23 30.75 36.59 36.16 35.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liaw</td>
<td>7265.62 6453.20 6170.27 12448.70 11371.75 11066.90 15459.00 14945.30 14682.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rawpermCF</td>
<td>151.26 148.13 146.80 159.90 158.93 158.62 161.62 161.39 161.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Party</td>
<td>4.32 0.03 0.002 12.67 0.01 0 0 0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUC</td>
<td>1.65 1.49 1.45 2.33 2.21 2.17 2.51 2.55 2.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mindepth39</td>
<td>1.52 1.55 1.57 1.49 1.47 1.48 1.52 1.51 1.51</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.14. Median of VIM and minimal for V₀ (associated uncorrelated variable) under Hₐ. Strong interaction study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

### Table A.15. Median of VIM medians and minimal depth medians for the correlated variables under Hₐ, Strong interaction study.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>( r = 0.80 )</th>
<th>( r = 0.40 )</th>
<th>( r = 0.10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>GINI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rawpermRF</td>
<td>1.33</td>
<td>0.28</td>
<td>0.16</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>1300.48</td>
<td>608.60</td>
<td>466.39</td>
</tr>
<tr>
<td>Liaw</td>
<td>62.63</td>
<td>31.17</td>
<td>24.11</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>0.0113</td>
<td>-0.00002</td>
<td>0</td>
</tr>
<tr>
<td>Party</td>
<td>0.121</td>
<td>0.019</td>
<td>0.0074</td>
</tr>
<tr>
<td>AUC</td>
<td>0.120</td>
<td>0.019</td>
<td>0.0073</td>
</tr>
<tr>
<td>mindepth39</td>
<td>3.49</td>
<td>5.62</td>
<td>6.32</td>
</tr>
<tr>
<td>mindepth27</td>
<td>3.31</td>
<td>5.42</td>
<td>6.19</td>
</tr>
</tbody>
</table>

### Table A.16. Median of VIM medians and minimal depth medians for the uncorrelated variables under Hₐ, Strong interaction study.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>( r = 0.80 )</th>
<th>( r = 0.40 )</th>
<th>( r = 0.10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>GINI</td>
<td>305.09</td>
<td>287.81</td>
<td>295.00</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>-0.0009</td>
<td>-0.0011</td>
<td>-0.0009</td>
</tr>
<tr>
<td>Liaw</td>
<td>-0.49</td>
<td>-0.59</td>
<td>-0.50</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>-0.00091</td>
<td>-0.00091</td>
<td>-0.00086</td>
</tr>
<tr>
<td>Party</td>
<td>-0.00011</td>
<td>-0.00011</td>
<td>-0.00012</td>
</tr>
<tr>
<td>AUC</td>
<td>-0.00011</td>
<td>-0.00011</td>
<td>-0.00012</td>
</tr>
</tbody>
</table>

### Table A.17. Median of VIM and minimal for \( V_2 \) (associated correlated variable) under Hₐ, Weak interaction study.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>( r = 0.80 )</th>
<th>( r = 0.40 )</th>
<th>( r = 0.10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>GINI</td>
<td>184.08</td>
<td>136.29</td>
<td>126.10</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>0.16</td>
<td>0.22</td>
<td>0.29</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>333.49</td>
<td>467.16</td>
<td>565.88</td>
</tr>
<tr>
<td>Liaw</td>
<td>17.29</td>
<td>24.09</td>
<td>29.05</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>0.01290</td>
<td>0.00008</td>
<td>0.00004</td>
</tr>
<tr>
<td>Party</td>
<td>0.00088</td>
<td>0.00054</td>
<td>0.00037</td>
</tr>
<tr>
<td>AUC</td>
<td>0.00086</td>
<td>0.00052</td>
<td>0.00037</td>
</tr>
<tr>
<td>mindepth39</td>
<td>6.31</td>
<td>6.98</td>
<td>7.18</td>
</tr>
<tr>
<td>mindepth27</td>
<td>6.27</td>
<td>6.90</td>
<td>7.07</td>
</tr>
</tbody>
</table>
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

### Table A.18. Median of VIM and minimal for $V_m$ (associated uncorrelated variable) under $H_A$.
Weak interaction study.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>Columna1 $r = 0.80$</th>
<th>Columna2</th>
<th>Columna3 $r = 0.40$</th>
<th>Columna4</th>
<th>Columna5 $r = 0.10$</th>
<th>Columna6</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>GINI</td>
<td>259.06</td>
<td>298.01</td>
<td>344.61</td>
<td>247.21</td>
<td>254.67</td>
<td>270.93</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>0.20</td>
<td>0.28</td>
<td>0.39</td>
<td>0.12</td>
<td>0.17</td>
<td>0.19</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>353.45</td>
<td>472.18</td>
<td>584.85</td>
<td>248.45</td>
<td>322.53</td>
<td>353.12</td>
</tr>
<tr>
<td>Liaw</td>
<td>18.31</td>
<td>24.36</td>
<td>29.97</td>
<td>12.92</td>
<td>16.73</td>
<td>18.27</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>0.19</td>
<td>0.28</td>
<td>0.38</td>
<td>0.13</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td>Party</td>
<td>0.00193</td>
<td>0.00261</td>
<td>0.00372</td>
<td>0.00145</td>
<td>0.00158</td>
<td>0.00194</td>
</tr>
<tr>
<td>AUC</td>
<td>0.00200</td>
<td>0.00262</td>
<td>0.00360</td>
<td>0.00138</td>
<td>0.00164</td>
<td>0.00189</td>
</tr>
<tr>
<td>mindepth39</td>
<td>5.58</td>
<td>5.17</td>
<td>4.73</td>
<td>5.69</td>
<td>5.66</td>
<td>5.43</td>
</tr>
<tr>
<td>mindepth27</td>
<td>5.65</td>
<td>5.28</td>
<td>4.91</td>
<td>5.81</td>
<td>5.76</td>
<td>5.54</td>
</tr>
</tbody>
</table>

Table A.19. Median of VIM medians and minimal depth medians for the correlated variables under $H_A$.
Weak interaction study.

<table>
<thead>
<tr>
<th>VIM median</th>
<th>Columna1 $r = 0.80$</th>
<th>Columna2</th>
<th>Columna3 $r = 0.40$</th>
<th>Columna4</th>
<th>Columna5 $r = 0.10$</th>
<th>Columna6</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>GINI</td>
<td>148.69</td>
<td>115.08</td>
<td>112.63</td>
<td>154.72</td>
<td>144.28</td>
<td>147.22</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>0.0918</td>
<td>0.1524</td>
<td>0.2246</td>
<td>0.0123</td>
<td>0.0391</td>
<td>0.0649</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>227.55</td>
<td>388.21</td>
<td>510.53</td>
<td>37.17</td>
<td>118.27</td>
<td>190.91</td>
</tr>
<tr>
<td>Liaw</td>
<td>11.85</td>
<td>20.10</td>
<td>26.23</td>
<td>1.94</td>
<td>6.14</td>
<td>9.93</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>-0.00042</td>
<td>-0.00030</td>
<td>-0.0002</td>
<td>-0.0036</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Party</td>
<td>0.00016</td>
<td>0.00007</td>
<td>0.00006</td>
<td>-0.00002</td>
<td>-0.00002</td>
<td>-0.00001</td>
</tr>
<tr>
<td>AUC</td>
<td>0.00016</td>
<td>0.00007</td>
<td>0.00006</td>
<td>-0.00002</td>
<td>-0.00002</td>
<td>-0.00001</td>
</tr>
<tr>
<td>mindepth39</td>
<td>5.65</td>
<td>5.17</td>
<td>4.73</td>
<td>5.69</td>
<td>5.66</td>
<td>5.43</td>
</tr>
<tr>
<td>mindepth27</td>
<td>6.60</td>
<td>7.20</td>
<td>7.34</td>
<td>6.60</td>
<td>6.75</td>
<td>6.79</td>
</tr>
</tbody>
</table>

Table A.20. Median of VIM medians and minimal depth medians for the uncorrelated variables under $H_A$.
Weak interaction study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

**Table A.21. Power of VIMs and minimal depth detecting $V_2$, permuting the outcome. Weak single study (WAC).**

<table>
<thead>
<tr>
<th></th>
<th>r = 0.80</th>
<th></th>
<th>r = 0.40</th>
<th></th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>GINI</td>
<td>34.29</td>
<td>9.97</td>
<td>4.73</td>
<td>67.06</td>
<td>48.77</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>76.06</td>
<td>75.12</td>
<td>52.07</td>
<td>74.17</td>
<td>66.86</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>72.61</td>
<td>68.08</td>
<td>40.48</td>
<td>66.04</td>
<td>61.29</td>
</tr>
<tr>
<td>Liaw</td>
<td>72.60</td>
<td>68.06</td>
<td>40.48</td>
<td>66.03</td>
<td>61.30</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>13.34</td>
<td>0.00</td>
<td>0</td>
<td>45.01</td>
<td>0.0028</td>
</tr>
<tr>
<td>Party</td>
<td>65.45</td>
<td>59.80</td>
<td>45.58</td>
<td>80.65</td>
<td>68.57</td>
</tr>
<tr>
<td>AUC</td>
<td>65.58</td>
<td>60.90</td>
<td>45.55</td>
<td>80.92</td>
<td>70.13</td>
</tr>
<tr>
<td>mindepth39</td>
<td>49.93</td>
<td>13.15</td>
<td>4.44</td>
<td>79.16</td>
<td>60.22</td>
</tr>
<tr>
<td>mindepth27</td>
<td>51.78</td>
<td>14.26</td>
<td>5.08</td>
<td>79.26</td>
<td>61.22</td>
</tr>
</tbody>
</table>

**Table A.22. Power of VIMs and minimal depth detecting $V_2$, permuting the outcome. Strong single study (SAC).**

<table>
<thead>
<tr>
<th></th>
<th>r = 0.80</th>
<th></th>
<th>r = 0.40</th>
<th></th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>GINI</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Liaw</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>100</td>
<td>55.3404</td>
<td>0.0172</td>
<td>100</td>
<td>32.7868</td>
</tr>
<tr>
<td>Party</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>AUC</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>mindepth39</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>mindepth27</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>V2 WAC</th>
<th>$r = 0.80$</th>
<th>$r = 0.40$</th>
<th>$r = 0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5 20 40 5 20 40 5 20 40</td>
<td>5 20 40 5 20 40 5 20 40</td>
<td>5 20 40 5 20 40</td>
</tr>
<tr>
<td>GINI</td>
<td>23.81 3.83 0.95 45.80 35.75 32.96 53.32 54.74 55.42</td>
<td>64.65 56.97 36.38 58.34 58.92 53.37 56.58 59.46 60.03</td>
<td></td>
</tr>
<tr>
<td>rawpermRF</td>
<td>66.27 53.50 30.43 55.59 56.63 48.15 53.74 56.25 56.74</td>
<td>66.26 53.51 30.41 55.59 56.62 48.15 53.73 56.25 56.74</td>
<td></td>
</tr>
<tr>
<td>BREIMAN</td>
<td>8.65 0.02 0.00 31.18 0.18 0 0 0</td>
<td>66.13 52.23 41.78 76.10 73.56 69.79 75.65 76.88 76.66</td>
<td></td>
</tr>
<tr>
<td>Liaw</td>
<td>64.63 51.53 42.08 74.75 72.14 70.11 75.59 77.89 76.45</td>
<td>27.94 4.91 0.46 51.25 42.42 36.37 59.15 58.40 57.34</td>
<td></td>
</tr>
<tr>
<td>rawpermCF</td>
<td>66.27 53.50 30.43 55.59 56.63 48.15 53.74 56.25 56.74</td>
<td>5.95 4.63 0.51 49.96 42.08 36.80 56.83 58.27 56.89</td>
<td></td>
</tr>
<tr>
<td>Party</td>
<td>64.63 51.53 42.08 74.75 72.14 70.11 75.59 77.89 76.45</td>
<td>66.27 53.50 30.43 55.59 56.63 48.15 53.74 56.25 56.74</td>
<td></td>
</tr>
<tr>
<td>AUC</td>
<td>63.77 52.55 41.16 53.95 55.05 41.57 57.87 54.58 59.03</td>
<td>66.27 53.50 30.43 55.59 56.63 48.15 53.74 56.25 56.74</td>
<td></td>
</tr>
<tr>
<td>mindepth39</td>
<td>66.97 81.48 89.49 59.67 68.21 77.74 0 0 0</td>
<td>82.01 91.49 94.18 77.79 84.97 84.67 77.39 76.04 78.37</td>
<td></td>
</tr>
<tr>
<td>mindepth27</td>
<td>65.83 79.99 89.23 58.60 65.83 74.26 60.04 57.69 63.56</td>
<td>67.02 79.84 88.25 60.42 66.64 74.98 61.28 58.59 65.61</td>
<td></td>
</tr>
</tbody>
</table>

Table A.23. Power of VIMs and minimal depth detecting $V_2$, permuting the outcome. Weak interaction study (WAC).

<table>
<thead>
<tr>
<th>V90 WAC</th>
<th>$r = 0.80$</th>
<th>$r = 0.40$</th>
<th>$r = 0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5 20 40 5 20 40 5 20 40</td>
<td>5 20 40 5 20 40 5 20 40</td>
<td></td>
</tr>
<tr>
<td>GINI</td>
<td>61.35 72.15 83.05 55.30 59.98 67.32 55.98 55.01 61.40</td>
<td>65.76 62.82 51.08 57.66 59.94 52.41 61.67 57.03 63.14</td>
<td></td>
</tr>
<tr>
<td>rawpermRF</td>
<td>63.77 52.55 41.16 53.95 55.05 41.57 57.87 54.58 59.03</td>
<td>63.78 52.58 41.20 53.96 55.05 41.59 57.87 54.58 59.04</td>
<td></td>
</tr>
<tr>
<td>BREIMAN</td>
<td>66.97 81.48 89.49 59.67 68.21 77.74 0 0 0</td>
<td>82.01 91.49 94.18 77.79 84.97 84.67 77.39 76.04 78.37</td>
<td></td>
</tr>
<tr>
<td>Liaw</td>
<td>63.78 52.58 41.20 53.96 55.05 41.59 57.87 54.58 59.04</td>
<td>82.10 91.47 94.40 78.61 84.86 86.61 76.97 75.39 79.18</td>
<td></td>
</tr>
<tr>
<td>rawpermCF</td>
<td>67.02 79.84 88.25 60.42 66.64 74.98 61.28 58.59 65.61</td>
<td>65.83 79.99 89.23 58.60 65.83 74.26 60.04 57.69 63.56</td>
<td></td>
</tr>
<tr>
<td>Party</td>
<td>82.01 91.49 94.18 77.79 84.97 84.67 77.39 76.04 78.37</td>
<td>67.02 79.84 88.25 60.42 66.64 74.98 61.28 58.59 65.61</td>
<td></td>
</tr>
<tr>
<td>AUC</td>
<td>82.10 91.47 94.40 78.61 84.86 86.61 76.97 75.39 79.18</td>
<td>65.83 79.99 89.23 58.60 65.83 74.26 60.04 57.69 63.56</td>
<td></td>
</tr>
</tbody>
</table>

Table A.24. Power of VIMs and minimal depth detecting $V_{90}$, permuting the outcome. Weak interaction study (WAC).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>V2 SAC</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>GINI</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Liaw</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>100</td>
<td>2.81</td>
<td>0.44</td>
</tr>
<tr>
<td>Party</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>AUC</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>mindepth39</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>mindepth27</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table A.25. Power of VIMs and minimal depth detecting V2, permuting the outcome. Strong interaction study (SAC).

<table>
<thead>
<tr>
<th>V90 SAC</th>
<th>r = 0.80</th>
<th>r = 0.40</th>
<th>r = 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>GINI</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>rawpermRF</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>BREIMAN</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Liaw</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>rawpermCF</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Party</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>AUC</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>mindepth39</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>mindepth27</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table A.26. Power of VIMs and minimal depth detecting V90, permuting the outcome. Strong interaction study (SAC).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Table A.27. Median of the depth threshold under each correlation condition when applying minimal depth under $H_0$. For both mtry values the median was the same.

<table>
<thead>
<tr>
<th>Median Depth threshold</th>
<th>$r=0.80$</th>
<th>$r=0.40$</th>
<th>$r=0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
</tbody>
</table>

Table A.28. Median of the depth threshold under each correlation condition when applying minimal depth under $H_A$ for each association study. SAC refers to strongly-associated studies, and WAC to weakly-associated studies. For both mtry values the median was the same.

<table>
<thead>
<tr>
<th>Median Depth threshold</th>
<th>$r=0.80$</th>
<th>$r=0.40$</th>
<th>$r=0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>5</td>
<td>20</td>
<td>40</td>
</tr>
</tbody>
</table>

Figure A.1. RF VIMs, minimal depth, VIM\textsubscript{AUC} and VIM\textsubscript{party} under $H_0$ for $V_2$, for two variable correlated $V_3$ and $V_6$, and for two independent variables $V_{42}$ and $V_{90}$ when $r=0.10$ and $N=20$. 
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.2. RF VIMs, minimal depth, VIM\textsubscript{AUC} and VIM\textsubscript{party} under H\textsubscript{0} for V\textsubscript{2}, for two variable correlated V\textsubscript{3} and V\textsubscript{6}, and for two independent variables V\textsubscript{42} and V\textsubscript{90} when r = 0.10 and N = 40.

Figure A.3. RF VIMs, minimal depth, VIM\textsubscript{AUC} and VIM\textsubscript{party} under H\textsubscript{0} for V\textsubscript{2}, for two variable correlated V\textsubscript{3} and V\textsubscript{6}, and for two independent variables V\textsubscript{42} and V\textsubscript{90} when r = 0.40 and N = 5.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.4. RF VIMs, minimal depth, VIM$_{\text{AUC}}$ and VIM$_{\text{party}}$ under H$_0$ for $V_2$, for two variable correlated $V_3$ and $V_6$, and for two independent variables $V_{42}$ and $V_{90}$ when $r = 0.40$ and $N = 40$.

Figure A.5. RF VIMs, minimal depth, VIM$_{\text{AUC}}$ and VIM$_{\text{party}}$ under H$_0$ for $V_2$, for two variable correlated $V_3$ and $V_6$, and for two independent variables $V_{42}$ and $V_{90}$ when $r = 0.80$ and $N = 5$. 
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.6. RF VIMs, minimal depth, VIM\textsubscript{AUC} and VIM\textsubscript{party} under H\textsubscript{0} for V\textsubscript{2}, for two variable correlated V\textsubscript{3} and V\textsubscript{6}, and for two independent variables V\textsubscript{42} and V\textsubscript{90} when r = 0.80 and N = 20.

Figure A.7. RF VIMs, minimal depth, VIM\textsubscript{AUC} and VIM\textsubscript{party} under H\textsubscript{A} for V\textsubscript{2}, for two variable correlated V\textsubscript{3} and V\textsubscript{6}, and for two independent variables V\textsubscript{42} and V\textsubscript{90} when r = 0.10 and N = 5. Strong single study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.8. RF VIMs, minimal depth, $\text{VIM}_\text{AUC}$ and $\text{VIM}_\text{party}$ under $H_A$ for $V_2$, for two variable correlated $V_3$ and $V_6$, and for two independent variables $V_{42}$ and $V_{90}$ when $r = 0.10$ and $N = 20$. Strong single study.

Figure A.9. RF VIMs, minimal depth, $\text{VIM}_\text{AUC}$ and $\text{VIM}_\text{party}$ under $H_A$ for $V_2$, for two variable correlated $V_3$ and $V_6$, and for two independent variables $V_{42}$ and $V_{90}$ when $r = 0.10$ and $N = 40$. Strong single study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.10. RF VIMs, minimal depth, VIM\textsubscript{AUC} and VIM\textsubscript{party} under H\textsubscript{A} for V\textsubscript{2}, for two variable correlated V\textsubscript{3} and V\textsubscript{6}, and for two independent variables V\textsubscript{42} and V\textsubscript{90} when r = 0.40 and N = 5. Strong single study.

Figure A.11. RF VIMs, minimal depth, VIM\textsubscript{AUC} and VIM\textsubscript{party} under H\textsubscript{A} for V\textsubscript{2}, for two variable correlated V\textsubscript{3} and V\textsubscript{6}, and for two independent variables V\textsubscript{42} and V\textsubscript{90} when r = 0.40 and N = 20. Strong single study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.12. RF VIMs, minimal depth, VIM\_AUC and VIM\_party under H\_A for V\_2, for two variable correlated V\_3 and V\_6, and for two independent variables V\_42 and V\_90 when r = 0.40 and N = 40. Strong single study.

Figure A.13. RF VIMs, minimal depth, VIM\_AUC and VIM\_party under H\_A for V\_2, for two variable correlated V\_3 and V\_6, and for two independent variables V\_42 and V\_90 when r = 0.80 and N = 5. Strong single study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.14. RF VIMs, minimal depth, VIM\textsubscript{AUC} and VIM\textsubscript{party} under $H_A$ for $V_2$, for two variable correlated $V_3$ and $V_6$, and for two independent variables $V_{42}$ and $V_{90}$ when $r = 0.80$ and $N = 20$. Strong single study.

Figure A.15. RF VIMs, minimal depth, VIM\textsubscript{AUC} and VIM\textsubscript{party} under $H_A$ for $V_2$, for two variable correlated $V_3$ and $V_6$, and for two independent variables $V_{42}$ and $V_{90}$ when $r = 0.10$ and $N = 20$. Weak single study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.16. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under $H_A$ for $V_3$, for two variable correlated $V_3$ and $V_6$, and for two independent variables $V_{42}$ and $V_{90}$ when $r = 0.10$ and $N = 40$. Weak single study.

Figure A.17. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under $H_A$ for $V_2$, for two variable correlated $V_3$ and $V_6$, and for two independent variables $V_{42}$ and $V_{90}$ when $r = 0.40$ and $N = 5$. Weak single study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.18. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_42 and V_90 when r = 0.40 and N = 40. Weak single study.

Figure A.19. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_42 and V_90 when r = 0.80 and N = 5. Weak single study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.20. RF VIMs, minimal depth, VIM\textsubscript{Gini} and VIM\textsubscript{party} under $H_A$ for $V_2$, for two variable correlated $V_3$ and $V_6$, and for two independent variables $V_{42}$ and $V_{90}$ when $r = 0.80$ and $N = 20$. Weak single study.

Figure A.21. RF VIMs, minimal depth, VIM\textsubscript{AUC} and VIM\textsubscript{party} under $H_A$ for $V_2$ and $V_{90}$, for two variables correlated $V_3$ and $V_6$, and for two independent ones $V_{42}$ when $r = 0.10$ and $N = 5$. Strong interaction study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.22. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.10 and N = 20. Strong interaction study.

Figure A.23. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.10 and N = 40. Strong interaction study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.24. RF VIMs, minimal depth, VIM$_{AUC}$ and VIM$_{party}$ under $H_A$ for $V_2$ and $V_{90}$, for two variables correlated $V_3$ and $V_6$, and for two independent ones $V_{42}$ when $r = 0.40$ and $N = 5$. Strong interaction study.

Figure A.25. RF VIMs, minimal depth, VIM$_{AUC}$ and VIM$_{party}$ under $H_A$ for $V_2$ and $V_{90}$, for two variables correlated $V_3$ and $V_6$, and for two independent ones $V_{42}$ when $r = 0.40$ and $N = 20$. Strong interaction study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.26. RF VIMs, minimal depth, VIM$_{AUC}$ and VIM$_{party}$ under $H_a$ for $V_2$ and $V_{90}$, for two variables correlated $V_3$ and $V_6$, and for two independent ones $V_{42}$ when $r = 0.40$ and $N = 40$. Strong interaction study.

Figure A.27. RF VIMs, minimal depth, VIM$_{AUC}$ and VIM$_{party}$ under $H_a$ for $V_2$ and $V_{90}$, for two variables correlated $V_3$ and $V_6$, and for two independent ones $V_{42}$ when $r = 0.80$ and $N = 5$. Strong interaction study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.28. RF VIMs, minimal depth, VIM\textsubscript{AUC} and VIM\textsubscript{party} under $H_A$ for $V_2$ and $V_{90}$, for two variables correlated $V_3$ and $V_6$, and for two independent ones $V_{42}$ when $r = 0.80$ and $N = 20$. Strong interaction study.

Figure A.29. RF VIMs, minimal depth, VIM\textsubscript{AUC} and VIM\textsubscript{party} under $H_A$ for $V_2$ and $V_{90}$, for two variables correlated $V_3$ and $V_6$, and for two independent ones $V_{42}$ when $r = 0.10$ and $N = 20$. Weak interaction study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.30. RF VIMs, minimal depth, VIM\text{AUC} and VIM\text{party} under H\text{A} for V_2 and V_90, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when \( r = 0.10 \) and \( N = 40 \). Weak interaction study.

Figure A.31. RF VIMs, minimal depth, VIM\text{AUC} and VIM\text{party} under H\text{A} for V_2 and V_90, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when \( r = 0.40 \) and \( N = 5 \). Weak interaction study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.32. RF VIMs, minimal depth, VIM_{\text{Gini}}, VIM_{\text{raperm-RF}} and VIM_{\text{Breiman}} under H_0 for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.40 and N = 40. Weak interaction study.

Figure A.33. RF VIMs, minimal depth, VIM_{\text{Gini}}, VIM_{\text{raperm-RF}} and VIM_{\text{Breiman}} under H_0 for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.80 and N = 5. Weak interaction study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Figure A.34. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.80 and N = 20. Weak interaction study.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code A.1. Function `rmvnormc` to generate multivariate normal predictors with correlation.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code A.2. Code of the weakly-associated single data generation, r=0.80 and N=40 as an example.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code A.3. Code of the weakly-associated interaction data generation, r=0.80 and N=40 as an example.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

```r
# Vector of the 100 mean, zero means
media<-rep(0,100)
# Matrix of the variance, identity matrix with 100x100 dimension
sigma<-diag(length(media))

# Number of variables correlated
H0=0
# Strength of correlation
correlation=0.80

# Correlation matrix
# First a matrix for the correlated variables
corr1<-matrix(0,nrow=H0,ncol=H0)
for(i in 1:H0){
  for(j in 1:H0){
    if(i==j){
      corr1[i,j]=1 # for the correlation of each variable with itself
    } else if(i>j){
      # Negative correlation of the 3rd variable with others
      corr1[i,j]=-correlation
    } else if(i<j){
      # The value of the correlation between the variable i and j is the correlation between j and i
      corr1[i,j]=corr1[j,i]
    } else if(i==3 & j==11){
      # The same between i, j and i and j
      corr1[i,j]=corr1[j,i]
    }
  }
}
# The rest of the correlation matrix has 0 values between different variables
correl1<-matrix(0,nrow=(100-H0),ncol=(100-H0))
correl1<-(diag(100-H0)
correl2<-diag(100-H0)
correl2<-(diag(100-H0)
correl3<-(diag(100-H0)
correl<-(diag(100-H0)

correl<-(diag(100-H0)

# Number of databases
n=100

# Loop to create the databases
for(i in 1:n){
  # Generate the 100 standard normal variables with 1000 observations
  x<-rnorm(1000,mean=media,sigma=sigma,cor=corr1)
  # Generate the error following a standard normal distribution
  e<-rnorm(1000,0.0,5)
  # The outcome is generated as the linear model
  y=x+e
  # Dataframe with the outcome and the predictors
  data<-data.frame(demand,y=x)
  # Save each database in each iteration
  write.table(data,quote="",row.names=FALSE,quote=FALSE)
}
```

Code A.4. Code of the data generation under $H_0, r=0.80$ and $N=40$ as an example.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code A.5. Code of the bias, coverage and p-values, $r=0.80$ and $N=40$ weakly-associated single study under $H_1$ as an example.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

```r
# number of variables correlated
N=40
# number of datasets
n=500
# Null vector for the median of the correlation
medic<-c(); medicnd<-c()
# loop for each database
for(i in 1:n) {
  data<-read.table(paste("data80_00strong","i",i,.R",header=T)) # input database
  x<-data[,1] # delete the first column as it is the outcome
  # create vectors of size the number of variables correlated
  mednum<-numeric(N); medicnum<-numeric(N)
  for(j in 1:N) { # loop for the number of variables correlated
    a<-x[,1:N] # take the matrix of correlated variables
    corr<-.cor(a) # estimate the correlation
    corri<-corr[,j] # for each variable take the correlation with others
    medic[j]<-median(cori) # take the median of the correlation with others
  }
  # save the median in the vector for each database
  medic<-append(medic,median(medic),after=1)
  forth in 1:(dim(x)[2]:N): # loop for the uncorrelated variables
    a<-x[,N+1:(dim(x)[2])] # take the matrix of uncorrelated variables
    corr<-cor(a) # estimate the correlation between them
    corri<-corr[,j] # for each variable take the correlation with others
    medicnd[j]<-median(cori) # take the median of the correlation with others
  }
  # take the median of the absolute value of the correlation with others
  medicnd<-append(medicnd,median(medicnd),after=1)
}
# median of the correlated
median(medic)
# median for the uncorrelated
median(medicnd)
```

Code A.6. Code for the correlation, r=0.80 and N=40 strong single study as an example.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code A.7. Code for VIMs and minimal depth under H0, 5% Cut-offs, and plot under H0 for all VIMs and minimal depth when \( r=0.80 \) and \( N=40 \) as an example.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code A.8. Code for the Power, VIMs and minimal depth under $H_A$, and plot under $H_A$ for all VIMs and minimal depth in the single association studies when $r=0.80$ and $N=40$ as an example.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Appendix B

Tables chapter 3

Table B.1. Median of VIM\textsubscript{Gini} for all ten variables in the four different studies under H\textsubscript{0}. Outcome continuous.

<table>
<thead>
<tr>
<th>H\textsubscript{0}, different studies</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xj<del>N(0,1), e</del>N(0,1)</td>
<td>62.9221</td>
<td>62.8311</td>
<td>62.7081</td>
<td>62.678</td>
<td>62.6089</td>
<td>62.3982</td>
<td>62.859</td>
<td>63.3346</td>
<td>62.8924</td>
<td>63.361</td>
</tr>
<tr>
<td>Different precision</td>
<td>52.4815</td>
<td>62.2177</td>
<td>64.0241</td>
<td>64.0603</td>
<td>64.2184</td>
<td>64.5318</td>
<td>64.6213</td>
<td>64.5983</td>
<td>64.3695</td>
<td>64.0536</td>
</tr>
<tr>
<td>e~N(0.0,5)</td>
<td>15.8618</td>
<td>15.611</td>
<td>15.5937</td>
<td>15.6527</td>
<td>15.8198</td>
<td>15.8144</td>
<td>15.7157</td>
<td>15.6832</td>
<td>15.6983</td>
<td>15.7261</td>
</tr>
</tbody>
</table>

Table B.2. Median of VIM\textsubscript{Gini} for all ten variables in the four different studies under H\textsubscript{0}. Outcome binary.

<table>
<thead>
<tr>
<th>HA, xj<del>N(0,1), e</del>N(0,1)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1 associated</td>
<td>63.1907</td>
<td>63.2225</td>
<td>63.237</td>
<td>63.411</td>
<td>63.731</td>
<td>63.3299</td>
<td>63.3741</td>
<td>63.5434</td>
<td>63.6112</td>
<td></td>
</tr>
<tr>
<td>X2 associated</td>
<td>62.5475</td>
<td>63.177</td>
<td>63.652</td>
<td>62.5191</td>
<td>62.1882</td>
<td>63.1568</td>
<td>63.3188</td>
<td>63.1333</td>
<td>63.0636</td>
<td></td>
</tr>
<tr>
<td>X3 associated</td>
<td>62.679</td>
<td>62.7183</td>
<td>63.499</td>
<td>63.1459</td>
<td>62.9821</td>
<td>63.2999</td>
<td>63.298</td>
<td>63.211</td>
<td>63.0888</td>
<td>63.0716</td>
</tr>
<tr>
<td>X4 associated</td>
<td>63.0417</td>
<td>62.527</td>
<td>62.5774</td>
<td>63.925</td>
<td>63.037</td>
<td>62.9427</td>
<td>62.7206</td>
<td>62.9106</td>
<td>63.3407</td>
<td>63.1398</td>
</tr>
<tr>
<td>X5 associated</td>
<td>60.273</td>
<td>62.8677</td>
<td>65.3761</td>
<td>62.9983</td>
<td>117.015</td>
<td>63.6788</td>
<td>63.2698</td>
<td>63.1303</td>
<td>63.2412</td>
<td>62.8546</td>
</tr>
<tr>
<td>X6 associated</td>
<td>62.8513</td>
<td>62.7945</td>
<td>62.564</td>
<td>63.3541</td>
<td>63.2972</td>
<td>117.479</td>
<td>63.9799</td>
<td>62.8141</td>
<td>63.0488</td>
<td>63.2025</td>
</tr>
<tr>
<td>X7 associated</td>
<td>62.2653</td>
<td>63.2159</td>
<td>63.1041</td>
<td>62.7904</td>
<td>63.3521</td>
<td>63.5549</td>
<td>116.711</td>
<td>62.8946</td>
<td>63.0605</td>
<td>62.9786</td>
</tr>
<tr>
<td>X8 associated</td>
<td>62.9558</td>
<td>62.3869</td>
<td>62.9407</td>
<td>62.7225</td>
<td>62.9905</td>
<td>63.088</td>
<td>62.5413</td>
<td>117.95</td>
<td>63.1133</td>
<td>62.8502</td>
</tr>
<tr>
<td>X10 associated</td>
<td>63.1169</td>
<td>62.5745</td>
<td>62.9598</td>
<td>62.7475</td>
<td>63.1775</td>
<td>63.1782</td>
<td>63.2238</td>
<td>63.4287</td>
<td>63.0448</td>
<td>117.561</td>
</tr>
</tbody>
</table>

Table B.3. Median of VIM\textsubscript{Gini} for all ten variables when all variables and the error are standard normal distributed under H\textsubscript{A} in the ten association studies. Outcome continuous.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>HA, ( s \sim \mathcal{N}(0,1) ), ( e \sim \mathcal{N}(0,1) )</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
<th>( X_5 )</th>
<th>( X_6 )</th>
<th>( X_7 )</th>
<th>( X_8 )</th>
<th>( X_9 )</th>
<th>( X_{10} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 ) associated</td>
<td>46.034</td>
<td>29.8585</td>
<td>29.9278</td>
<td>29.8821</td>
<td>29.7525</td>
<td>29.7971</td>
<td>29.932</td>
<td>29.8084</td>
<td>29.8824</td>
<td>29.8144</td>
</tr>
<tr>
<td>( X_2 ) associated</td>
<td>29.8273</td>
<td>45.857</td>
<td>29.822</td>
<td>29.855</td>
<td>29.8362</td>
<td>29.8717</td>
<td>29.9198</td>
<td>29.7442</td>
<td>29.639</td>
<td>29.9039</td>
</tr>
<tr>
<td>( X_3 ) associated</td>
<td>29.9533</td>
<td>29.7351</td>
<td>45.9971</td>
<td>29.8604</td>
<td>29.8681</td>
<td>29.9071</td>
<td>29.8305</td>
<td>29.735</td>
<td>29.899</td>
<td>29.7792</td>
</tr>
<tr>
<td>( X_4 ) associated</td>
<td>29.8286</td>
<td>29.9293</td>
<td>29.8451</td>
<td>46.0523</td>
<td>29.7872</td>
<td>29.8421</td>
<td>29.7835</td>
<td>29.8908</td>
<td>29.9046</td>
<td>29.8158</td>
</tr>
<tr>
<td>( X_5 ) associated</td>
<td>29.7303</td>
<td>29.7013</td>
<td>29.9555</td>
<td>29.8653</td>
<td>45.9434</td>
<td>29.8614</td>
<td>29.9575</td>
<td>29.8405</td>
<td>29.7226</td>
<td>30.0451</td>
</tr>
<tr>
<td>( X_7 ) associated</td>
<td>29.8547</td>
<td>29.7806</td>
<td>29.767</td>
<td>29.8967</td>
<td>29.7567</td>
<td>29.7129</td>
<td>46.2134</td>
<td>29.896</td>
<td>29.7968</td>
<td>29.8351</td>
</tr>
<tr>
<td>( X_8 ) associated</td>
<td>29.869</td>
<td>29.7679</td>
<td>29.8648</td>
<td>29.8724</td>
<td>29.8507</td>
<td>29.75</td>
<td>29.7657</td>
<td>46.4996</td>
<td>29.6813</td>
<td>29.9282</td>
</tr>
<tr>
<td>( X_9 ) associated</td>
<td>29.8809</td>
<td>29.7786</td>
<td>29.8447</td>
<td>29.8601</td>
<td>29.824</td>
<td>29.764</td>
<td>29.7746</td>
<td>46.2893</td>
<td>29.6852</td>
<td></td>
</tr>
<tr>
<td>( X_{10} ) associated</td>
<td>29.8222</td>
<td>29.7605</td>
<td>29.7045</td>
<td>29.8975</td>
<td>29.848</td>
<td>29.7855</td>
<td>29.8227</td>
<td>29.849</td>
<td>29.8035</td>
<td>46.3124</td>
</tr>
</tbody>
</table>

Table B.4. Median of VIM\(_{\text{Gini}}\) for all ten variables when all variables and the error are standard normal distributed under H\(_A\) in the ten association studies. Outcome binary.

<table>
<thead>
<tr>
<th>HA, Different precision</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
<th>( X_5 )</th>
<th>( X_6 )</th>
<th>( X_7 )</th>
<th>( X_8 )</th>
<th>( X_9 )</th>
<th>( X_{10} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 ) associated</td>
<td>100.339</td>
<td>63.3089</td>
<td>65.3545</td>
<td>65.3852</td>
<td>65.2904</td>
<td>65.5001</td>
<td>65.2534</td>
<td>65.0637</td>
<td>65.605</td>
<td>65.0978</td>
</tr>
<tr>
<td>( X_2 ) associated</td>
<td>53.3511</td>
<td>116.49</td>
<td>64.3514</td>
<td>64.662</td>
<td>64.1794</td>
<td>64.5586</td>
<td>64.5335</td>
<td>64.5422</td>
<td>65.089</td>
<td>64.328</td>
</tr>
<tr>
<td>( X_3 ) associated</td>
<td>53.2001</td>
<td>62.6837</td>
<td>119.033</td>
<td>64.445</td>
<td>64.213</td>
<td>64.3208</td>
<td>64.4411</td>
<td>64.4951</td>
<td>64.8469</td>
<td>64.0126</td>
</tr>
<tr>
<td>( X_4 ) associated</td>
<td>53.3683</td>
<td>62.9945</td>
<td>63.8045</td>
<td>119.937</td>
<td>63.9796</td>
<td>64.7709</td>
<td>64.3922</td>
<td>64.0794</td>
<td>64.2445</td>
<td>64.2556</td>
</tr>
<tr>
<td>( X_5 ) associated</td>
<td>53.4403</td>
<td>62.7712</td>
<td>64.0174</td>
<td>64.1662</td>
<td>120.141</td>
<td>64.7326</td>
<td>64.5422</td>
<td>64.2086</td>
<td>64.427</td>
<td>64.0494</td>
</tr>
<tr>
<td>( X_6 ) associated</td>
<td>53.1937</td>
<td>62.5086</td>
<td>63.8525</td>
<td>64.3627</td>
<td>64.0856</td>
<td>64.3952</td>
<td>63.9776</td>
<td>64.448</td>
<td>64.2555</td>
<td>64.2571</td>
</tr>
<tr>
<td>( X_7 ) associated</td>
<td>53.3783</td>
<td>62.5451</td>
<td>63.7194</td>
<td>64.1364</td>
<td>64.4073</td>
<td>64.7211</td>
<td>120.262</td>
<td>64.4709</td>
<td>64.3918</td>
<td>64.0277</td>
</tr>
<tr>
<td>( X_8 ) associated</td>
<td>52.7411</td>
<td>62.6389</td>
<td>63.9117</td>
<td>64.3752</td>
<td>64.3256</td>
<td>64.3619</td>
<td>64.1936</td>
<td>121.116</td>
<td>64.2059</td>
<td>64.3668</td>
</tr>
<tr>
<td>( X_9 ) associated</td>
<td>53.0516</td>
<td>62.5846</td>
<td>63.7718</td>
<td>64.4168</td>
<td>64.2923</td>
<td>64.6197</td>
<td>64.3321</td>
<td>64.366</td>
<td>119.014</td>
<td>64.1118</td>
</tr>
<tr>
<td>( X_{10} ) associated</td>
<td>53.3321</td>
<td>62.4944</td>
<td>63.736</td>
<td>64.0733</td>
<td>64.5771</td>
<td>64.7573</td>
<td>64.8163</td>
<td>64.4405</td>
<td>64.3617</td>
<td>120.48</td>
</tr>
</tbody>
</table>

Table B.5. Median of VIM\(_{\text{Gini}}\) when all variables follow \( \mathcal{N}(0,1) \) but each one with different precision, under H\(_A\) in the ten association studies. Each variable \( X_i \) has \( i \) number of decimal places. Outcome continuous.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>HA, Different precision</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>X10 associated</td>
<td>24.9239</td>
<td>29.737</td>
<td>30.4726</td>
<td>30.5497</td>
<td>30.541</td>
<td>30.4284</td>
<td>30.6795</td>
<td>30.4513</td>
<td>30.4251</td>
<td>46.5528</td>
</tr>
</tbody>
</table>

Table B.6. Median of VIM\textsubscript{Gini} when all variables follow N(0,1) but each one with different precision, under H\textsubscript{A} in the ten association studies. Each variable X\textsubscript{i} has i number of decimal places. Outcome binary.

<table>
<thead>
<tr>
<th>HA, Different variance</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>X2 associated</td>
<td>104.939</td>
<td>2228.02</td>
<td>104.579</td>
<td>104.744</td>
<td>104.589</td>
<td>105.195</td>
<td>104.383</td>
<td>105.144</td>
<td>104.703</td>
<td>105.62</td>
</tr>
<tr>
<td>X4 associated</td>
<td>94.1938</td>
<td>94.0224</td>
<td>93.4793</td>
<td>1760.41</td>
<td>94.5221</td>
<td>94.0474</td>
<td>93.4961</td>
<td>94.147</td>
<td>93.9245</td>
<td>93.341</td>
</tr>
<tr>
<td>X5 associated</td>
<td>89.746</td>
<td>88.3718</td>
<td>89.314</td>
<td>89.0518</td>
<td>1535.15</td>
<td>88.5215</td>
<td>89.2763</td>
<td>88.9912</td>
<td>89.8794</td>
<td>88.5399</td>
</tr>
<tr>
<td>X6 associated</td>
<td>84.8603</td>
<td>84.8254</td>
<td>85.0341</td>
<td>84.428</td>
<td>85.2924</td>
<td>1281.3</td>
<td>85.607</td>
<td>85.2379</td>
<td>85.6085</td>
<td>85.295</td>
</tr>
<tr>
<td>X7 associated</td>
<td>79.706</td>
<td>79.2596</td>
<td>79.2412</td>
<td>80.1114</td>
<td>80.1693</td>
<td>79.5691</td>
<td>1035.64</td>
<td>79.7442</td>
<td>80.4661</td>
<td>79.3398</td>
</tr>
<tr>
<td>X8 associated</td>
<td>75.4034</td>
<td>75.4049</td>
<td>75.4757</td>
<td>75.4771</td>
<td>75.2774</td>
<td>75.228</td>
<td>75.5661</td>
<td>79.8789</td>
<td>76.1737</td>
<td>76.2123</td>
</tr>
<tr>
<td>X9 associated</td>
<td>70.1521</td>
<td>70.2061</td>
<td>69.9071</td>
<td>70.0853</td>
<td>70.4992</td>
<td>70.5206</td>
<td>70.7234</td>
<td>70.2437</td>
<td>559.598</td>
<td>70.276</td>
</tr>
</tbody>
</table>

Table B.7. Median of VIM\textsubscript{Gini} for all ten variables when all variables follow a standard normal distribution but each one with different variance (\(\Sigma = \text{diag}(50,45,40,35,30,25,20,15,10,1)\)), under H\textsubscript{A} in the ten association studies. Outcome continuous.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Table B.8. Median of VIM Gini for all variables that follow a normal distribution but each one with different variance ($\Sigma = \text{diag}(50,45,40,35,30,25,20,15,10,1)$), under H$_A$ in the ten association studies. Outcome binary.

<table>
<thead>
<tr>
<th>HA, Different variance</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
<tbody>
<tr>
<td>X4 associated</td>
<td>15.4285</td>
<td>15.4204</td>
<td>15.4359</td>
<td>176.281</td>
<td>15.5682</td>
<td>15.3853</td>
<td>15.4746</td>
<td>15.5394</td>
<td>15.5274</td>
<td>15.4204</td>
</tr>
<tr>
<td>X6 associated</td>
<td>17.2422</td>
<td>17.2055</td>
<td>17.2622</td>
<td>17.3333</td>
<td>17.2865</td>
<td>159.901</td>
<td>17.22</td>
<td>17.2538</td>
<td>17.2085</td>
<td>17.311</td>
</tr>
<tr>
<td>X7 associated</td>
<td>18.4044</td>
<td>18.5416</td>
<td>18.6428</td>
<td>18.4906</td>
<td>18.401</td>
<td>18.3707</td>
<td>149.184</td>
<td>18.3864</td>
<td>18.367</td>
<td>18.547</td>
</tr>
<tr>
<td>X9 associated</td>
<td>22.4644</td>
<td>22.2372</td>
<td>22.4644</td>
<td>22.3366</td>
<td>22.2368</td>
<td>22.491</td>
<td>22.391</td>
<td>114.405</td>
<td>22.3227</td>
<td></td>
</tr>
<tr>
<td>X10 associated</td>
<td>29.9166</td>
<td>29.9386</td>
<td>29.7888</td>
<td>29.731</td>
<td>29.8383</td>
<td>29.7921</td>
<td>29.8748</td>
<td>29.7025</td>
<td>29.7869</td>
<td>46.4456</td>
</tr>
</tbody>
</table>

Table B.9. Median of VIM Gini for all ten variables when all variables ~ N(0,1) but error has less variance N(0,0.5), under H$_A$ in the ten association studies. Outcome continuous.

<table>
<thead>
<tr>
<th>HA, xj<del>N(0,1), e</del>N(0,0.5)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
</table>
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

<table>
<thead>
<tr>
<th>HA, xj<del>N(0,1), e</del>N(0,0.5)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
<th>X10</th>
</tr>
</thead>
</table>

Table B.10. Median of VIM\textsubscript{Gini} for all ten variables when all variables follow a standard normal distribution but error has less variance N(0,0.5), under H\textsubscript{A} in the ten association studies. Outcome binary.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

**Codes chapter 3**

```r
# Number of database
n=500

# Import the library to generate multivariate normal distribution (independent)
library(mvnorm)

# Import the library for the binary data, to dichotomize with threshold zero.
library(biased)

# All ten variables are standard normal
mu0=rep(0,10)
sigma=diag(length(mu0))

# Create each database
for(i in 1:n){
  # 10 independent standard normal with 1000 observations
  x=rmvnorm(1000,mean=mu0,sigma=sigma)
  e=rnorm(1000,0,1) # the error follow a standard normal
  ye=rbinom(1000,1,0.5) # the outcome
  y=ye+r=ifelse(ye,1,0) # Create a binary outcome with with threshold 0
  data=data.frame(cbind(x,y)) dataframe with the outcome and the predictors
  # Save each database
  write.table(data, paste("DATA/chapter3/chapter3\s",i,"_\s",sep=""), row.names=FALSE, quote=FALSE)
}

# Now with an error with less variance
for(i in 1:n){
  # 10 independent standard normal with 1000 observations
  x=rmvnorm(1000,mean=mu0,sigma=sigma)
  # Generate normal distributed error with mean 0 and standard deviation 0.5
  e=rnorm(1000,0,0.5)
  ye=rbinom(1000,1,0.5) # the outcome
  y=ye+r=ifelse(ye,1,0) # Create a binary outcome with with threshold 0
  data=data.frame(cbind(x,y)) dataframe with the outcome and the predictors
  # Save each database
  write.table(data, paste("DATA/chapter3/chapter3\s",i,"_\s",sep=""), row.names=FALSE, quote=FALSE)
}

### predictor with different variance

# The mean 0 and each variable different variance
mu1=rep(0,10)
sigma=.diag(c(90,40,40,35,30,27,23,19,15,10,1))

# Loop to create 500 databases in each association study
for(i in 1:n){
  # The ten predictors have different variance and the error follow standard normal
  x=rmvnorm(1000,mean=mu1,sigma=sigma)
  e=rnorm(1000,0,1)
  ye=rbinom(1000,1,0.5) # the outcome under the null
  y=ye+r=ifelse(ye,1,0) # Create a binary outcome with with threshold 0
  data=data.frame(cbind(x,y))
  # Save each database
  write.table(data, paste("DATA/chapter3/chapter3\s",i,"_\s",sep=""), row.names=FALSE, quote=FALSE)
}

```
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code B.1. Generation of the databases with binary outcome under $H_0$, when all predictors and the error are standard normal, when predictors have different variance, when predictors have different precision, and when predictors $\sim N(0,1)$ and the error has less variance.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code B.2. Generation of the databases with binary outcome under $H_0$, when all predictors and the error are standard normal, when predictors have different variance, when predictors have different precision, and when predictors $\sim N(0,1)$ and the error has less variance.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code B.3. Generation of the databases with the continuous outcome under $H_0$, when all predictors and the error are standard normal, when predictors have different variance, when predictors have different precision, and when predictors ~ $N(0,1)$ and the error has less variance.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code B.4. Generation of the databases with continuous outcome under H1a, when all predictors and the error are standard normal, when predictors have different variance, when predictors have different precision, and when predictors \( \sim N(0,1) \) and the error has less variance.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code B.5. Extract the bias, coverage, and p-values when the outcome was binary under $H_0$ for all different studies. Example for the binary outcome. When the outcome was continuous the regressions where general linear models.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code B.6. Extract the bias, coverage, and p-values when the outcome was continuous under $H_A$ for all different studies. Example for the continuous outcome. When the outcome was binary the regressions were general logistic models with the probit link.
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code B.7. Generation of the figures to compare the VIMs when all predictors follow a standard normal distribution and when all variables have different number of decimal places under H₀. Also, extract the median of the VIMgini for each predictor in each case. Illustration of this particular case, to make the other plots and extract the median when all variables have different variance and when the error have difference variance, t² input was changed for those cases, and also medgini2 was the median VIMgini output for the particular case (all variables with different variance, or error with lower variance).
Studying the ability of finding single and interaction effects with Random Forest, and its application in Psychiatric genetics.

Code B.8. Generation of the figures under $H_A$ when all predictors and the error follow a standard normal distribution. Also, extract the median of the VIMgini for each predictor in case for all 10 association studies. Illustration of this particular case, to make the other plots and extract the median when all variables have different variance, different precision, and when the error have difference variance, 11 input was changed, and also medgini2 was the median VIMgini output for those cases (all variables with different variance, different precision and error with lower variance).