THE UNIVERSITY
of EDINBURGH

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 80
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 92

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 126
3.2.1.1 Data simulation under H_0 126
3.2.1.1.1 All variables follow a standard normal distribution 126
3.2.1.1.1.1 Continuous outcome 126
3.2.1.1.1.2 Binary outcome 127
3.2.1.1.2 All variables follow a normal distribution with different variances 127
3.2.1.1.2.1 Continuous outcome 127
3.2.1.1.2.2 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

3.2.3 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

5.5 Conclusions 206

References 207

Appendix A 242

Appendix B 284

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 V_2 and V_{90} (interacting variables), for two variable correlated V_3 and V_6, and for one independent variable V_{42}, when r = 0.80 and N = 40..... 93

Figure 2.12. RF VIMs under HA in the weakly-associated interaction study for V_2 and V_{90} (interacting variables), for two variable correlated V_3 and V_6, and for one independent variable V_{42}, when r = 0.10 and N = 5..... 96

Figure 2.13. RF VIMs under HA in the weakly-associated interaction study for V_2 and V_{90} (interacting variables), for two variable correlated V_3 and V_6, and for one independent variable V_{42}, when r = 0.40 and N = 20..... 97

Figure 2.14. RF VIMs under HA in the weakly-associated interaction study for V_2 and V_{90} (interacting variables), 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. 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 ($\sum$=diag(50,45,40,35,30,25,20,15,10,1), $\sum$ 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 ~ N(0,0.5). Continuous outcome. Each number i of the X axis corresponds to the subscript of the variable X_i. 164

Figure 3.14. 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 ~ N(0,0.5). Binary outcome..... 166

Figure 3.15. 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 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_i 170167

Figure 4.1. Illustration of the study design and methods. 181

Table 1.1. Summary of genome-wide significant findings for schizophrenia, bipolar disorder and MDD identified by GWAS (2008-2016).…… 27

Table 2.1. The nine different correlation conditions with the 3 different strengths of correlation (r) and 3 different number of correlated variables (N)……...61

Table 2.2. Bias and coverage of V2 (associated variable) in the strongly associated study (SAC)..... 69

Table 2.3. Bias and coverage of V2 (associated variable) in the weakly associated study (WAC). 70

Table 2.4. Number of p-values less than 0.0001. WAC mean weakly associated continuous studies. 70

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

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 1x10-5 on the weakly-associated interaction model study. 72

Table 2.8. Power of detecting V2 in the single strongly-associated study (SAC), VIMs, mtry=39 and mtry=27 Mindepth. 78

Table 2.9. Power of detecting V2 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 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 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 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 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 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 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 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..... 143

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. 144

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. 144

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. 155

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..... 155

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..... 158

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. 158

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. 160

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 bother 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, α 1C 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

H_0 : Null hypothesis

H_A : 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

$VIM_{rawperm\text{-}RF}$: unconditional unscaled permutation variable importance measure

$VIM_{Breiperm\text{-}RF}$: Breiman scaled permutation variable importance measure

$VIM_{Liawperm\text{-}RF}$: Liaw scaled permutation variable importance measure

$VIM_{rawperm\text{-}CF}$: 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

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

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. Murawieck 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).

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 disoder; 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 the100% 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

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

(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

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 refers to the genetic and E 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 $H^2 = S^2_G / S^2_P$, 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.

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

$$Period\ prevalence = \frac{\begin{matrix} Number\ of\ cases\ with\ the\ disease \\ at\ some\ point\ over\ that\ period \end{matrix}}{Number\ of\ population\ at\ risk\ over\ that\ period} \times 100$$

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

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

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).

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), but it is not a subtype of schizophrenia anymore. Since catatonia can be present in many other disorders, in DSM-V it is treated as a specifier for psychotic disorders. Moreover, catatonia can be diagnosed in cases where the medical disorder or psychiatric condition is unknown, so now it is a new clinical entry instead (Tandon *et al.* 2013).

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

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

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 (DTNBP1)* (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

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 (R^2 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

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).

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

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

(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 (DTNBP1)*, 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).

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 *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 *et al.* 2014) (rare variants that have relativily 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).

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 *et al.* 2008) found significant associations in schizophrenia passing the GWAS threshold p-value 5×10^{-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 *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

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 = 4.79×10−8) 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 = 1.1×10^{-9}); two SNPS which were the most significant ones in *protease, serine 16 (PRSS16)* with *p*-values 1.3×10^{-10} and 1.4×10^{-12}; and the other two in *pterin-4 alpha-carbinolamine dehydratase 1 (PGBD1)* and the *NOTCH4,* with *p*-values 8.3×10^{-11} and 2.3×10^{-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 = 2.4×10^{-9}) and in *transcription factor 4 (TCF4)* (*p*-value = 4.1×10^{-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 = 2.65×10^{-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 = 3.48×10^{-31}) was on chromosome 6, in the MHC region.

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*

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 4.53×10^{-9} and 5.87×10^{-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 = 3.4×10^{-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 x 10^{-4}). Moreover, the SNP showed significant association in meta-analysis including the validation sample and each replication cohort (p-value = 5.2×10^{-11}, OR = 1.16, 95% CI: (1.11, 1.21)).

STUDY	DISEASE(S) OF INTEREST	FINDINGS
Baum *et al.* (2008)	BP	Reported genome-wide significant association between bipolar disorder and DGHK in two independent samples.
O'donovan *et al.* (2008)	Schizophrenia and schizophrenia and BP combined	Reported strong evidence for association between ZNF804A and schizophrenia, attaining genome-wide significance when both schizophrenia and bipolar disorder were considered.
Ferreira *et al.*, (2008)	BP	First study to report genome-wide significant associations with bipolar disorder implicating CACNA1C and ANK3.
Stefansson *et al.* (2009)	Schizophrenia	Identified genome-wide significant association with schizophrenia at loci in the major histocompatibility complex (MHC), NRGN, and TCF4.
Steinberg *et al.* (2011)	Schizophrenia	Provided evidence in support of association between schizophrenia and NRGN and TCF4. Identified two novel loci associated with schizophrenia at CCDC68 and VRK2

Cichon *et al.* (2011)	BP	Identified a genome-wide association between bipolar disorder and NCAN.
PGC (Ripke *et al.*, 2011)	Schizophrenia and schizophrenia and BP combined	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.
PGC (Sklar *et al.*, 2011)	BP and schizophrenia and BP combined	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.
Chen *et al.* (2013)	BP	Genome wide significant association with bipolar disorder reported near TRANK1, LMAN2L and PTGFR. Also

		provided support for association with ANK3.
Ruderfer *et al.* (2014)	BP and schizophrenia	Identified a novel association between both disorders and (PIK3C2A), as well as five previously-identified loci (TRANK1, MHC, MAD1L1, and CACNA1C)
PGC (Ripke *et al.*, 2014)	Schizophrenia	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
Converge Consortium (Cai *et al.*, 2015)	MDD	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.
Hou *et al.* (2016)	BP	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.
Hyde *et al.* (2016)	MDD	15 novel genome-wide significant loci were identified in a meta-analysis of a

		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 = 9.4×10^{-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 = 3.0×10^{-19}, including early-onset cases vs controls; R^2 = 0.14%, p-value = 3.9×10^{-5} late-onset cases vs controls) and BP (R^2 = 0.41%, p-value = 1.4×10^{-12} early-onset cases vs controls; R^2 = 0.16%, p = 1.9×10^{-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).

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* =

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 *NDEL1/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).

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).*

1.3. 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

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 year (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 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

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 (5×10^{-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. 5×10^{-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

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 (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

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

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,

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 $(X_1, Y_1), \ldots, (Xn, Yn)$, 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

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:

$$D = \{y_i, x_{i1}, x_{i2}, \dots, x_{iN} \vee i = 1, \dots, n\} = \{(Y, X_j) | j = 1, \dots, N\}$$

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,θ) where θ 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:

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

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:

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

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

$$Gini\ index = 1 - \sum_{j=1}^{N} P_j^2$$

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

called mtry (X_{mtry} C X, X_{nxN} 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 € {1, … 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).

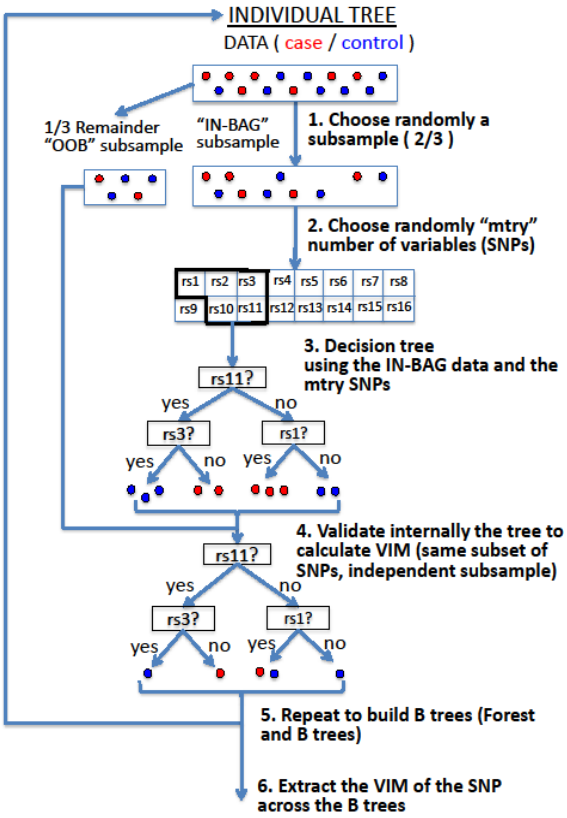

Figure 1.1. Illustration of the process of RF algorithm. The exemplificative SNPs are annotated as in dbSNP database, reference SNP ID number. In the third step the split criteria is based on whether the individual carries the risk allele at the SNP chosen in that node. Note that mtry is resampled at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

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 $w.x + b = 0$ 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).

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

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

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.

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 H_0. 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 H_0 and two conditions under H_A: weakly and strongly associated. The simulation study considered variations in the (a) number of correlated predictors and (b) strength of correlation between predictors.

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

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 (Janitza *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).

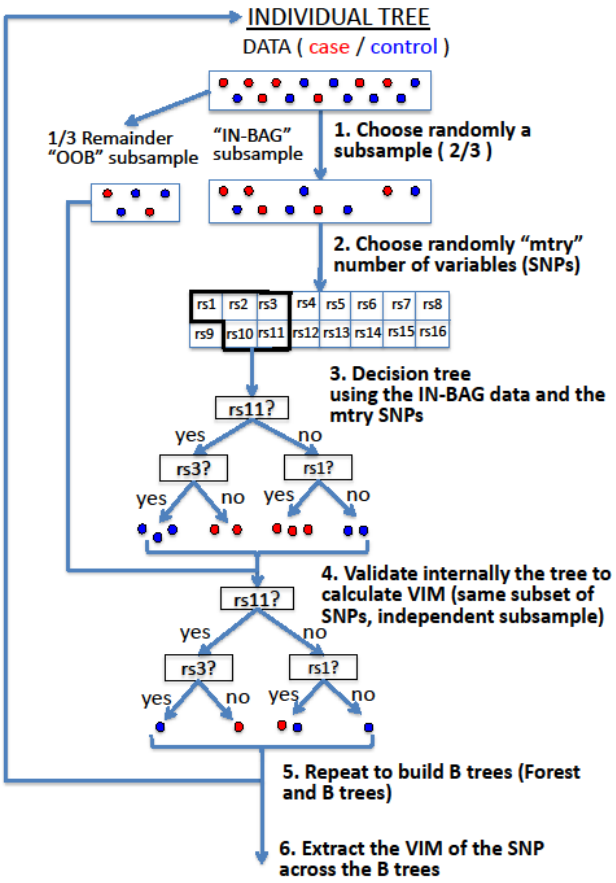

Figure 2.1. Illustration of the process of RF algorithm. The exemplificative SNPs are annotated as in dbSNP database, reference SNP ID number. In the third step the split criteria is based on whether the individual carries the risk allele at the SNP chosen in that node. Note that mtry is resampled at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

2.2.1.1. Variable Importance Measures

The two fundamentally different VIMs in RF are the Gini importance ($VIM_{Gini\text{-}RF}$) and the permutation importance (PVIM) (Breiman 2001). In the $VIM_{Gini\text{-}RF}$, at each split in each tree b, with b∈{1,…,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∈{1,…,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 $VIM_{Gini\text{-}RF}$ (computed in the in-bag sample) of a predictor variable X_i is the total decrease in impurity Δ 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,

$$\Delta\, I = \sum\nolimits_k \Delta\, i_k$$

The standard PVIM, $VIM_{rawperm\text{-}RF}$, 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:

$$VI_{x_i}^{ER} = \frac{1}{|T|}\sum_{t\in T} A_t - A_{t^*}$$

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 ($VIM_{Mengperm\text{-}RF}$) 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 $VIM_{rawperm\text{-}RF}$, except the average is taken across all the trees in the forest containing that predictor instead of across all trees, hence the $VIM_{Mengperm\text{-}RF}$ is:

$$VI_{x_i}^{ER} = \frac{1}{|T_{X_i}|}\sum_{t\in T_{X_i}} A_t - A_{t^*}$$

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 $VIM_{rawperm\text{-}RF}$, as expected. Other studies have proposed different VIMs based on the $VIM_{rawperm\text{-}RF}$, such as scaled PVIMs, which divide the PVIM by its empirical standard error over all the trees in the forest. More precisely,

$$VI_{x_i}^{SER} = \frac{VI_{X_i}^{ER}}{\sqrt{\frac{s^2}{|T|}}}$$

I compared two scaled PVIMs, Breiman's ($VIM_{Breiperm\text{-}RF}$) and Liaw's ($VIM_{Liawperm\text{-}RF}$), the difference between them is the variance estimator. The estimator of $VIM_{Breiperm\text{-}RF}$ is defined as

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

where $N_{OOB,t}$ is the number of samples in OOB of the tree t. And the estimator of $VIM_{Liawperm\text{-}RF}$ 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 $VIM_{Breiperm\text{-}RF}$ 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 ($VIM_{rawperm\text{-}CF}$) which is the same as the $VIM_{rawperm\text{-}RF}$ 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 $VIM_{rawperm\text{-}RF}$ 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 $VIM_{rawperm\text{-}CF}$ 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).

The VIM_{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 $VIM_{rawperm\text{-}RF}$ where the AUC is the probability to detect a random true value as a true positive rather than a false positive (Janitza *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 *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, *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_{Xi} an X_i -subtree of T if the root node of T_{Xi} is split using X_i. Call T_{Xi} a maximal X_i -subtree if T_{Xi} is not a subtree of a larger X_i -subtree (Ishwaran *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 *et al.* 2010). It

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

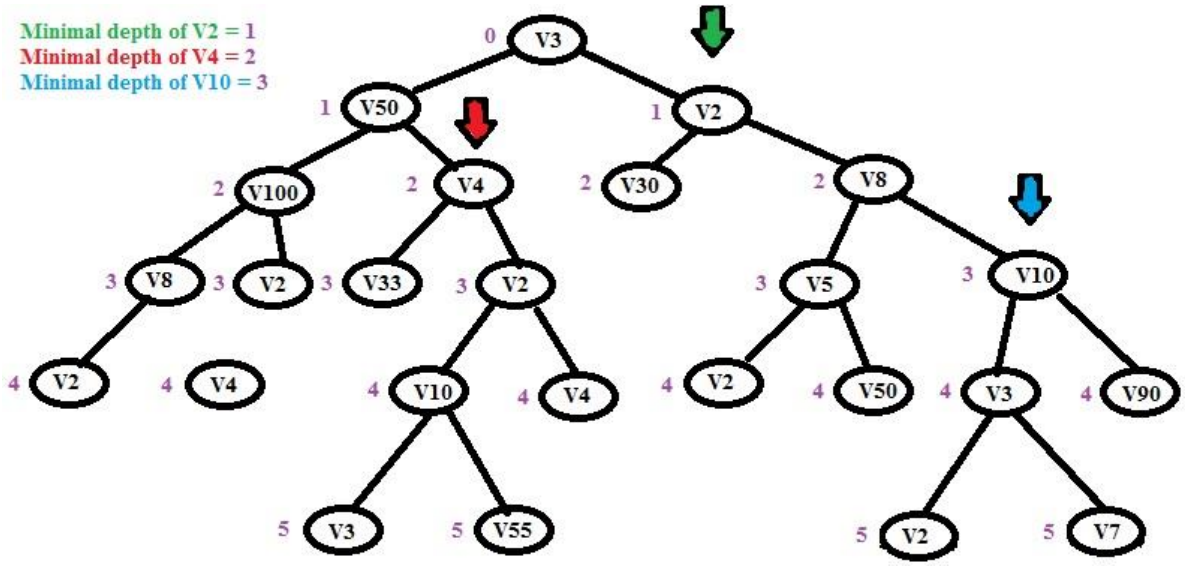

Figure 2.2. Illustration of RF based in minimal depth. As an example, it shows the minimal depth of V2, V4 and V10.

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 χ^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 $VIM_{rawperm\text{-}RF}$. 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

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

r	**80**	**80**	**80**	**40**	**40**	**40**	**10**	**10**	**10**
N	5	20	40	5	20	40	5	20	40

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

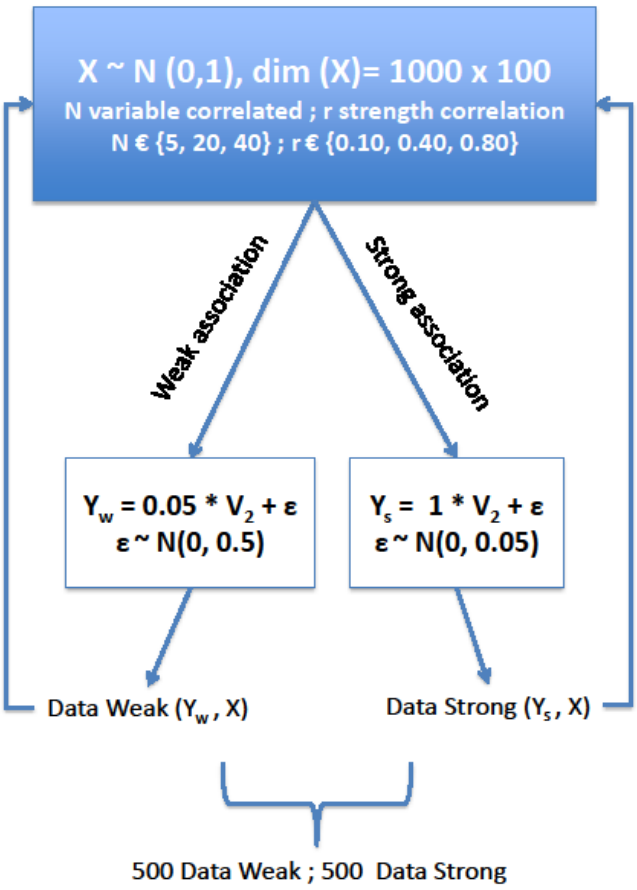

Figure 2.3. Generation of the single association study as an example of the data simulation.

2.2.2.1. Simulation under H_A

2.2.2.1.1. Single association

I performed the simulations as follows:

$$y = \beta_1 * V_2 + \varepsilon$$

where β_1 was set to 1 in the strongly-associated case and to 0.05 in the weakly associated case. The error was set to $\varepsilon \in N(0,0.05)$ for the strongly-associated case and $\varepsilon \in N(0,0.5)$ for the weakly-associated case.

2.2.2.1.2. Interaction association

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

$$y = \beta_1 * V_2 + \beta_2 * V_{90} + \beta_3 * V_2 * V_{90} + \varepsilon$$

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

$$y = 0 * V_2 + \varepsilon$$

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 $VIM_{Gini\text{-}RF}$ 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

V_{90} 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 H_0 and H_A (including weakly and strongly associated conditions when a single predictor V_2 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 β_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.

SAC	**r=0.80**		**r=0.40**		**r=0.10**	
N	BIAS	COV%	BIAS	COV%	BIAS	COV%
5	0.00012	94.8	0.00008	94.0	-0.00006	95.2
20	0.000001	91.8	0.000022	93.8	0.000058	96.2
40	0.00007	93.8	0.000068	94.4	0.000023	96.0

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

WAC	r=0.80		r=0.40		r=0.10	
N	BIAS	COV%	BIAS	COV%	BIAS	COV%
5	0.00110	95.6	0.00122	94.8	-0.00027	97.6
20	0.00102	94.8	-0.00139	94.8	0.00135	95.8
40	-0.00125	94.8	-0.00131	94.4	0.00129	95.6

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

p-value	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
WAC	183	211	171	201	170	173	117	196	200

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.

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).

SAC	**r=0.80**		**r=0.40**		**r=0.10**	
N	BIAS	COV%	BIAS	COV%	BIAS	COV%
5	-0.00002	94.4	0.000005	95	0.000009	94.6
20	-0.00004	94.6	-0.00003	94.4	0.0000008	94.4
40	-0.00009	93.4	-0.00007	94.6	-0.000003	93.8

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

WAC	**r=0.80**		**r=0.40**		**r=0.10**	
N	BIAS	COV%	BIAS	COV%	BIAS	COV%
5	0.00062	92.6	0.00144	95.8	0.00009	93.8
20	-0.00139	93.6	0.00045	95.4	0.00094	94.2
40	0.00043	94.8	-0.00110	94.8	0.00096	93.4

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

p-value	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
WAC	459	452	457	461	460	460	457	456	464

Table 2.7. Number of p-values less than p-value threshold 1x10-5 on the weakly-associated interaction model study.

2.3.2. Distributions under H_0

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 H_0 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 $VIM_{rawperm\text{-}CF}$ 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_{AUC} and VIM_{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

minimal depth. In this section, three conditions (r = 0.10 and N = 5, r = 0.40 and N = 20, r = 0.80 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.).

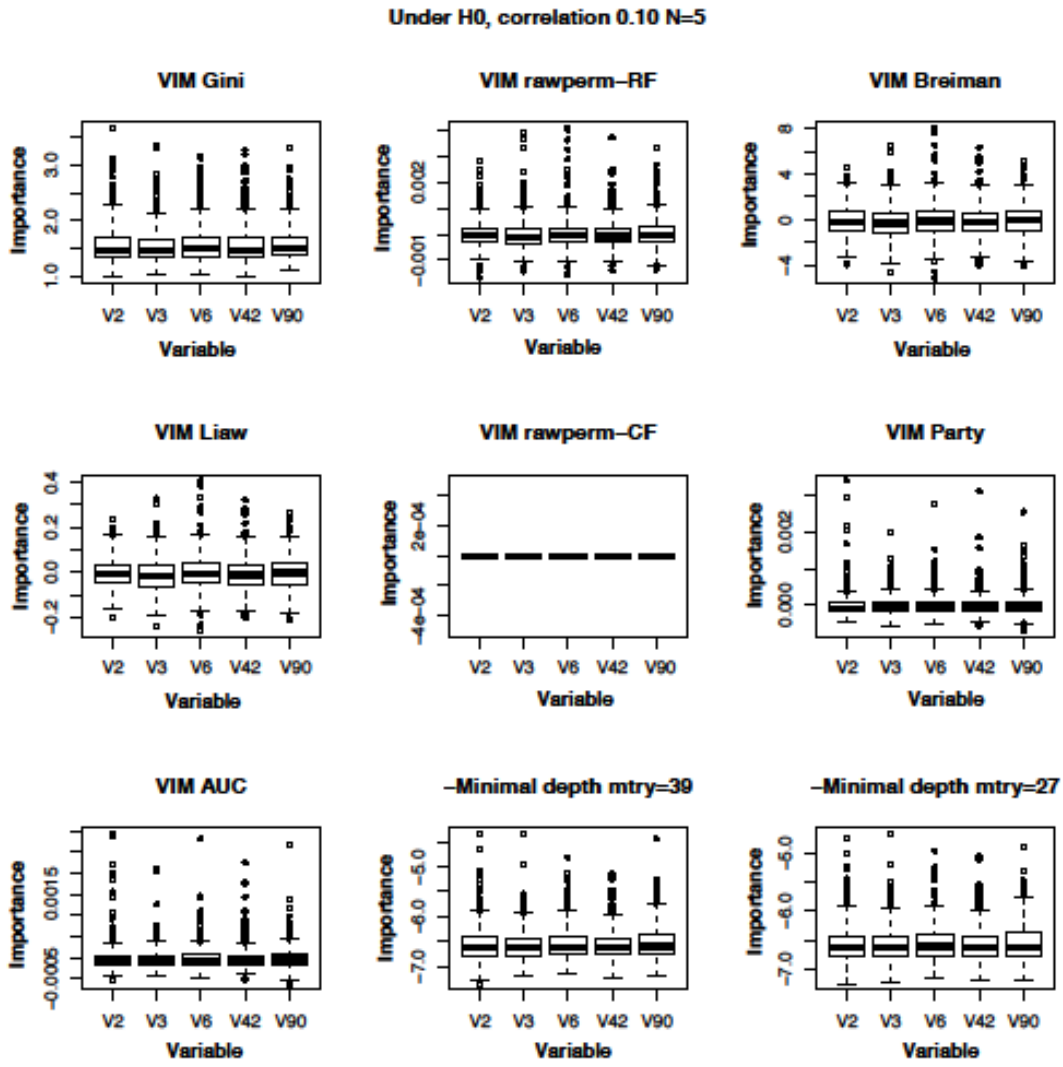

Figure 2.4. RF VIMs, minimal depth, VIMAUC and VIMparty under H0 for V_2, two variable correlated V_3 and V_6, and two independent variables V_{42} and V_{90} when r = 0.10 and N = 5.

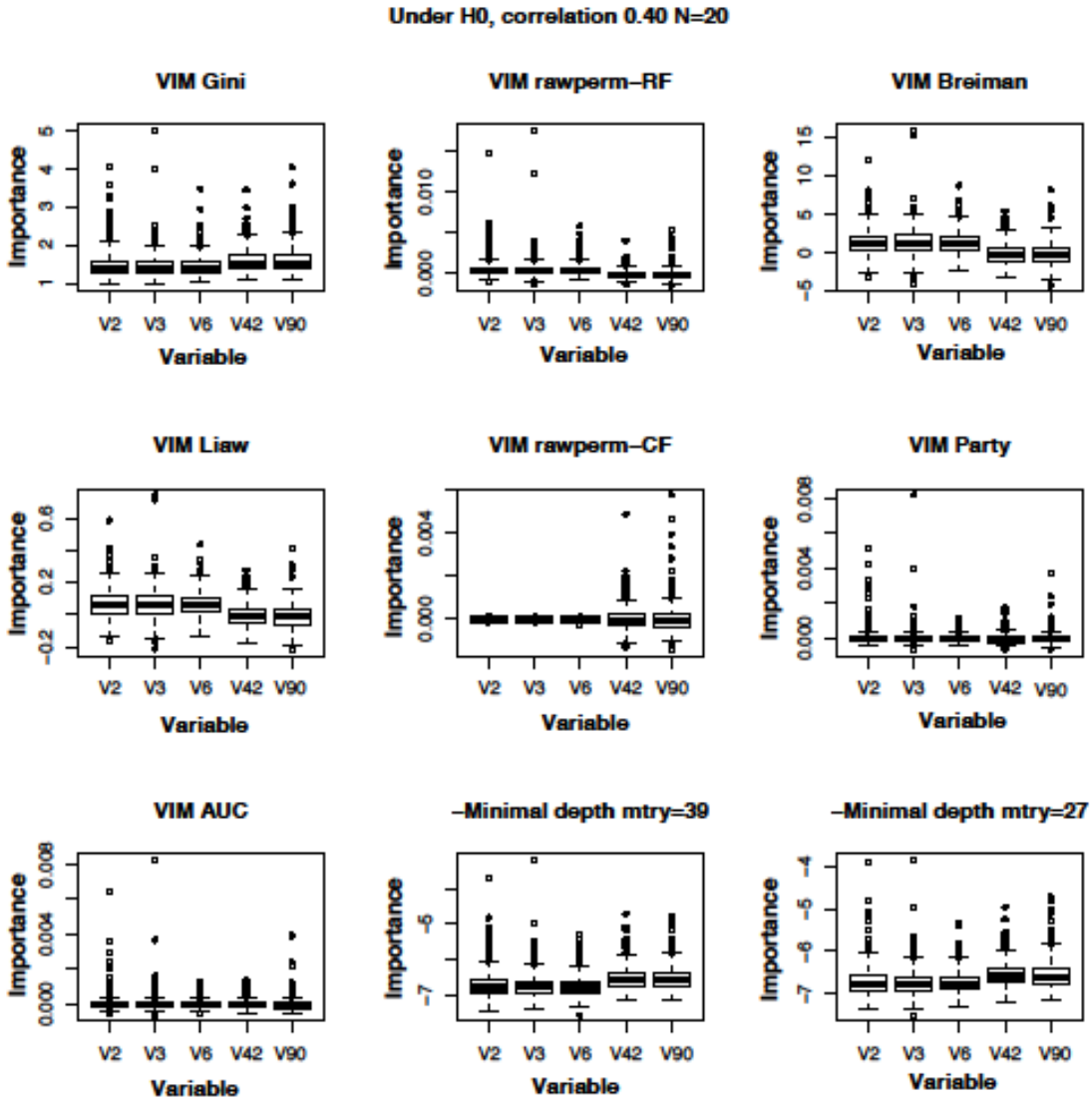

Figure 2.5. RF VIMs, minimal depth, VIMAUC and VIMparty under H0 for V_2, two variable correlated V_3 and V_6, and two independent variables V_{42} and V_{90} when r = 0.40 and N = 20.

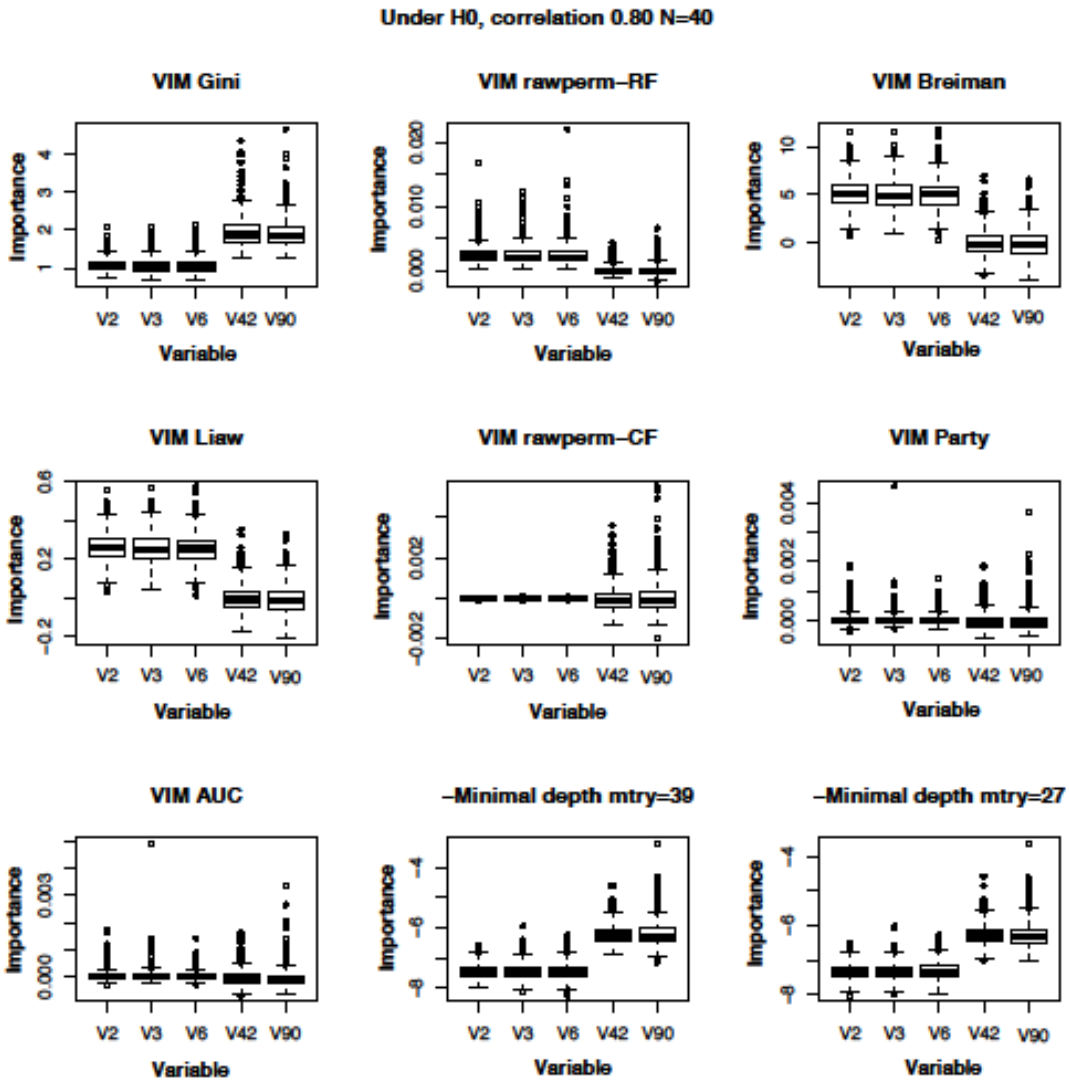

Figure 2.6. RF VIMs, minimal depth, VIMAUC and VIMparty under H0 for V_2, two variable correlated V_3 and V_6, and two independent variables V_{42} and V_{90} when r = 0.80 and N = 40.

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 Wilcoxson test were extracted to test that difference formally (under predictor correlation conditions, all showed to be less than 0.05). VIM_{Gini} and scaled PVIMs were biased under predictor correlation under H_0, 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

2.6). VIM_{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 $VIM_{Breiman\text{-}RF}$ and $VIM_{Liawman\text{-}RF}$ 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_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, $VIM_{rawperm\text{-}RF}$, VIM_{AUC} and VIM_{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, $VIM_{rawperm\text{-}RF}$ 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). $VIM_{rawperm\text{-}CF}$ was observed to be unbiased, but with more variability for uncorrelated than for the correlated predictors. This small inflation on the $VIM_{rawperm\text{-}RF}$ was also reported by Nicodemus (2010); the authors also

showed more inflation in $VIM_{rawperm\text{-}RF}$ than the unconditional unscaled PVIM from CIF and the conditional PVIM under H_0 (they used CIF rather than RF). In this study, $VIM_{rawperm\text{-}CF}$ 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 $VIM_{rawperm\text{-}CF}$ under predictor correlation than the original study (Strobl *et al.* 2008). $VIM_{rawperm\text{-}CF}$ 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 $VIM_{rawperm\text{-}CF}$: 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 $VIM_{rawperm\text{-}CF}$ is applied in each tree (see section 2.3.5. for the explanation of the behaviour).

SAC	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
GINI	100	100	100	100	100	100	100	100	100
rawpermRF	100	100	100	100	100	100	100	100	100
BREIMAN	100	100	100	100	100	100	100	100	100
LIAW	100	100	100	100	100	100	100	100	100
rawpermCF	100	53.75	0.034	100	32.46	5.03	0	0	0
Party	100	100	100	100	100	100	100	100	100
AUC	100	100	100	100	100	100	100	100	100
mindepth 39	100	100	100	100	100	100	100	100	100
mindepth 27	100	100	100	100	100	100	100	100	100

Table 2.8. Power of detecting V2 in the single strongly-associated study (SAC), VIMs, mtry=39 and mtry=27 Mindepth.

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 V_2 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

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 VIM_{party} (unconditional unscaled PVIM from CIF) but not the behavior of VIM_{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 VIM_{AUC} had the highest power when the correlation was medium-low, however when the correlation was r = 0.80, the unconditional unscaled PVIM ($VIM_{rawperm-RF}$) 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). $VIM_{rawperm-CF}$, 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.

WAC	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
GINI	34.65	10.10	4.89	69.66	48.76	43.64	70.69	74.38	73.42
rawpermRF	75.93	74.77	50.40	74.62	66.58	61.93	66.28	71.92	73.07
BREIMAN	72.81	67.98	39.21	66.86	61.33	52.62	57.00	63.80	65.79
LIAW	72.80	67.94	39.21	66.86	61.33	52.64	57.00	63.81	65.78
rawpermCF	13.25	0.00	0.00	45.21	0.00	0.00	0.00	0.00	0.00
Party	65.43	59.75	45.75	80.67	68.57	67.06	79.54	82.36	81.67
AUC	65.47	60.89	45.41	81.03	69.81	68.45	79.68	82.91	81.93
mindepth 39	49.55	12.89	4.45	79.15	59.94	54.01	80.91	82.44	81.33
mindepth 27	51.52	14.08	5.10	79.22	60.56	56.20	80.35	82.22	80.67

Table 2.9. Power of detecting V2 in the single weakly-associated study (WAC), VIMs, mtry=39 and mtry=27 Mindepth.

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). $VIM_{rawperm\text{-}CF}$ 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.

SAC V_2	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
GINI	100	100	100	100	100	100	100	100	100
rawpermRF	100	100	100	100	100	100	100	100	100
BREIMAN	100	100	100	100	100	100	100	100	100
LIAW	100	100	100	100	100	100	100	100	100
rawpermCF	100	2.56	0.43	100	20.61	2.07	0	0	0
Party	100	100	100	100	100	100	100	100	100
AUC	100	100	100	100	100	100	100	100	100
mindepth 39	100	100	100	100	100	100	100	100	100
mindepth 27	100	100	100	100	100	100	100	100	100

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

SAC V_{90}	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
GINI	100	100	100	100	100	100	100	100	100
rawpermRF	100	100	100	100	100	100	100	100	100
BREIMAN	100	100	100	100	100	100	100	100	100
LIAW	100	100	100	100	100	100	100	100	100
rawpermCF	100	100	100	100	100	100	0	0	0
Party	100	100	100	100	100	100	100	100	100
AUC	100	100	100	100	100	100	100	100	100
mindepth 39	100	100	100	100	100	100	100	100	100
mindepth 27	100	100	100	100	100	100	100	100	100

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

In the weakly-associated study, VIM_{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.

WAC V_2	**r=0.80**			**r=0.40**			**r=0.10**		
N	5	20	40	5	20	40	5	20	40
GINI	27.66	4.83	1.24	52.23	39.96	37.49	57.94	59.63	59.20
rawpermRF	64.59	55.66	36.19	58.20	58.58	53.53	56.43	59.62	60.05
BREIMAN	66.48	51.89	30.99	55.84	56.95	48.68	53.45	56.29	56.88
LIAW	66.46	51.87	30.94	55.85	56.96	48.68	53.45	56.29	56.88
rawpermCF	8.65	0.02	0	31.20	0.16	0	0	0	0
Party	67.01	53.29	43.62	76.76	73.63	71.05	76.32	77.65	77.17
AUC	65.40	52.89	43.86	75.86	72.25	71.43	76.19	79.00	76.81
mindepth 39	28.01	4.88	0.45	51.09	41.61	36.69	58.76	57.77	56.48
mindepth 27	30.01	4.59	0.49	49.59	40.95	37.44	56.52	57.55	55.68

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_{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*

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, $VIM_{rawperm\text{-}CF}$ 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 $VIM_{rawperm\text{-}CF}$ showed low power (Table 2.13). $VIM_{rawperm\text{-}CF}$ was dramatically more powerful in detecting the signal from the uncorrelated interacting predictor than from the correlated one. This suggests that $VIM_{rawperm\text{-}CF}$ 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. $VIM_{rawperm\text{-}CF}$ 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.

WAC V_{90}	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
GINI	65.73	75.98	86.53	61.99	65.07	71.84	60.87	59.48	65.53
rawpermRF	65.78	61.93	50.90	57.53	59.57	52.64	61.55	57.15	63.03
BREIMAN	63.91	51.54	41.44	54.23	55.43	42.01	57.57	54.58	59.16
LIAW	63.91	51.54	41.47	54.24	55.45	42.03	57.57	54.58	59.17
rawpermCF	67.05	81.34	89.74	59.75	67.57	77.36	0.00	0.00	0.00
Party	82.59	91.77	94.71	78.59	85.00	85.40	77.97	77.00	78.94
AUC	82.58	91.82	94.84	79.68	84.92	87.42	77.44	76.47	79.57
mindepth 39	67.06	79.65	88.57	60.27	65.71	75.29	60.78	57.99	64.74
mindepth 27	65.87	79.94	89.46	58.34	64.21	74.98	59.70	56.98	62.24

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_{Gini}, $VIM_{rawperm\text{-}CF}$ 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

2.2.2.2.). 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 H_A

2.3.4.1. Single association study

After applying different VIMs and minimal depth from different implementations of RF under H_A, 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 VIM_{Gini} 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 V_2 and the outcome is weak. See Appendix A for the VIM median values for V_2, 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 H_A, RF based on the different VIMs, VIM_{AUC}, VIM_{party} 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 $VIM_{rawperm\text{-}CF}$ when the correlation was low (r = 0.10) that gave no importance to all variables. In general, it

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.

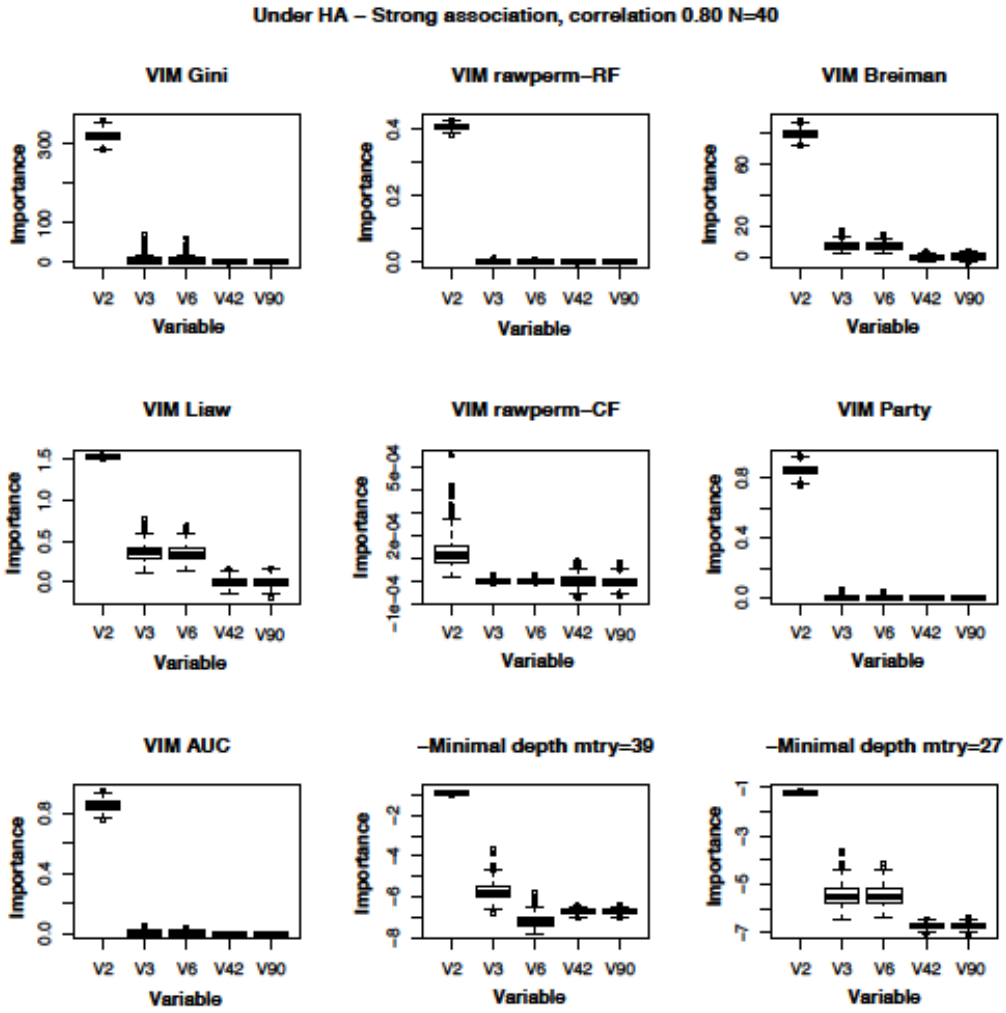

Figure 2.7. RF VIMs under HA for V_2, two variable correlated V_3 and V_6, and two independent variables V_{42} and V_{90} when r = 0.80 and N = 40 in the strongly-association single study.

In spite of the fact that in the strongly-associated condition VIM_{Gini} ranked the non-associated correlated predictors higher than uncorrelated ones, VIM_{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, VIM_{Gini} resulted in lower values for the influential variable than the uncorrelated predictors. This inflation for the uncorrelated predictors by VIM_{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 VIM_{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 VIM_{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 $VIM_{rawperm\text{-}RF}$ 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 ($VIM_{rawperm\text{-}RF}$ 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. VIM_{party} and VIM_{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 VIM_{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 $VIM_{rawperm\text{-}CF}$ in this study also showed less variability for correlated predictors than $VIM_{rawperm\text{-}RF}$, 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 V_2, 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.

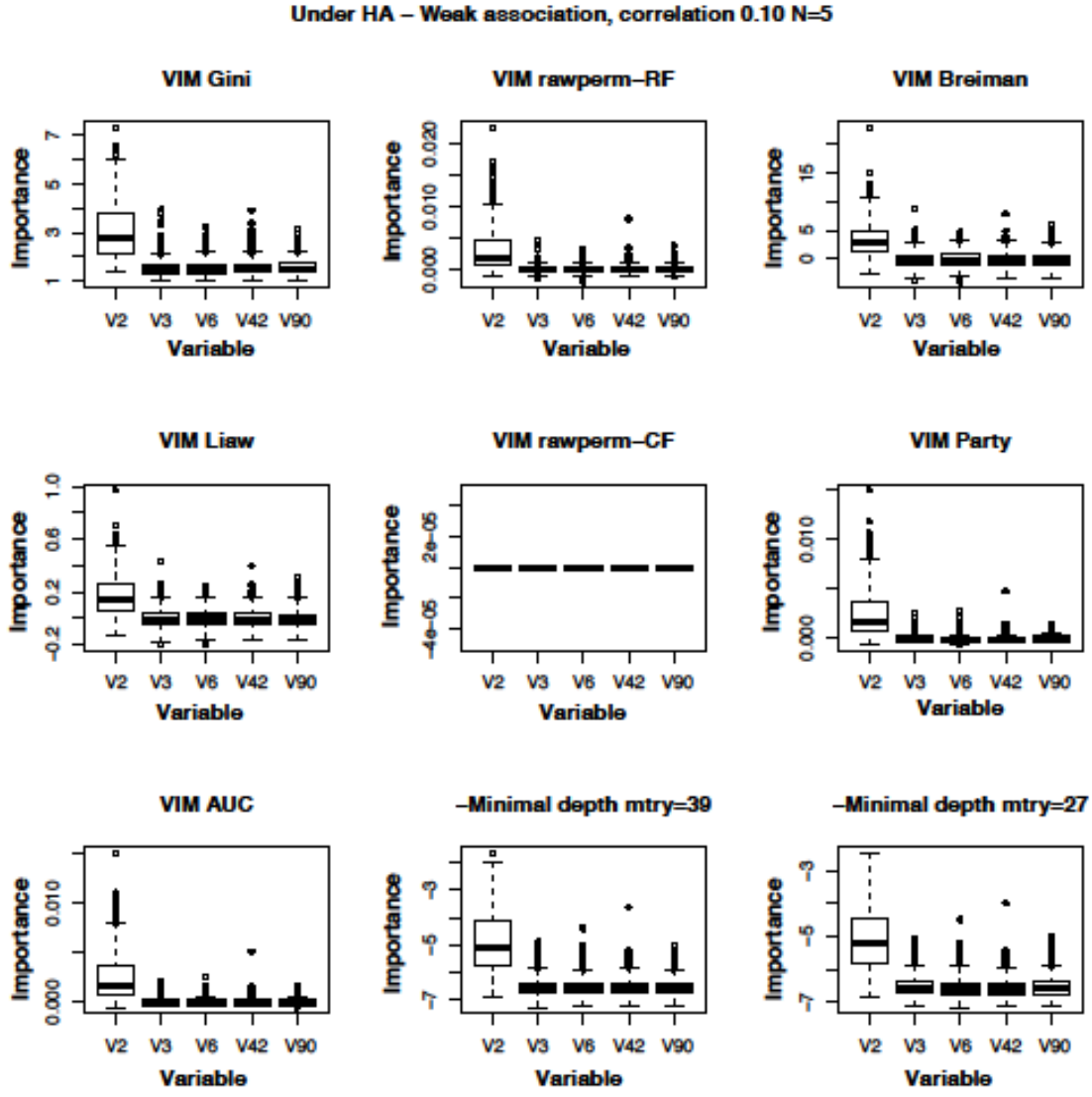

Figure 2.8. RF VIMs under HA for V_2, two variable correlated V_3 and V_6, and two independent variables V_{42} and V_{90} when r = 0.10 and N = 5 in the weakly-association single study.

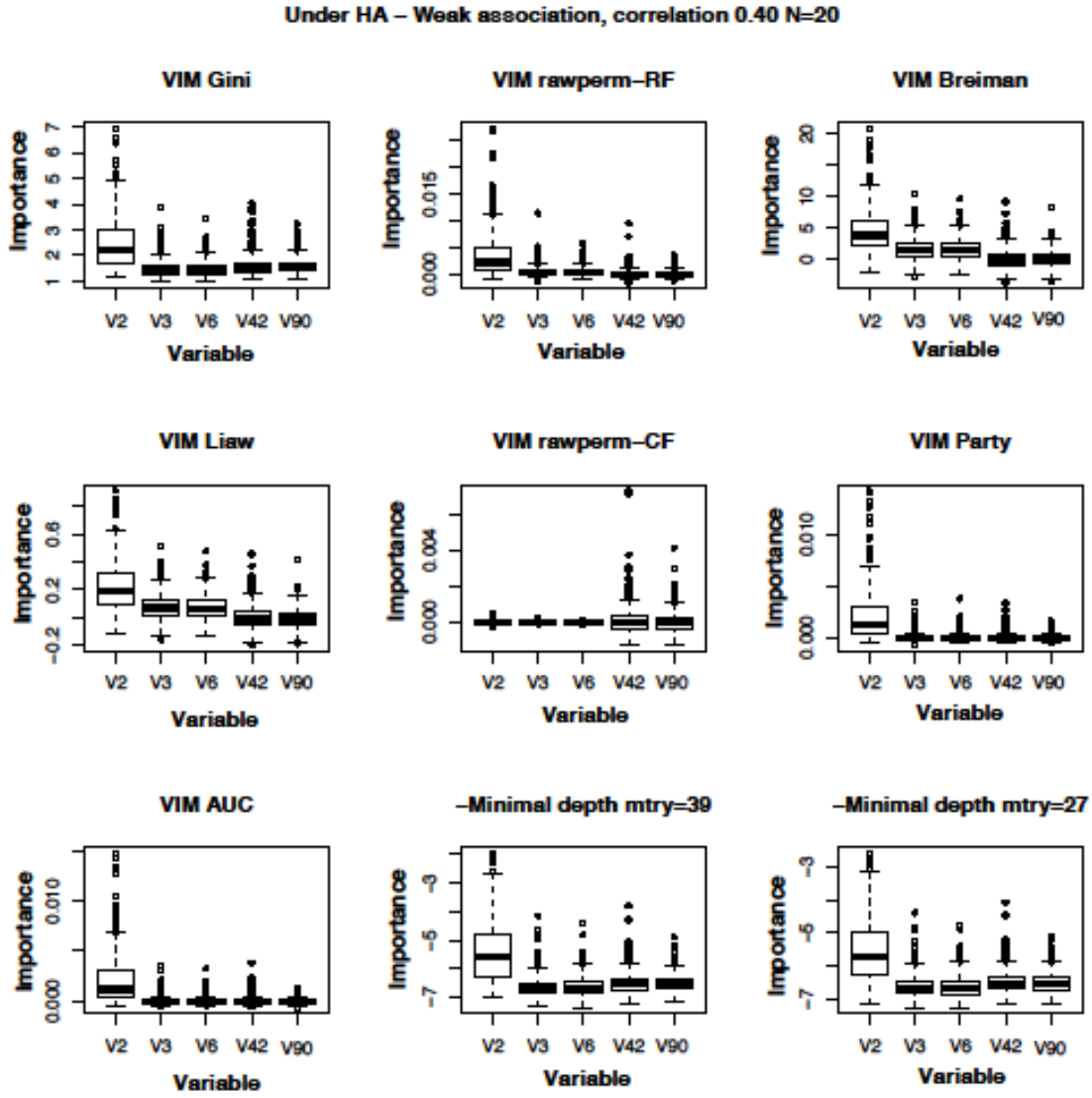

Figure 2.9. RF VIMs under HA for V2, for two variable correlated V3 and V6, and for two independent variables V42 and V90 when r = 0.40 and N = 20 in the weakly-association single study.

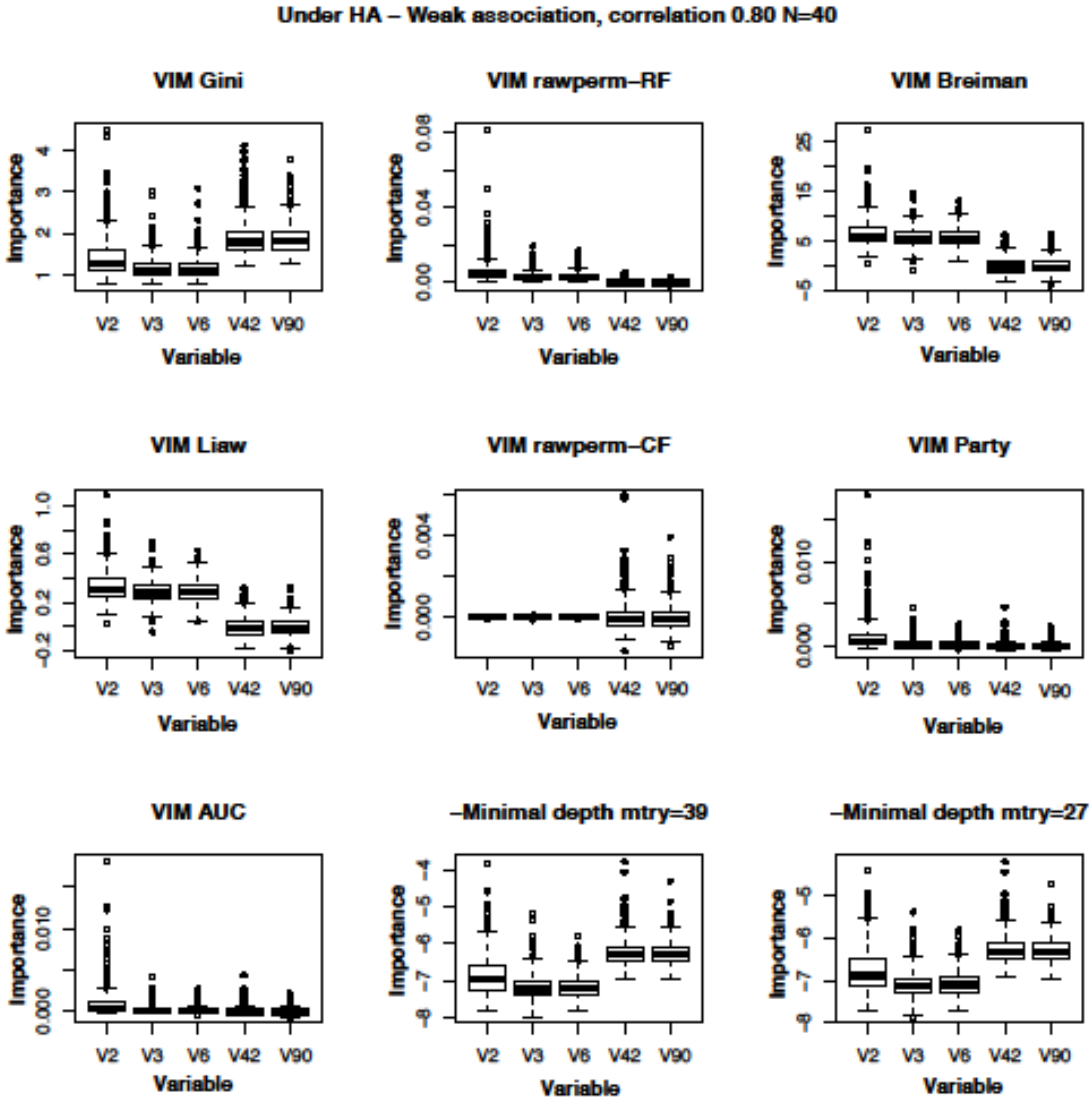

Figure 2.10. RF VIMs under HA for V2, for two variable correlated V3 and V6, and for two independent variables V42 and V90 when r = 0.80 and N = 40 in the weakly-association single study.

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 assocation studies (Figures A.21 - A.28 in the strong assocation 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 $VIM_{rawperm\text{-}CF}$ 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.

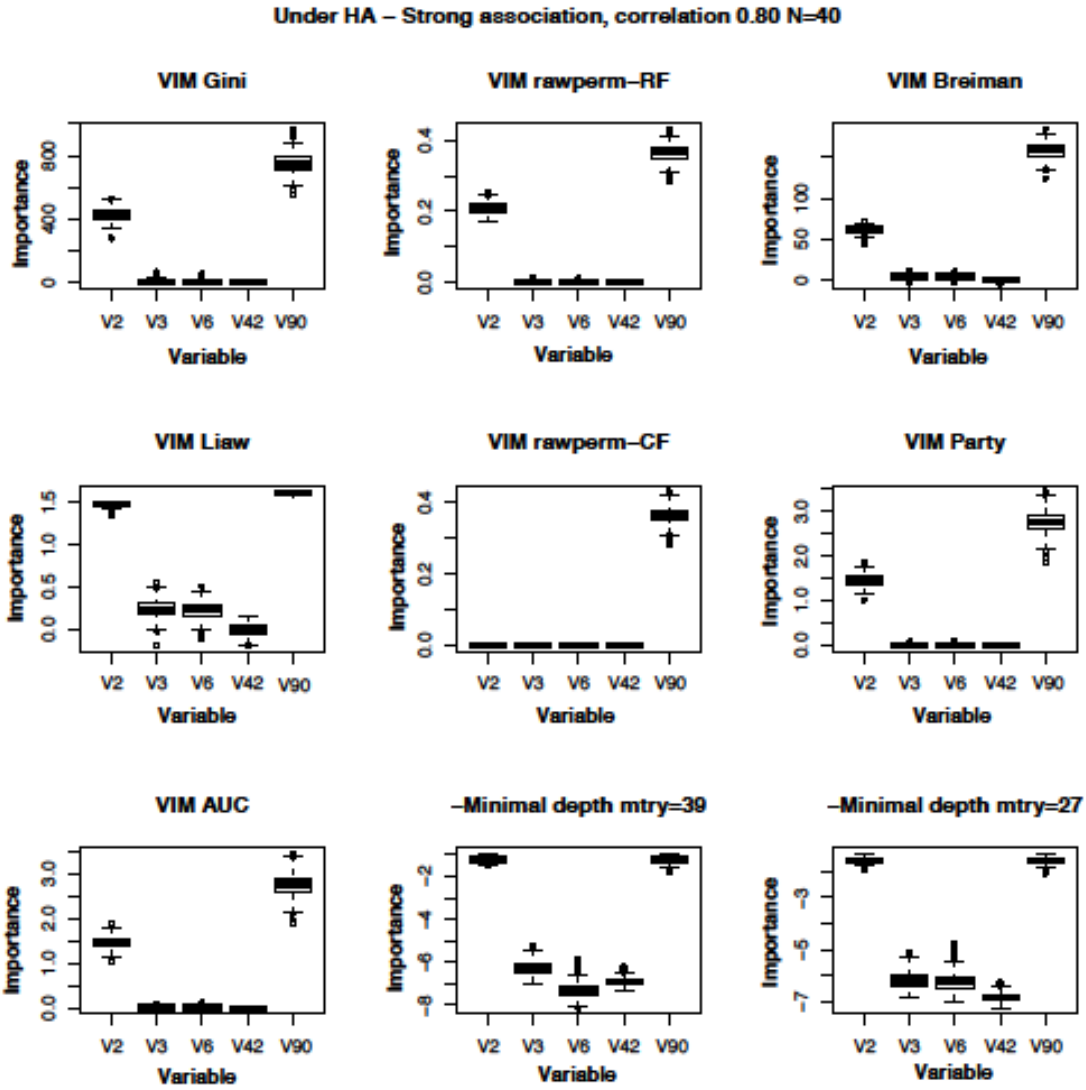

Figure 2.11. RF VIMs under HA in the strongly-associated interaction study for V_2 and V_{90} (interacting variables), two variable correlated V_3 and V_6, and one independent variable V_{42}, when r = 0.80 and N = 40.

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 associted 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. VIM_{Gini} and minimal depth gave greater importance to uncorrelated variables than to correlated ones, including the interacting correlated one. Minimal depth with mtry = 27 ranked V_2 slighty higher than with mtry =39 when r = 0.80 (for V_{90} it was the opposite). Under high correlation, permuted unconditional PVIMs, VIM_{AUC} and VIM_{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 VIM_{AUC} and VIM_{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 $VIM_{rawperm\text{-}CF}$ only suggested association with V90, the uncorrelated interacting variable. The inflation of $VIM_{rawperm\text{-}CF}$ 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 $VIM_{rawperm\text{-}CF}$ 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

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_{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_{Gini} also preferred both interacting predictors with low-medium correlation than non-associated predictors. However, with high correlation (r = 0.80), VIM_{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_{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). Mininal depth showed slightly larger scores for both interacting variables wih 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 *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

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

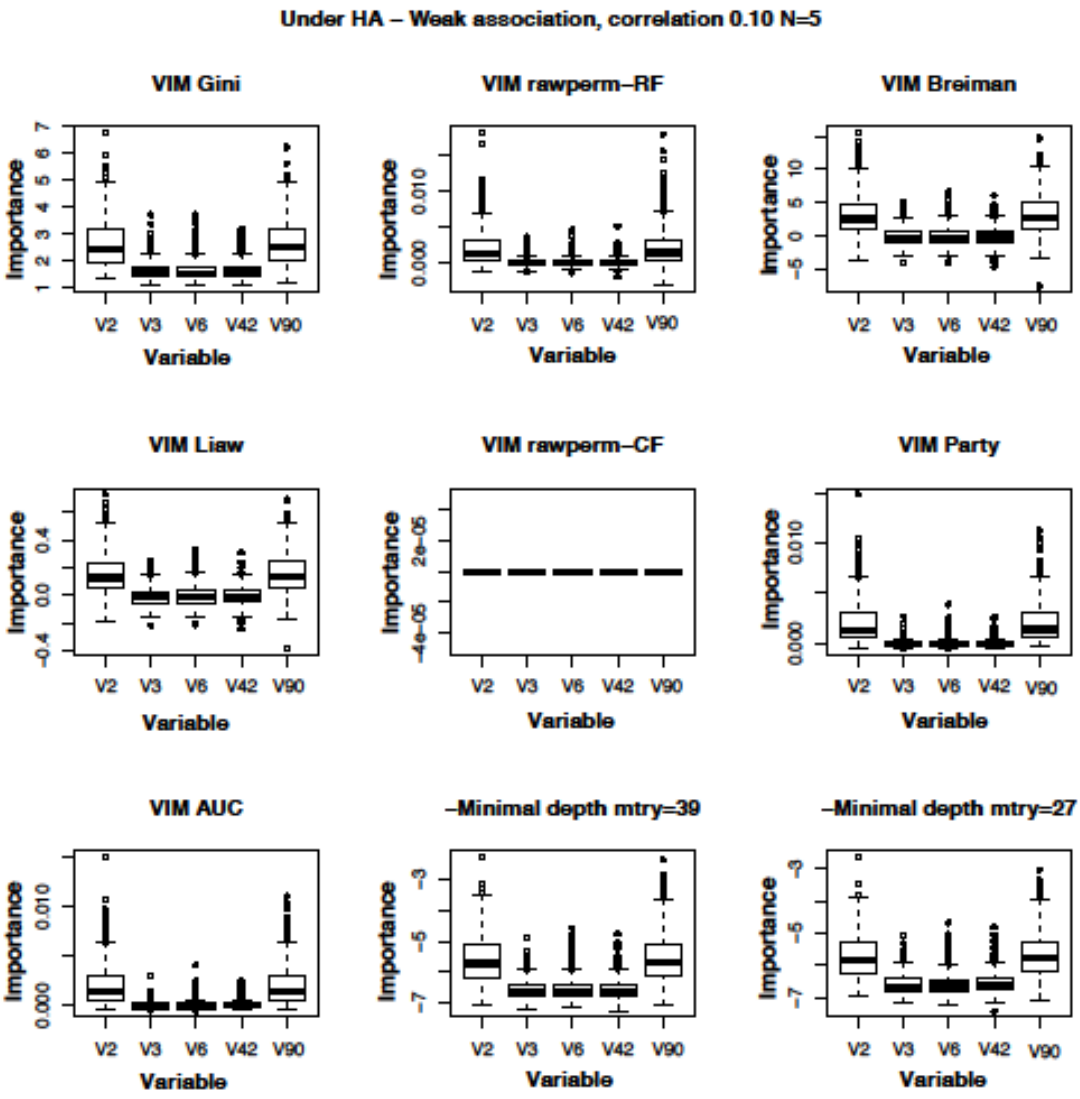

Figure 2.12. RF VIMs under HA in the weakly-associated interaction study for V_2 and V_{90} (interacting variables), two variable correlated V_3 and V_6, and one independent variable V_{42}, when r = 0.10 and N = 5.

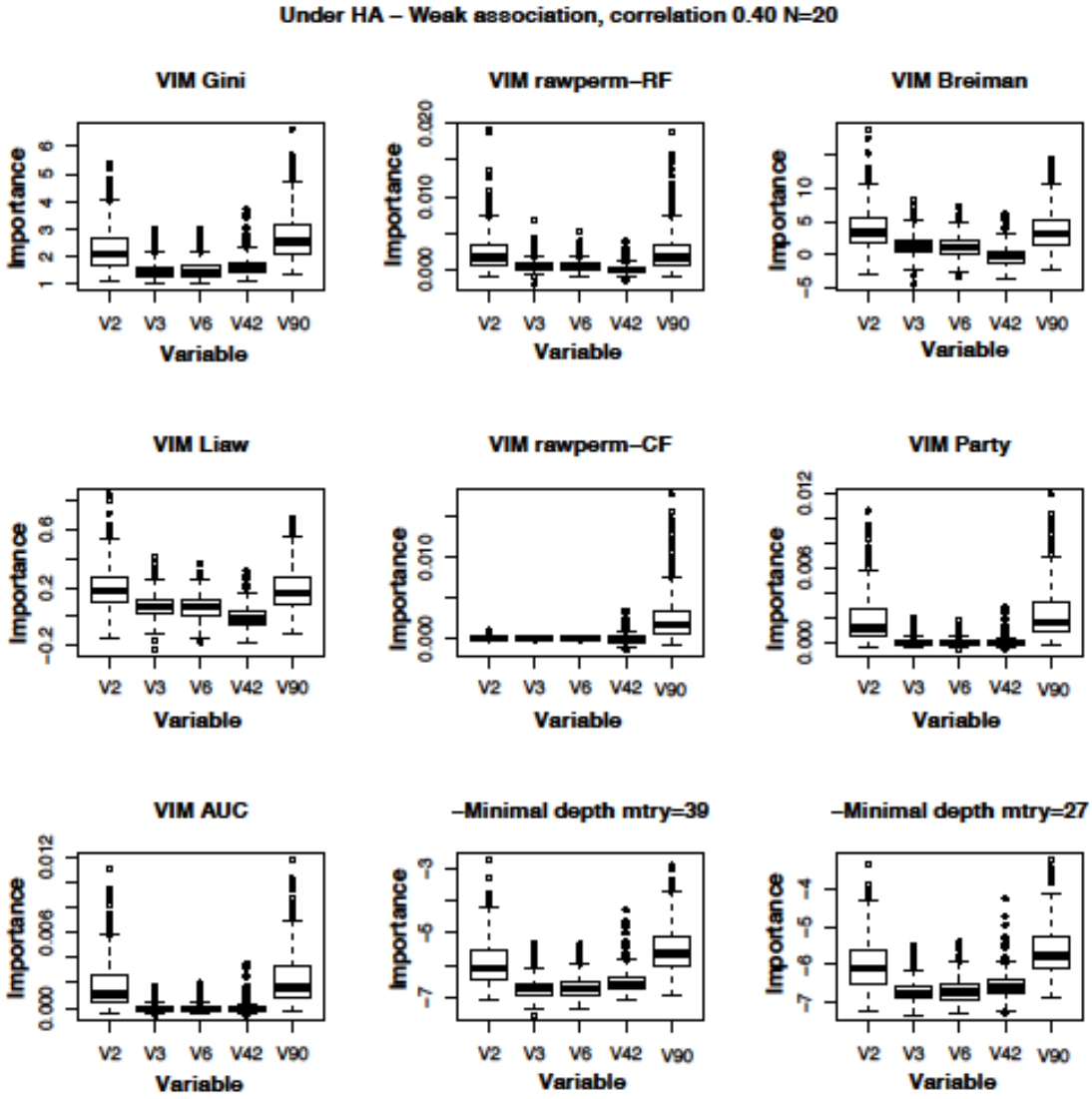

Figure 2.13. RF VIMs under HA in the weakly-associated interaction study for V_2 and V_{90} (interacting variables), two variable correlated V_3 and V_6, and one independent variable V_{42}, when r = 0.40 and N = 20.

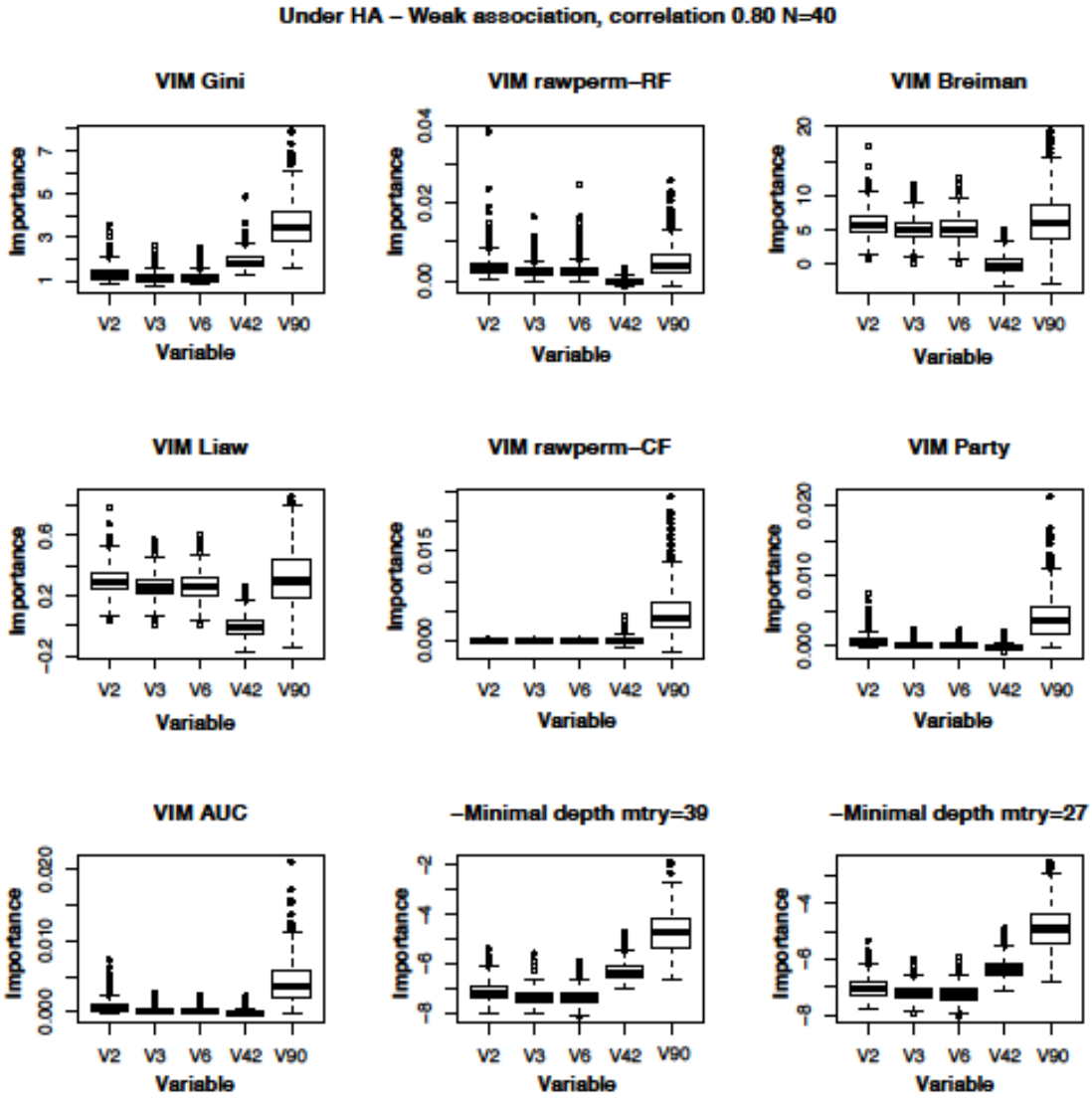

Figure 2.14. RF VIMs under HA in the weakly-associated interaction study for V_2 and V_{90} (interacting variables), two variable correlated V_3 and V_6, and one independent variable V_{42} and V_{90} when r = 0.80 and N = 40.

2.3.5. Conditional PVIM with different correlation cut-off

$VIM_{rawperm\text{-}CF}$ 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, $VIM_{rawperm\text{-}CF}$ 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 $VIM_{rawperm\text{-}CF}$ 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 $VIM_{rawperm\text{-}RF}$ 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

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

V2 weak	**r = 0.80**			**r = 0.40**			**r = 0.10**		
N	5	20	40	5	20	40	5	20	40
K=0.05	0	0	0	0	0	0	0	0	0
K=0.35	13.99	0.0056	0	45.21	0.0012	0	66.22	72.06	73.61
K=0.75	13.25	0	0	74.62	66.58	61.93	66.28	71.92	73.07

Table 2.14. Power of $VIM_{rawperm\text{-}CF}$ in detecting V_2 under all correlation conditions with the three different cut-offs in the weak single association study.

V2 strong	**r = 0.80**			**r = 0.40**			**r = 0.10**		
N	5	20	40	5	20	40	5	20	40
K=0.05	0	0	0	0	0	0	0	0	0
K=0.35	100	49.71	0.0028	100	32.46	5.03	100	100	100
K=0.75	100	53.75	0.0344	100	100	100	100	100	100

Table 2.15. Power of $VIM_{rawperm\text{-}CF}$ in detecting V_2 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 V_2 when the correlation was medium, but in the extreme situation still did not show larger scores for V_2 because of a large number of correlated variables.

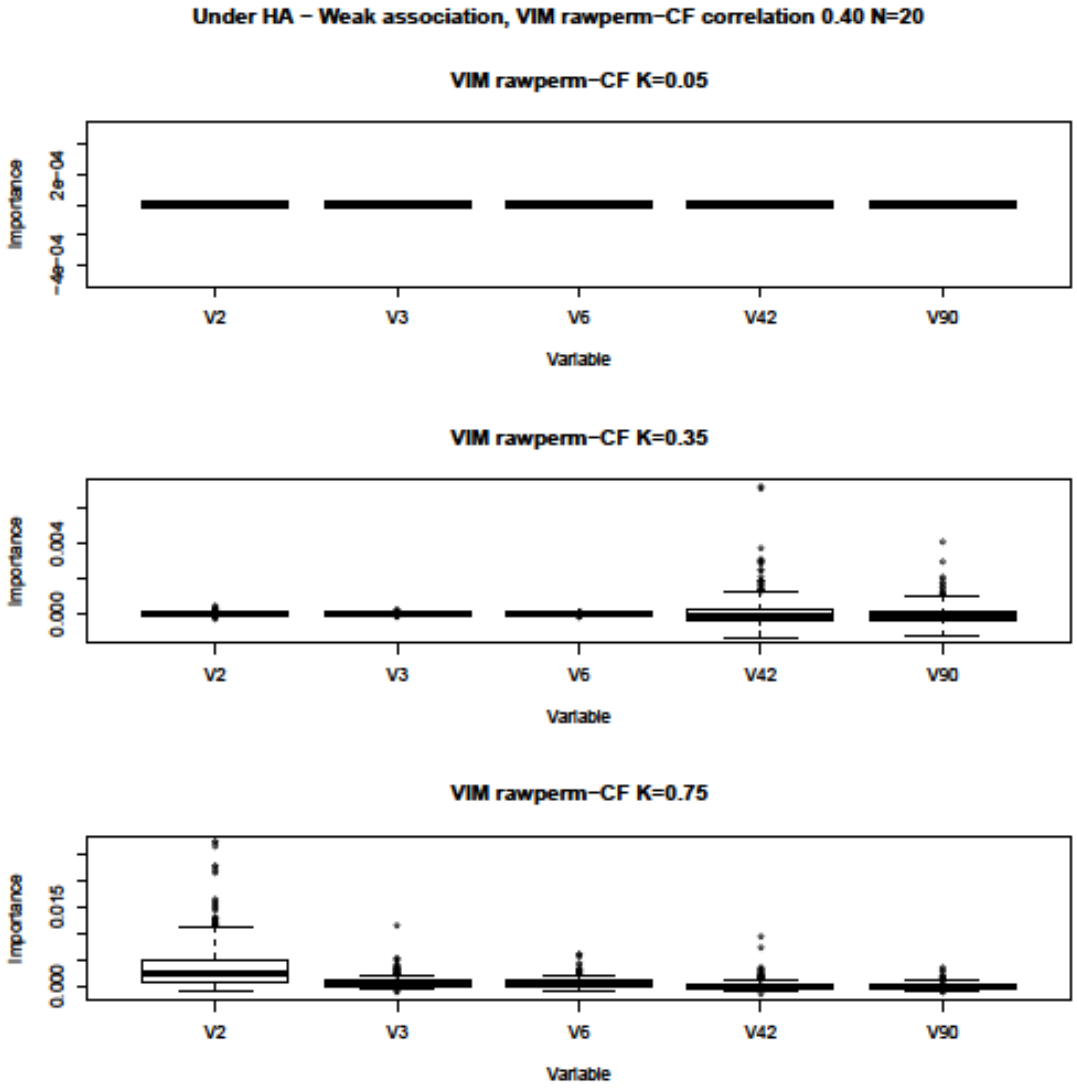

Figure 2.15. $VIM_{rawperm\text{-}CF}$ under HA in the weakly-associated single study for V_2, two variable correlated V_3 and V_6, and one independent variable V_{42} and V_{90} when r = 0.40 and N = 20.

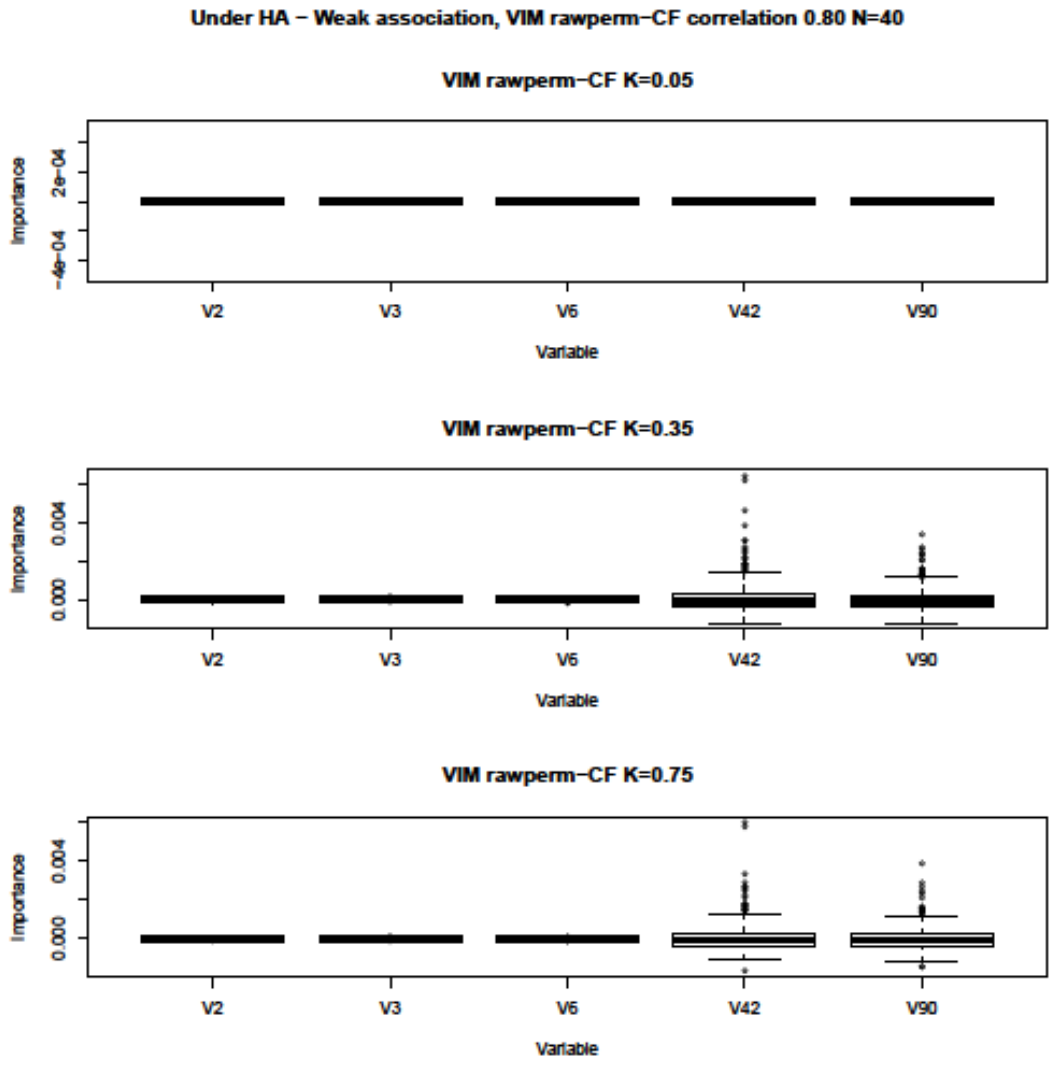

Figure 2.16. $VIM_{rawperm-CF}$ under HA in the weakly-associated single study for V_2, two variable correlated V_3 and V_6, and one independent variable V_{42} and V_{90} when r = 0.80 and N = 40.

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,

$VIM_{rawperm\text{-}CF}$ had the same behavior as $VIM_{rawperm\text{-}RF}$ 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 $VIM_{rawperm\text{-}CF}$ 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) $VIM_{rawperm\text{-}CF}$ 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.

V2 weak	**r = 0.80**			**r = 0.40**			**r = 0.10**		
N	5	20	40	5	20	40	5	20	40
K=0.05	0	0	0	0	0	0	0	0	0
K=0.35	8.13	0.08	0	31.20	0.16	0	56.42	59.39	60.39
K=0.75	13.25	0.02	0	74.62	66.58	61.93	66.28	71.92	73.07

Table 2.16. Power of $VIM_{rawperm\text{-}CF}$ in detecting V_2 under all correlation conditions with the three different cut-offs in the weakly-associated interaction study.

V90 weak	**r = 0.80**			**r = 0.40**			**r = 0.10**		
N	5	20	40	5	20	40	5	20	40
K=0.05	0	0	0	0	0	0	0	0	0
K=0.35	66.86	81.83	90.64	59.75	67.57	77.36	61.54	57.26	62.77
K=0.75	67.05	81.34	89.74	57.53	59.57	52.64	61.55	57.15	63.03

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

V2 strong	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
K=0.05	0	0	0	0	0	0	0	0	0
K=0.35	100	15.66	0.02	100	20.61	2.07	100	100	100
K=0.75	100	25.64	0.44	100	100	100	100	100	100

Table 2.18. Power of $VIM_{rawperm\text{-}CF}$ in detecting V_2 under all correlation conditions with the three different cut-offs in the strong interaction association study.

V90 weak	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
K=0.05	0	0	0	0	0	0	0	0	0
K=0.35	100	100	100	100	100	100	100	100	100
K=0.75	100	100	100	100	100	100	100	100	100

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

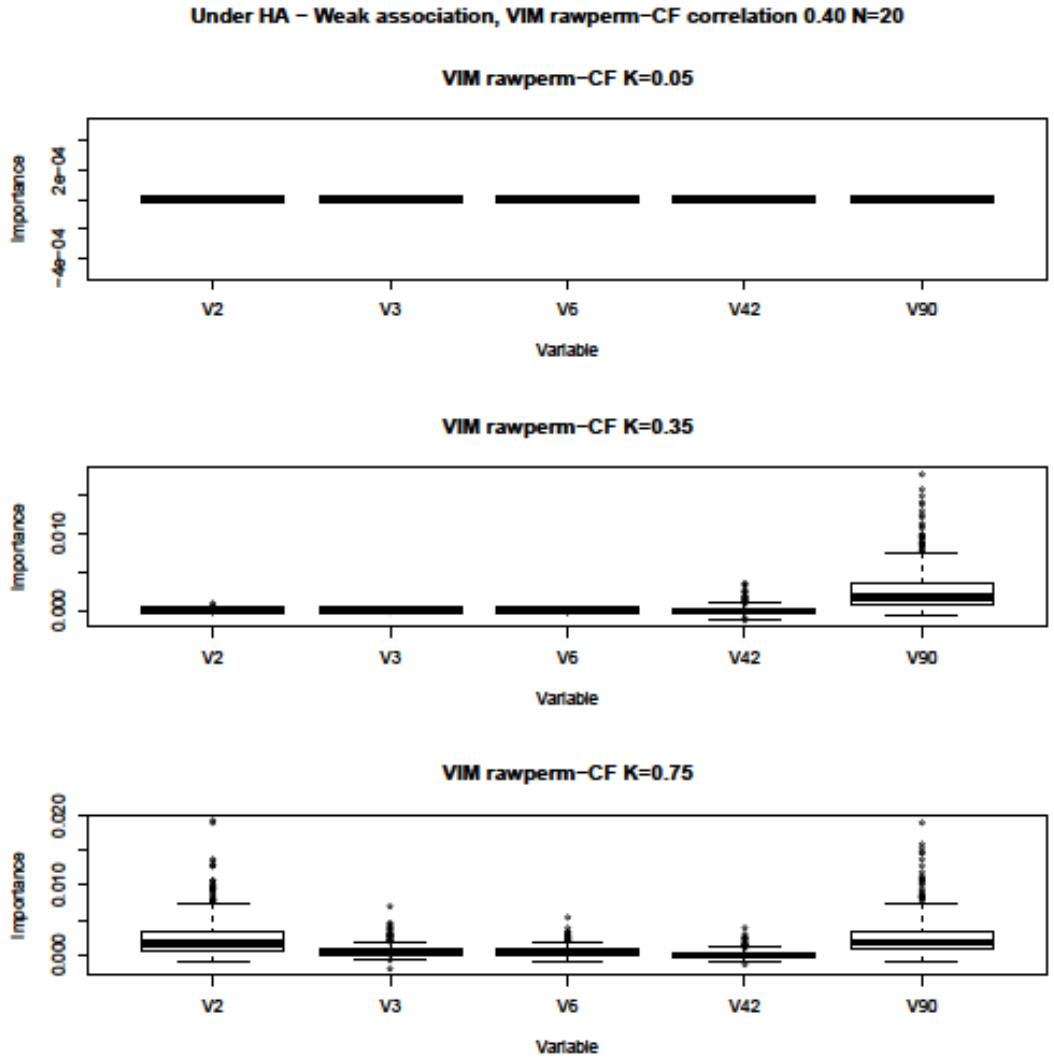

Figure 2.17. $VIM_{rawperm-CF}$ under HA in the weakly-associated interaction study for V_2, two variable correlated V_3 and V_6, and one independent variable V_{42} and V_{90} when r = 0.40 and N = 20.

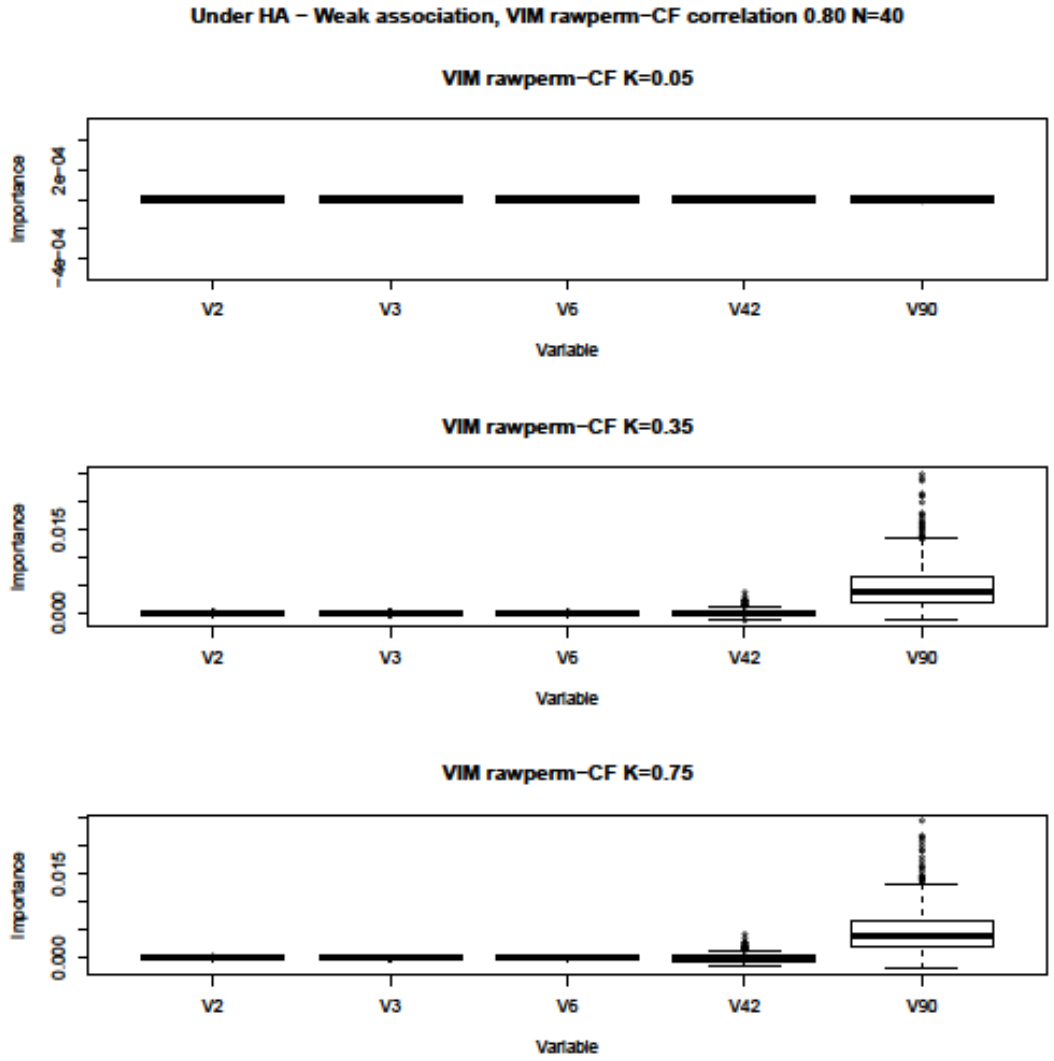

Figure 2.18. $VIM_{rawperm-CF}$ under HA in the weakly-associated interaction study for V_2, two variable correlated V_3 and V_6, and one independent variable V_{42} and V_{90} when r = 0.80 and N = 40.

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_{Gini} under predictor correlation, as has been shown in previous studies (Nicodemus and Malley 2009); (Nicodemus 2011). Under H_0, VIM_{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

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_{Gini} 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

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 $VIM_{rawperm-CF}$ (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, $VIM_{rawperm-CF}$ 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, $VIM_{rawperm-CF}$ performs the same as $VIM_{rawperm-RF}$ 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 $VIM_{rawperm-CF}$. 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, $VIM_{rawperm-CF}$ 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 $VIM_{rawperm-RF}$ 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 VIM_{AUC} and VIM_{party}, which are based on CIF, may be better to apply in

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_{Gini}, PVIMs from RF, VIM_{AUC}, VIM_{party} and minimal depth improve their ability to detect the uncorrelated associated predictor if the correlation among other predictors is higher. VIM_{AUC} and VIM_{party} showed the highest power for capturing V_{90} in the extreme correlation condition, followed by $VIM_{rawperm\text{-}CF}$.

Wright *et al.* (2016) previously found that VIM_{Gini} gives higher importance to both interacting predictors than the unconditional unscaled PVIM. The reason of the contradictory behaviour of VIM_{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_{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_{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_{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

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}, $VIM_{rawperm\text{-}RF}$ under correlation conditions. The time consumed applying $VIM_{rawperm\text{-}RF}$ is significantly lower than performing VIM_{AUC} or VIM_{party}. Although $VIM_{rawperm\text{-}CF}$ 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 $VIM_{rawperm\text{-}RF}$; $VIM_{rawperm\text{-}RF}$ was chosen due to the computational constraints (mostly timing).

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.

TARGET DATASETS	TARGET SOURCE	NCASES	NCONTROLS	NTOTAL
ABER	UK (Aberdeen)	719	697	1416
AJSZ	Israel (Lencz/Darvasi Sample)	894	1594	2488
ASRB	Australia	456	287	743
BULS	Bulgaria (case control)	195	608	803
BUTR	Bulgaria (trios)	608	613	1221
CATI	US (CATIE)	397	203	600
CAWS	UK (Cardiff)	396	284	680

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

UMEB	Sweden (Umeå)	341	577	918
UMES	Sweden (Umeå)	193	704	897
ZHH1	US (New York)	190	190	380
TOTAL (ALL STUDIES)	Total (All studies)	**26476**	**31804**	**58280**

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.

MOLECULAR FUNCTION	**ANALYTE**	**GENES**
LIPID TRANSPORT	Apolipoprotein H	APOH
	Apolipoprotein A1	APOA1
INFLAMMATORY RESPONSE	Macrophage migration inhibitory factor	MIF
	Carcinoembryonic antigen	CEACAM5
	Tenascin C	TNC
	Interleukin-10	IL10
	Interleukin-1 receptor antagonist	IL1RN
	Receptor for advanced glycosylation end products	AGER
	Interleukin-8	CXCL8

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

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

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.

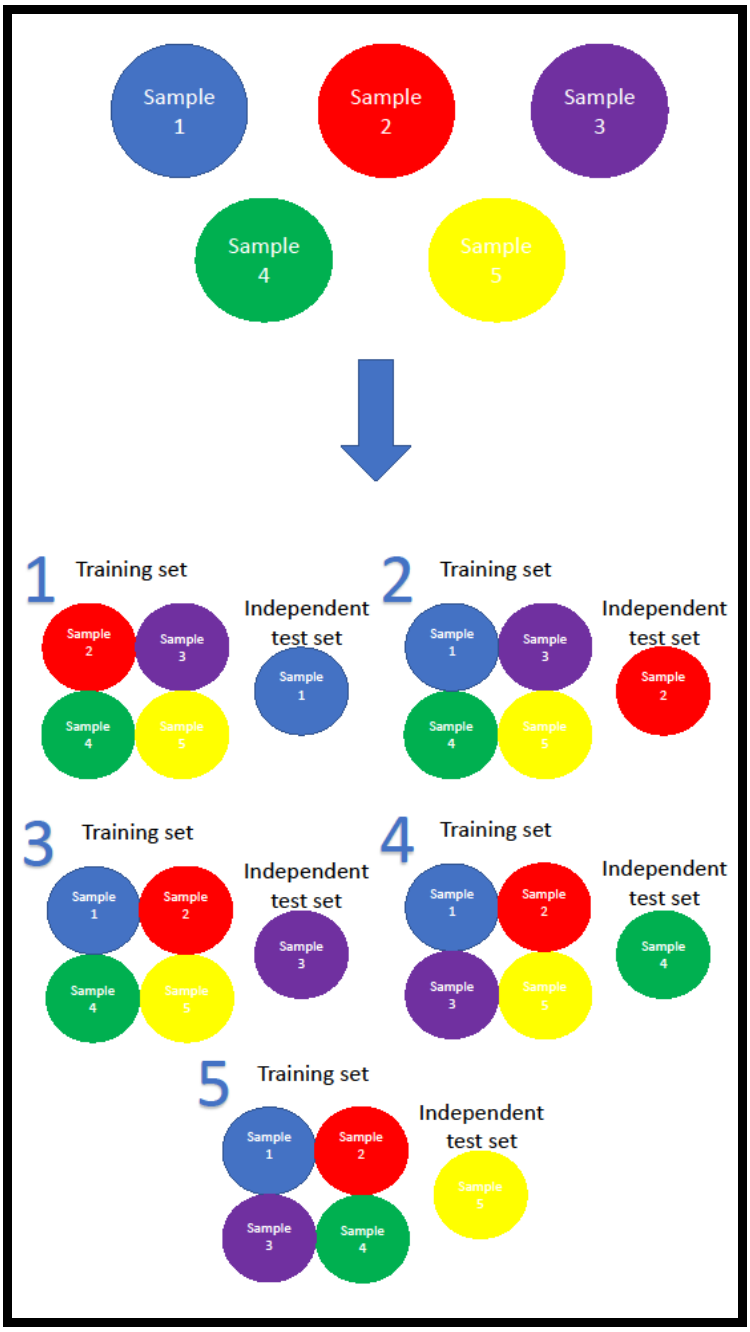

Figure 2.19. Example of leave-one-out cross validation in a study design with 5 cohorts, instead with 39.

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

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: $Y \sim \beta 1\ SNPi + \beta 2\ SNPj + \beta 3\ SNPi * SNPj$

Reduced Model: $Y \sim \beta 1\ SNPi + \beta 2\ SNPj$

Where

SNPi , 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).

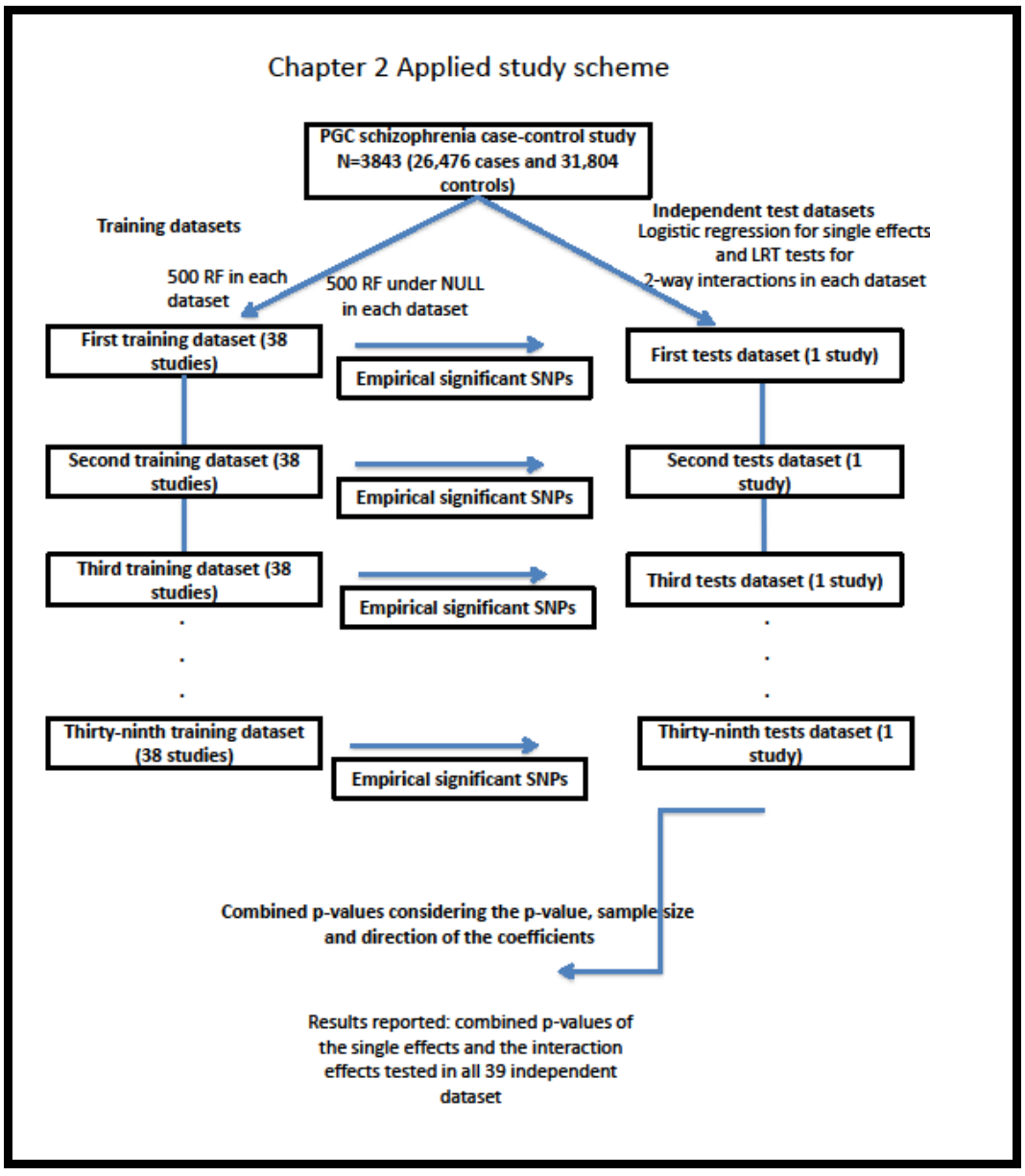

Figure 2.20. Illustration of the approach taken on the applied study in section 2.5. This illustration shows the methods taken in the study and the studies which were combined to report the final results (only the effects which were tested in all independent datasets were combined).

2.5.7. Results

After calculating the empirical *p*-values over all RF iterations based on $VIM_{rawperm\text{-}RF}$ 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.

SNP	CHR	GENE	P-VALUE	%R^2
rs3798211	6	*ACAT2*	2.62 x 10^{-5}	[0.0023%,0.7%]
rs3789875	9	*TNC*	0.0004	[0.0001%,2.5%]

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

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 (Brellier 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

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 considered 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).

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_{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 *et al.* 2011) performs VIM_{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_{Gini} has been previously shown to be biased. VIM_{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 *et al.* 2007b); (Nicodemus and Malley 2009); (Nicodemus 2011); (Boulesteix *et al.* 2012a)).

With regard to the second bias, Strobl *et al* (2007b) showed that VIM_{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_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

major allele), VIM_{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 VIM_{Gini} under H_0. Under H_A, SNPs with larger MAF showed higher VIM_{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 VIM_{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 VIM_{Gini} when all predictors followed a normal distribution (all continuous) but with different variances, and were also independent of each other. If VIM_{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 VIM_{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 VIM_{Gini}. So, it may happen that more noise has an impact on the VIM_{Gini} and, therefore, inflates the importance scores due to noise rather than association. In real situations, the error

(noise) cannot be distinguished or quantified, but it is still important to investigate whether it has an impact on RF based on VIM_{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_{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_{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) = 1000x10, $\sum_1$= I (identity matrix) (corr$(x_j, x_k)_{j \neq k}$=0).

$$y = e_1$$

where $e_1 \sim N(0,1)$ (annotation N(μ,σ)).

3.2.1.1.1.2. 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 ($\beta = 0$):

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

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 $Z = [z_{ij}] \sim N(0, \sum_2)$ where dim(Z) = 1000x10, $\sum_2$ = diag(50,45,40,35,30,25,20,15,10,1) (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(μ,σ)).

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

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 $X = [x_{ij}] \sim N(0, \sum_1)$ where dim(X) = 1000x10, $\sum_1$ = I (identity matrix) (corr(x_j, x_k)$_{j \neq k}$=0)

$$y_j = 0.3 * x_j + e_1$$

where $e_1 \sim N(0,1)$ (annotation N(μ,σ)); 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_{Ginin} 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:

$$y_{bin_j} = \begin{cases} 1 \; when \; 0.3 * x_j + e_1 > 0 \\ 0 \qquad Otherwise \end{cases}$$

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 $Z = [z_{ij}] \sim N(0, \sum_2)$ where dim(Z) = 1000x10, $\sum_2$ = diag(50,45,40,35,30,25,20,15,10,1) (corr(z_j, z_k)$_{j \neq k}$ = 0)

$$y_j = 0.3 * z_j + e_1$$

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_{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_{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

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 $X = [x_{ij}] \sim N(0, \sum_1)$ where dim(X) = 1000x10, $\sum_1$= I (identity matrix) (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 3.2.1.1.1.2.:

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

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.

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 $X = [x_{ij}] \sim N(0, \sum_1)$ where dim(X) = 1000x10, $\sum_1$ = I (identity matrix) (corr(x_j, x_k)$_{j \neq k}$=0)

$$y_j = 0.3 * x_j + e_1$$

where $e_1 \sim N(0,1)$; x_j is one of the ten predictors, and $j = 1, \dots, 10$. 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:

$$y_{bin_j} = \begin{cases} 1 \; when \; 0.3 * x_j + e_1 > 0 \\ 0 \qquad Otherwise \end{cases}$$

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

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 VIM_{Gini} when there is no association. In the subsections above, I simulated different

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 $X = [x_{ij}] \sim N(0, \sum_1)$ where dim(X) = 1000x10, $\sum_1$= I (identity matrix) (corr$(x_j, x_k)_{j \neq k}$=0)

$$y = e_2$$

where $e_2 \sim N(0,0.5)$ (annotation N(μ,σ)).

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:

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

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.

3.2.3.2.1. All variables follow a standard normal distribution

3.2.3.2.1.1. Continuous outcome

Let $X = [x_{ij}] \sim N(0, \sum_1)$ where dim(X) = 1000x10, $\sum_1$ = I (identity matrix) (corr(x_j, x_k)$_{j \neq k}$=0)

$$y_j = 0.3 * x_j + e_2$$

where $e_2 \sim N(0,0.5)$ (annotation N(μ,σ)); 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

$$y_{bin_j} = \begin{cases} 1 \; when \; 0.3 * x_j + e_2 > 0 \\ 0 \qquad Otherwise \end{cases}$$

where $e_2 \sim N(0,0.5)$ (annotation N(μ,σ)); all predictors x_j and the non-associated ones followed a standard normal distribution.

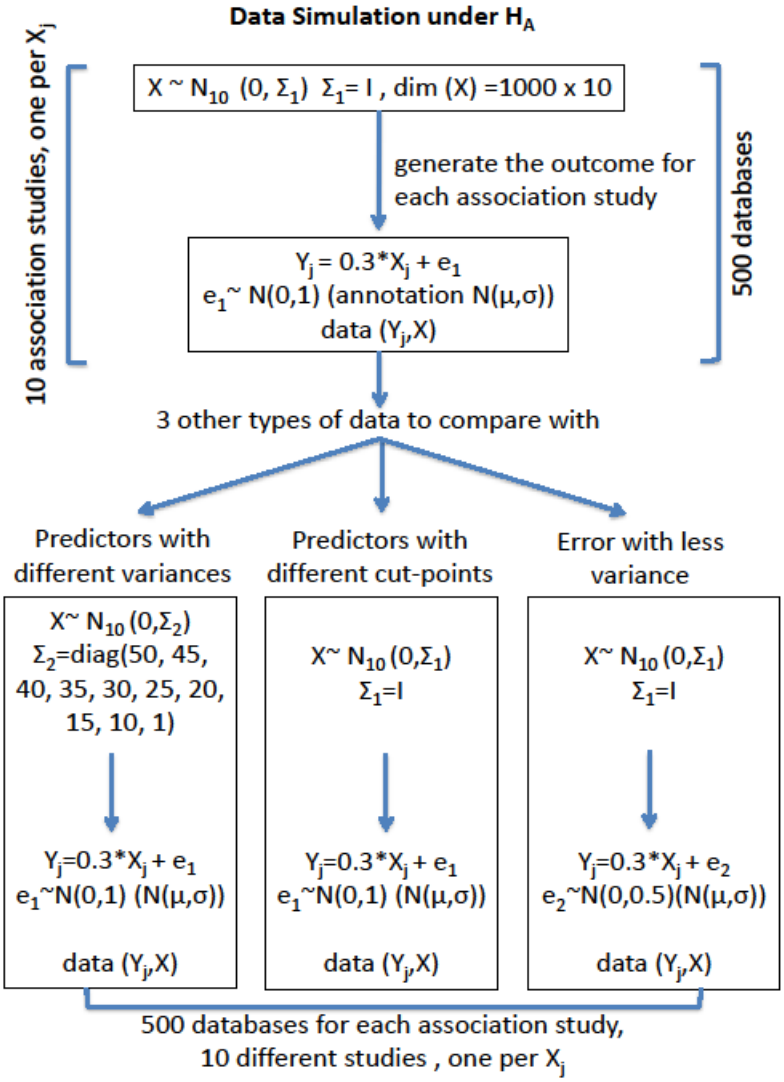

Figure 3.1. Illustration of the data generation under H_A in all different conditions when the outcome is continuous. The top one corresponds to when all predictors and the error follow a standard normal distribution. The three bottom conditions are, from left to right, when predictors follow a standard 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.

CONTINUOUS AND BINARY OUTCOME	UNDER H_0	UNDER H_A
X_{10} ~ N(0,I) , e_1 ~ 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, … 500 when X_{10} is associated). **Study of reference**
X_{10} ~ N(0,$\sum_2$) , e_1 ~ 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, … 500 when X_{10} is associated)
X_{10} ~ N(0,I) , e_1 ~ 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, … 500 when X_{10} is associated)
X_{10} ~ N(0,I) , e_1 ~ 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, … 500 when X_{10} is associated)

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.

3.2.4. Random Forest based on VIM_{Gini} simulation.

For details of RF method and how VIM_{Gini} is defined see the Methods sections of Chapter 2. The performance of RF based on VIM_{Gini} was investigated under the H_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_{Gini} in the absence of association under the null hypothesis (H_0) for both outcomes. Under both H_0 and H_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_{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_{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 *ntree* = 1000, the *mtry* (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 mtry 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 (mtry value is an integer).

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 (μ) = 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 ($\sum$=diag(50,45,40,35,30,25,20,15,10,1), where $\sum$ 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

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).

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

Table 3.2. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under H_0, when all predictors and the error followed a standard normal distribution. Continuous outcome.

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

Table 3.3. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under H_0, when all predictors followed a normal distribution but with different amounts of variance. The error followed a standard normal distribution. Continuous outcome.

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

Table 3.4. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under 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.

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

Table 3.5. Bias, coverage, number of p-values less than 0.05 and less than Bonferroni correction threshold (0.0001) under H_0, when all predictors followed a standard normal distribution. The error followed a normal distribution with 0.5 standard deviation. Continuous outcome

In addition, under H_A, 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 VIM_{Gini} 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

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.

N(0,1)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
BIAS	0.001	-0.002	0.0008	0.0001	-0.0021	-0.001	-0.003	-0.0001	-0.0054	-0.0024
%COVER	95.8	94.6	95.4	96	95.2	95.2	96	95.8	92.4	94.2

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

Diff variance	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
BIAS	-0.0001	-0.0003	0.0001	-0.00001	0.0004	-0.0001	-0.0005	-0.0001	-0.0007	0.0013
%COVER	94.8	96.6	95.4	95.4	95.6	93.8	96	93.4	95.8	95.2

Table 3.7. 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. Continuous outcome. All models were significant before and after correction.

cutpoints	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
BIAS	0.0014	0.0026	-0.0001	0.0024	-0.0005	-0.0005	0.001	0.0029	-0.0013	0.0026
%COVER	96.2	94.8	94.6	94.6	93.8	97.4	95	94.2	94.8	93.6

Table 3.8. Bias and coverage under 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.

N(0,0.5)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
BIAS	0.0002	-0.0016	0.0001	-0.00003	0.0007	-0.0002	-0.0009	0.0003	0.0007	0.0009
%COVER	94.6	94.6	96	95.6	93.6	93.8	95.4	96.2	96.8	94.6

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.

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 ($\sum$=diag(50,45,40,35,30,25,20,15,10,1), $\sum$ 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.

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

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.

Diff variance	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
BIAS	-0.0002	0.0001	0.0002	-0.0004	0.0002	-0.0006	0.0007	0.00001	0.0004	-0.0004
%COVER	93.8	94.2	95.2	94	95.2	97	95	94.6	95	95
P<0.05	31	29	24	30	24	15	25	27	25	25
P<0.0001	0	0	0	0	0	1	0	0	0	0

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.

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

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.

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

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.

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.

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

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.

Diff Variance	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
BIAS	0.0007	0.0022	0.0015	0.0014	0.0012	0.0016	0.0025	-0.0009	0.0013	0.004
%COVER	95	95.6	95.4	94.6	94.8	95.2	93.2	93.6	94	95.2

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.

cutpoints	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
BIAS	0.0006	0.0062	0.0024	0.0002	-0.0024	-0.0007	0.001	0.0014	0.0027	-0.0029
%COVER	94.6	94.2	94.2	96.4	96.6	94	94.6	93	95.4	96.4

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.

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

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.

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.

3.3.2. VIM_{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_{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_{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 ($\sum$=diag(50,45,40,35,30,25,20,15,10,1), $\sum$ was the variance matrix of uncorrelated predictors). In addition, VIM_{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_{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).

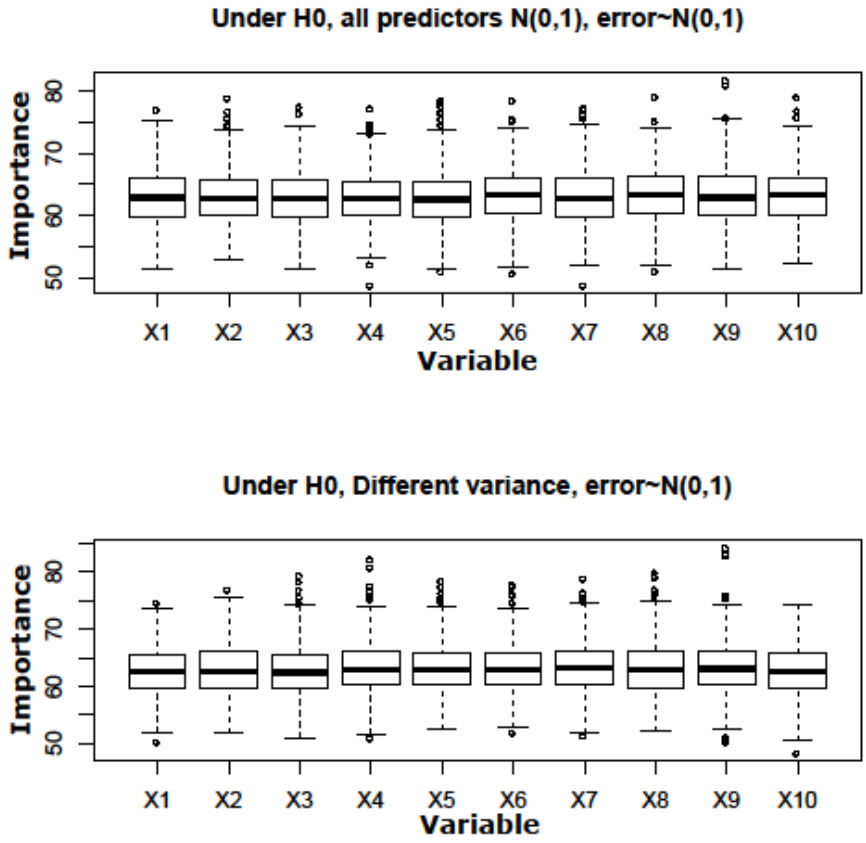

Figure 3.2. VIM_{Gini} under H_0. The top plot illustrates the VIM when all predictors follow a standard normal distribution. The bottom plot shows the VIM when all predictors follow a normal distribution, but each one with a different variance. Continuous outcome.

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.

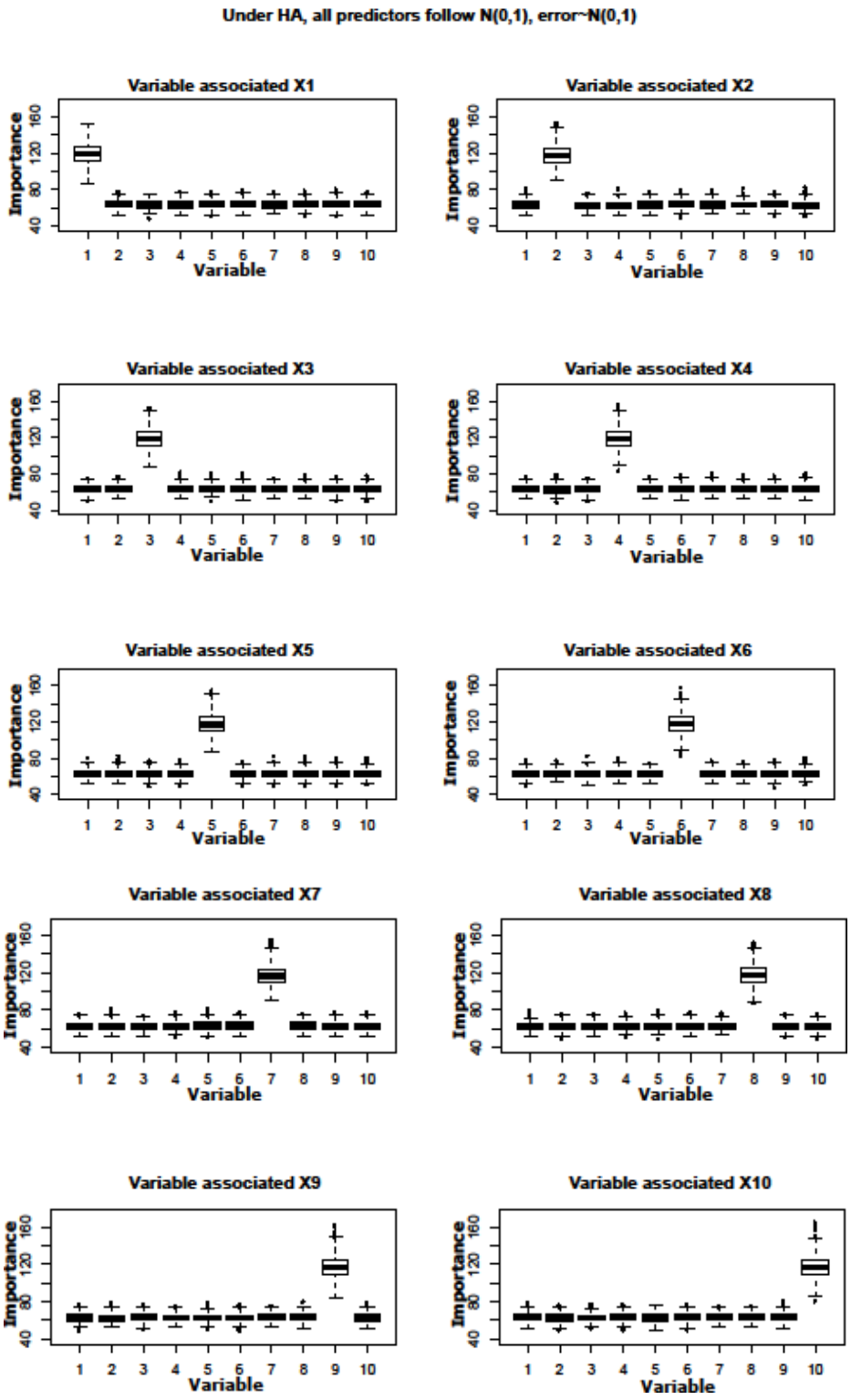

Figure 3.3. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single association models, depending on which variable is associated, when all predictors follow a standard normal distribution. Continuous outcome. Each number i of the X axis corresponds to the subscript of the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

However, if all predictors had different variance ($\sum$ = 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.

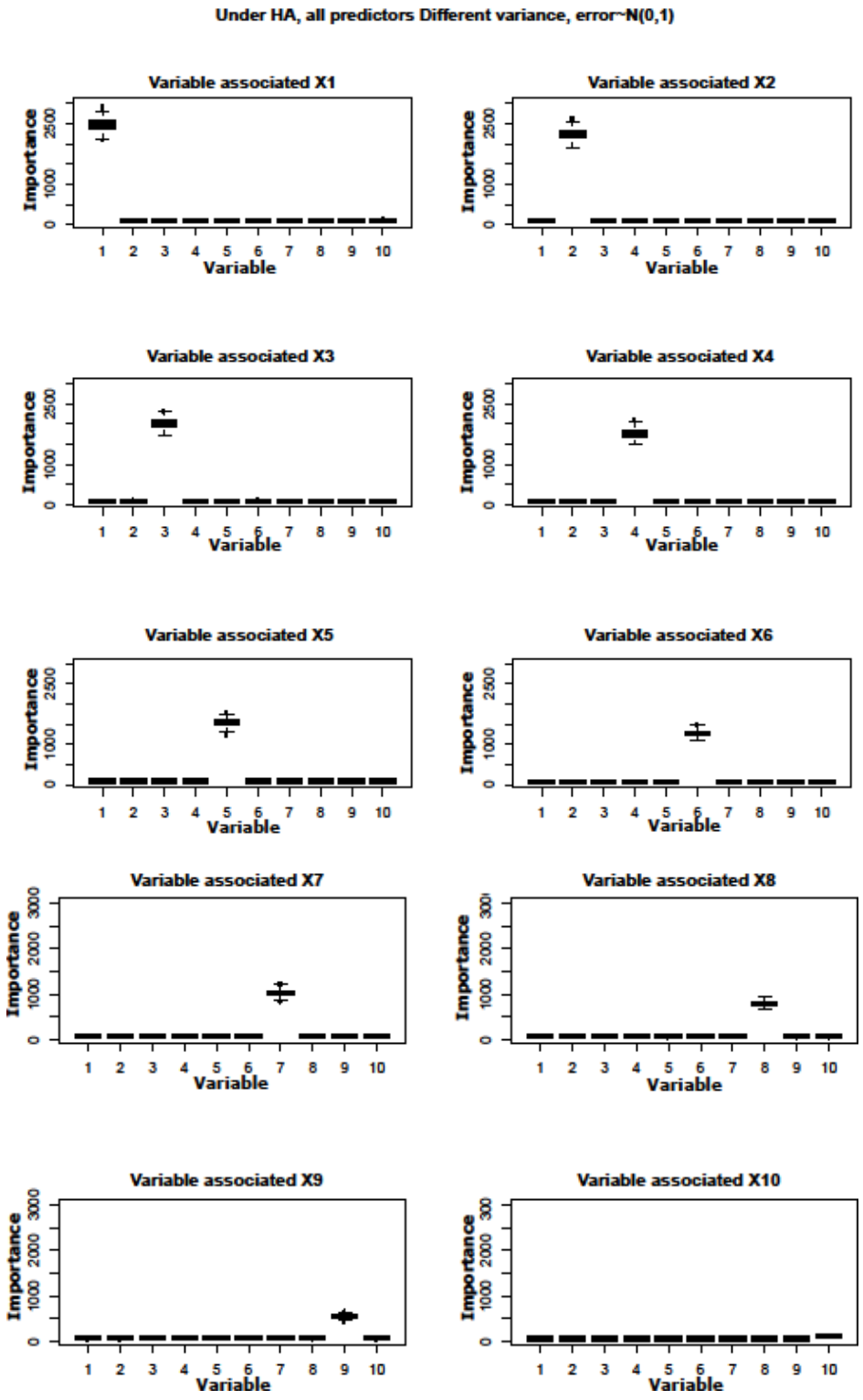

Figure 3.4. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a normal distribution, but each one with different variance. Continuous outcome. Each number i of the X axis corresponds to the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

3.3.2.2. Binary outcome

3.3.2.2.1. Under the null hypothesis

When the outcome was binary, VIM_{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.

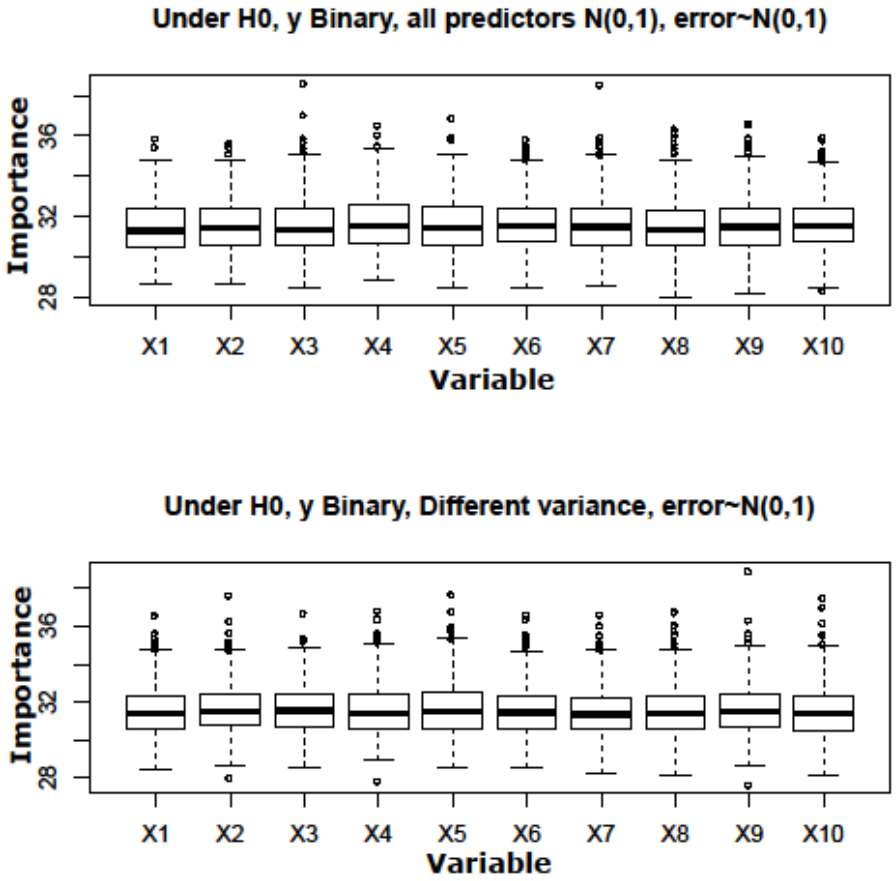

Figure 3.5. VIM_{Gini} under H_0. The top plot illustrates the VIM when all predictors follow a standard normal distribution. The bottom plot shows the VIM when all predictors follow a normal distribution, but each one with a different variance. Binary outcome.

3.3.2.2.2. Under the alternative hypothesis

As when the outcome was continuous, VIM_{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_{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 ($\sum$=diag(50,45,40,35,30,25,20,15,10,1), $\sum$ is the variance matrix of the predictors), VIM_{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_{Gini} inflation (as the measure is based on the Gini index).

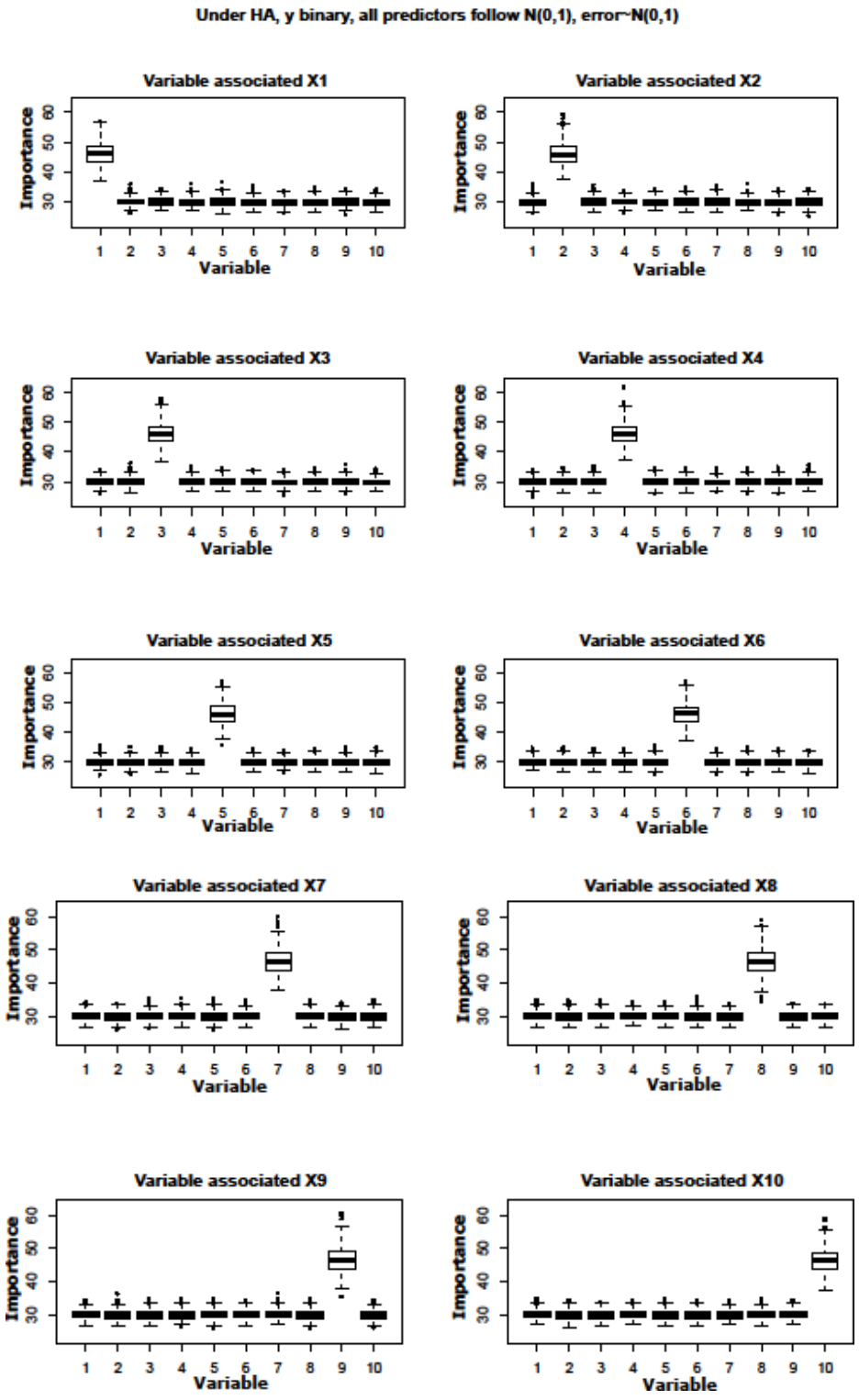

Figure 3.6. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a standard normal distribution. Binary outcome. Each number i of the X axis corresponds to the subscript of the variable X_i.

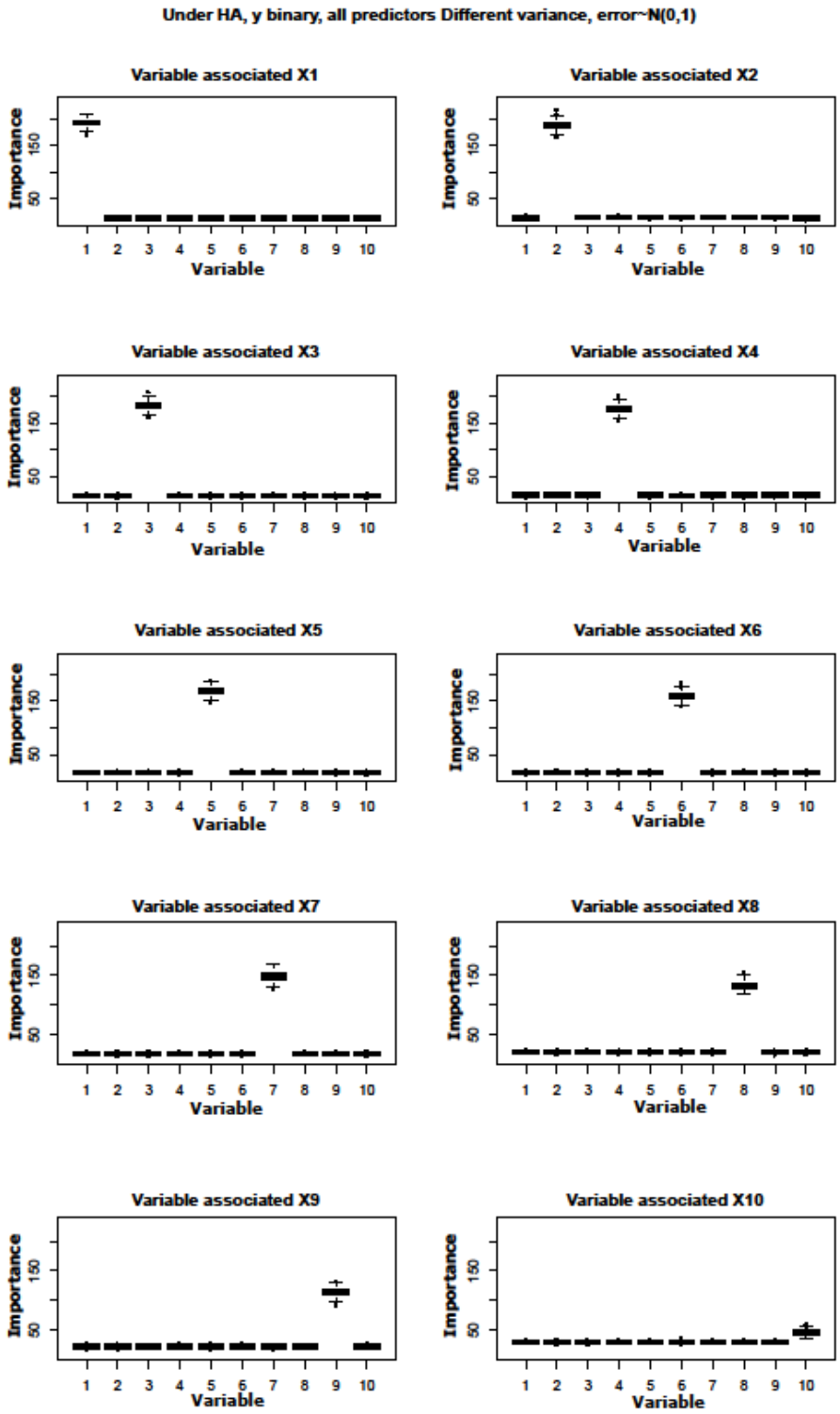

Figure 3.7. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a normal distribution, but with different variances ($\sum$=diag(50,45,40,35,30,25,20,15,10,1), $\sum$ is the variance matrix of the predictors). Binary outcome. Each number i of the X axis corresponds to the subscript of the variable X_i.

3.3.3. VIM_{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.

3.3.3.1. Continuous outcome

3.3.3.1.1. Under the null hypothesis

VIM_{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_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.

cutpoints	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Median	58	373	873	986	999	1000	1000	1000	1000	1000

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.

N(0,1)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Median	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

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.

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_{Gini} did not prefer one predictor over another. However, VIM_{Gini} behave differently if the number of decimal places for each predictor was different. VIM_{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_{Gini} showed an inflation for predictors

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

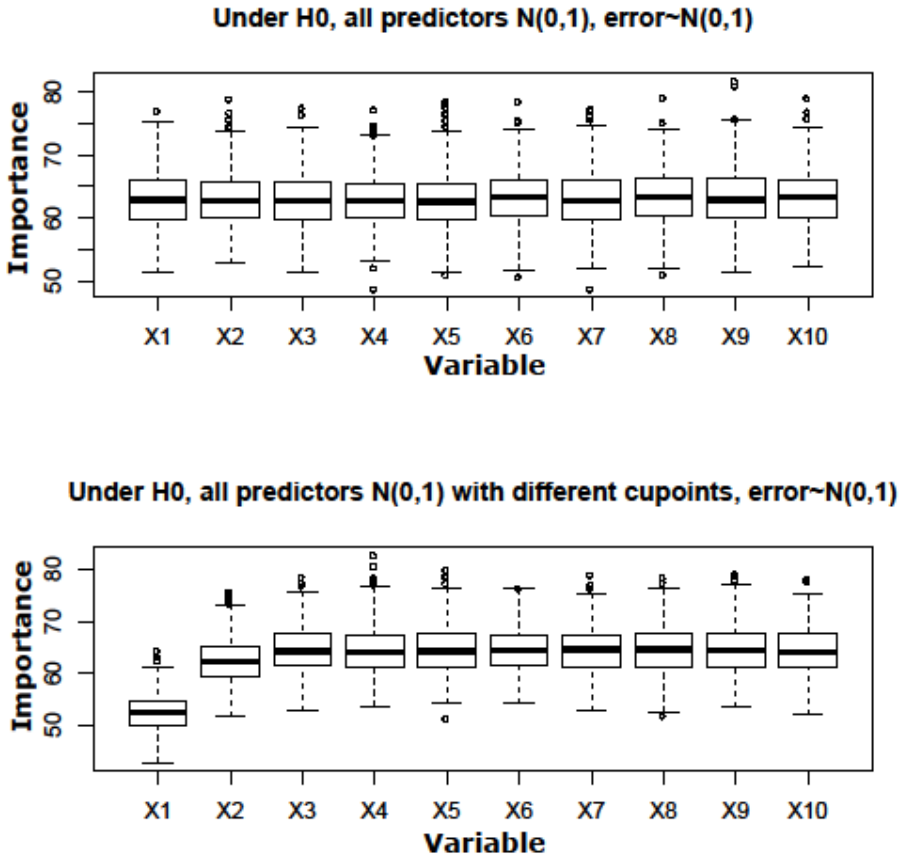

Figure 3.8. VIM_{Gini} under H_0. The top plot illustrates the VIM when all predictors follow a standard normal distribution. The bottom plot shows the VIM when all predictors follow a normal distribution, but each one with different number of decimal places. X1 has one decimal place, X2 has two decimal places, …, X10 has ten decimal places. Continuous outcome.

This bias towards continuous predictors with more cut-points is related to the bias towards variables with more categories found by Strobl *et al.* (2007b), due to the bias of the variable selection in each individual tree because of the Gini index criterion (Boulesteix 2006; Strobl *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

testing case, the predictors with more precision (more tests) are more likely to have a good Gini index value by chance, so the VIM_{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_{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_{Gini} did not show a preference for any of the variables, and all VIM_{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_{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_{Gini} value compared to VIM_{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_{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_{Gini} for influential predictors with more than 3 decimal places was approximately the same

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.

N(0,1)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Median	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

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.

cutpoints	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Median	58	373	871.5	986	999	1000	1000	1000	1000	1000

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.

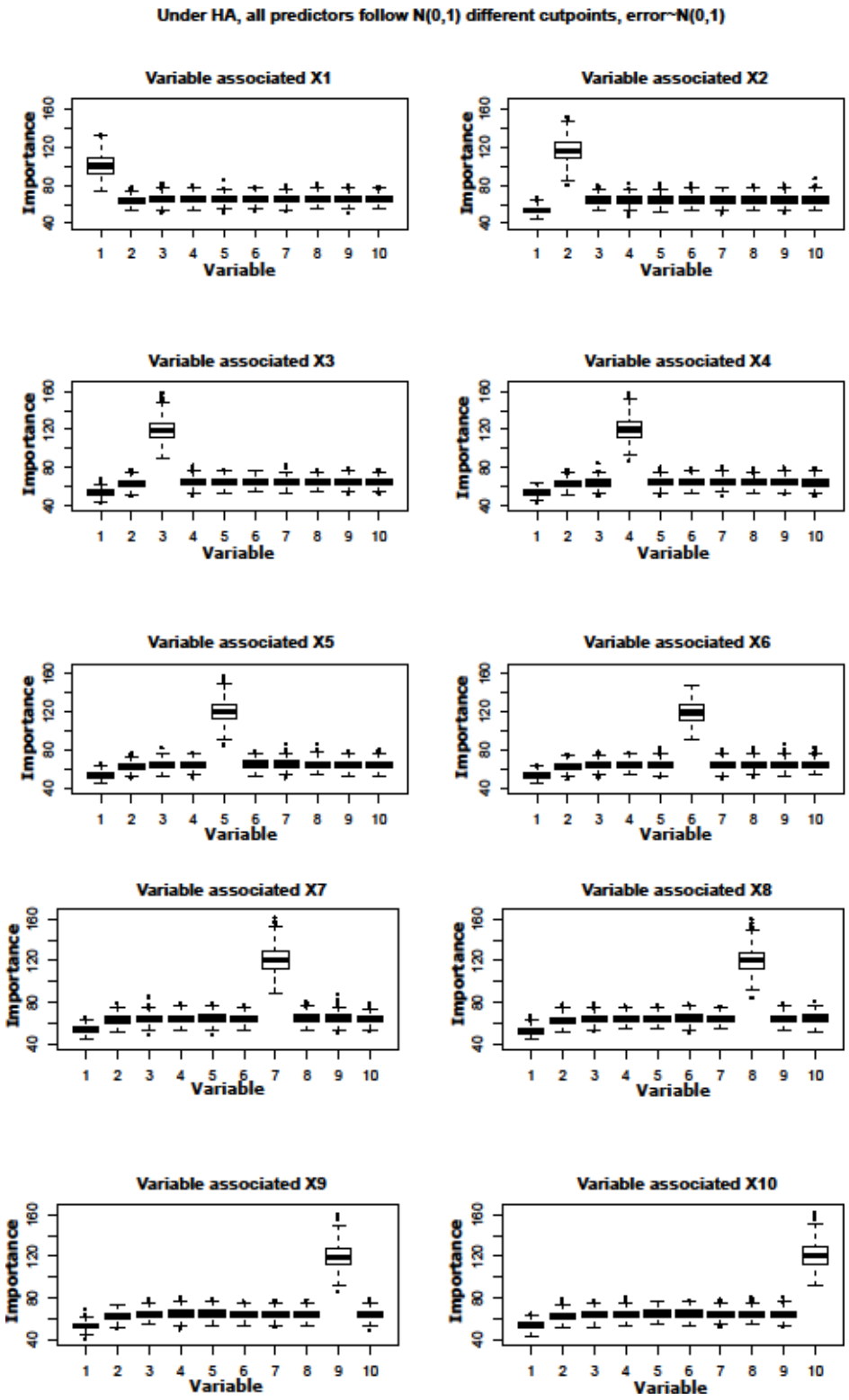

Figure 3.9. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a standard normal distribution, but with different precision. Continuous outcome. Each number i of the X axis corresponds to the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

The fact that VIM_{Gini} showed lower scores for predictors with less than 3 decimal places was because VIM_{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 VIM_{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, VIM_{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 VIM_{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.

N(0,1)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Median	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

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.

cutpoints	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Median	58	373	874	987	999	1000	1000	1000	1000	1000

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.

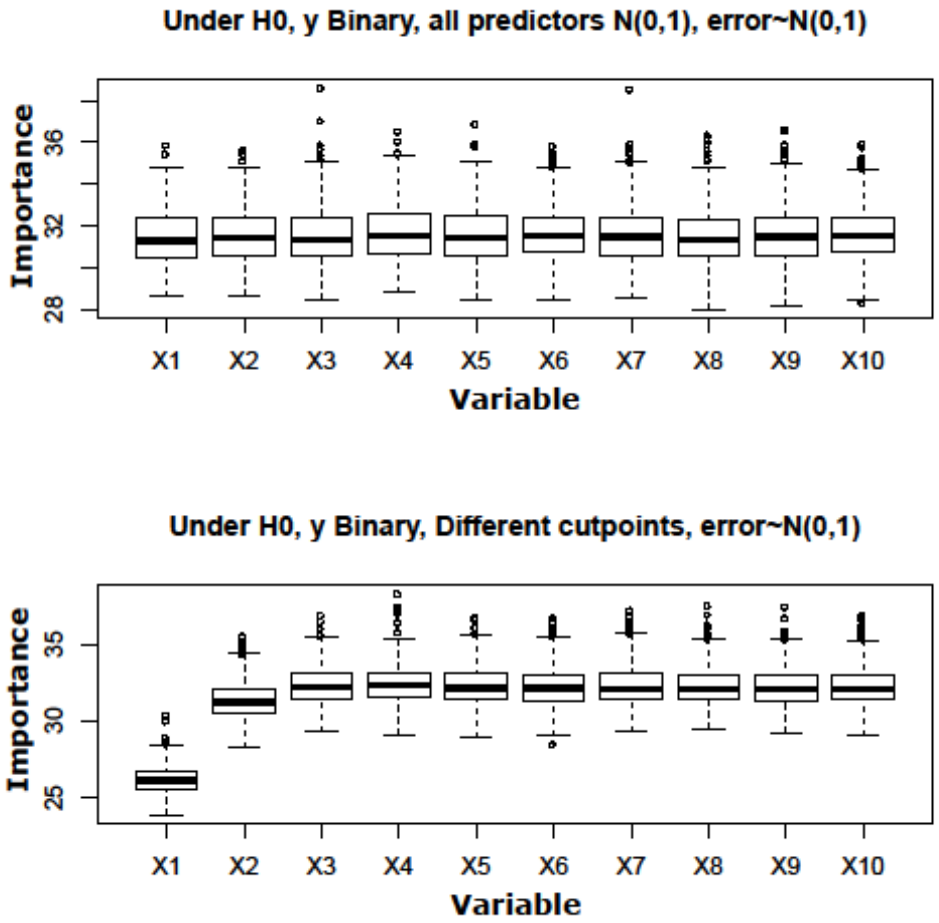

Figure 3.10. VIM_{Gini} under H_0. The top plot illustrates the VIM when all predictors follow a standard normal distribution. The bottom plot shows the VIM when all predictors follow a normal distribution, but each one with different number of decimal places. X1 has one decimal place, X2 has two decimal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

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.

N(0,1)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Median	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

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.

cutpoints	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Median	58	372.5	873	986	999	1000	1000	1000	1000	1000

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.

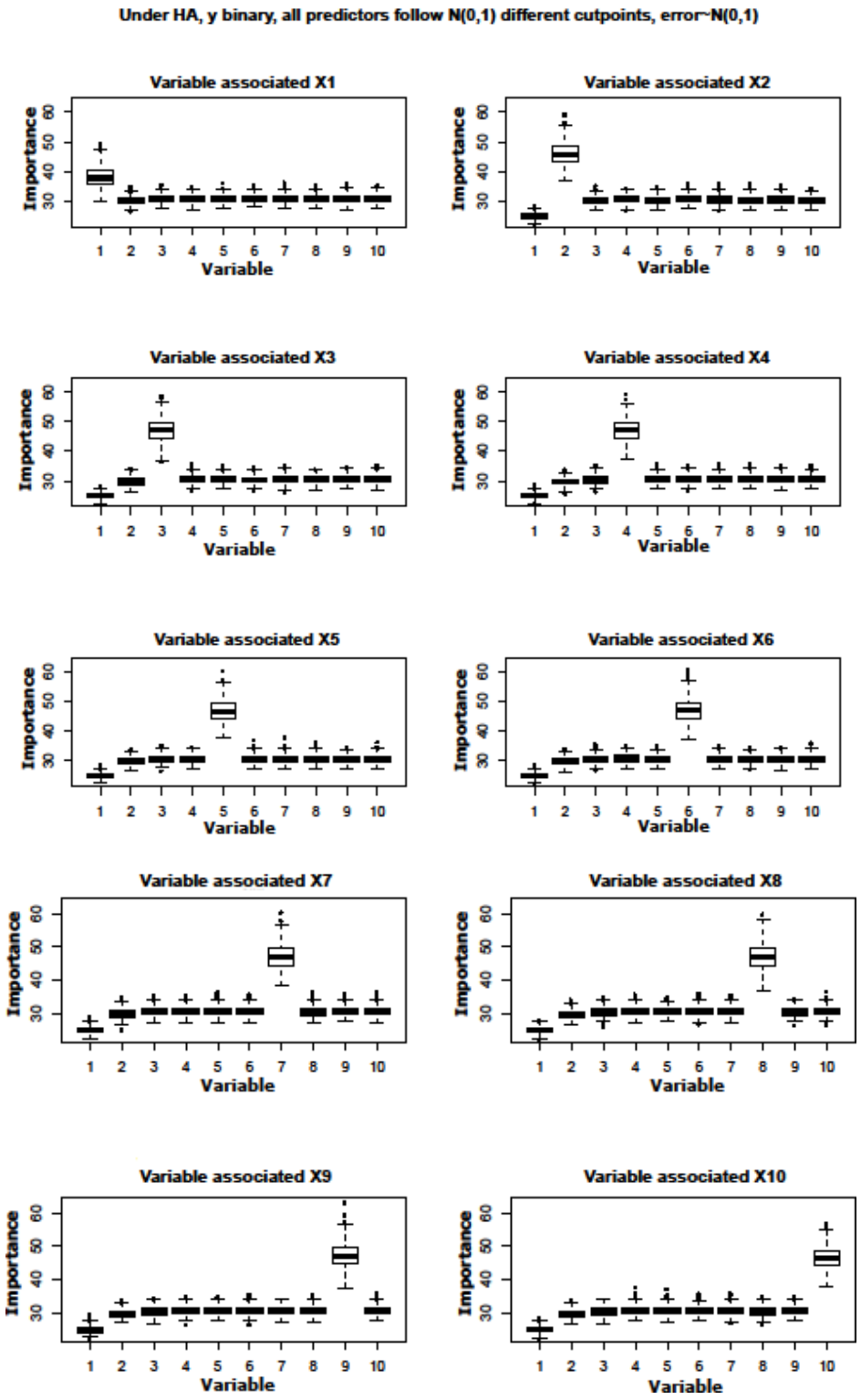

Figure 3.11. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a standard normal distribution, but each one has different number of decimal places. Binary outcome. Each number i of the X axis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

3.3.4. VIM_{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_{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_{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_{Gini} was higher than 60. With decreased variance of the error, VIM_{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_{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_{Gini}, in this case, towards more error variance. It is difficult to know the reason for this inflation, as VIM_{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.

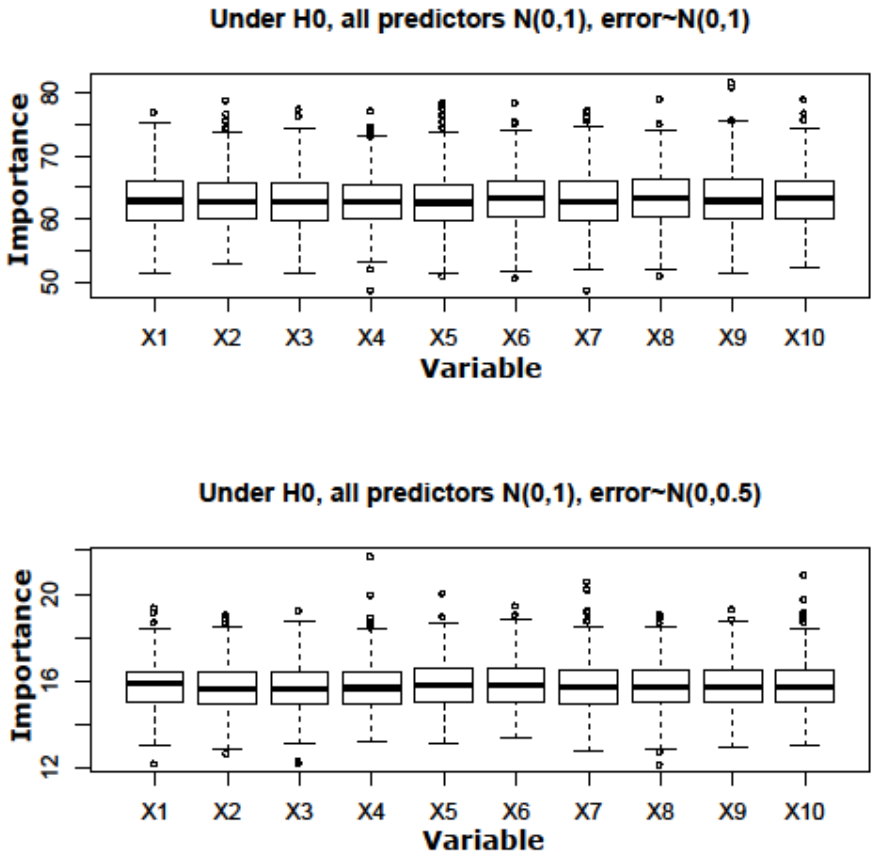

Figure 3.12. VIM_{Gini} under H_0. The top plot illustrates the VIM when the error follows a standard normal distribution. The bottom plot shows the VIM when error has a variance of 0.25. Continuous outcome.

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 VIM_{Gini}. So, the decrease of the variability of the outcome because of lower error variance may be the reason for the decrease on the VIM_{Gini} scores. Here, any predictor had higher scores than any other as their variability was the same, but in general the VIM_{Gini} was lower. This bias was unexpected as VIM_{Gini} was supposed to be checking the variables and it should be blind to the noise.

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 $0.3^2 * 1 + 0.25$ ($\beta^2 * \mu + \sigma^2$)), and was observed to be 1.09 when the error had variance one (as expected: $0.3\text{^}2 * 1 + 1(\beta^2 * \mu + \sigma^2)$). 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.

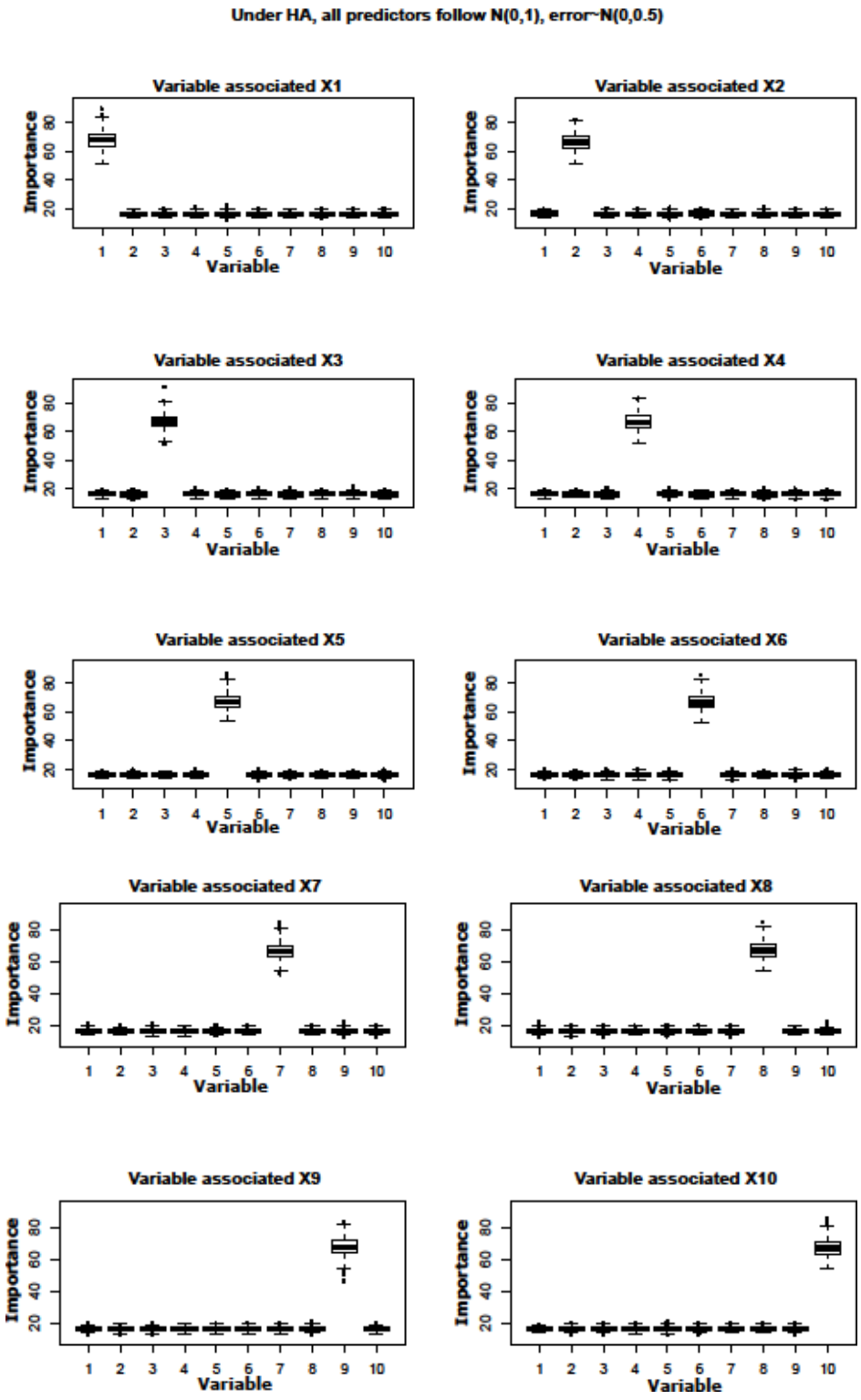

Figure 3.13. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors follow a standard normal distribution, but the error ~ N(0,0.5). Continuous outcome. Each number i of the X axis corresponds to the subscript of the variable X_i.

Therefore, the overall inflation of VIM_{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 *et al.* (2012b) suggested, VIM_{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_{Gini} scores. So, under association (signal not equal to 0), VIM_{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_{Gini} under H_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_{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_{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).

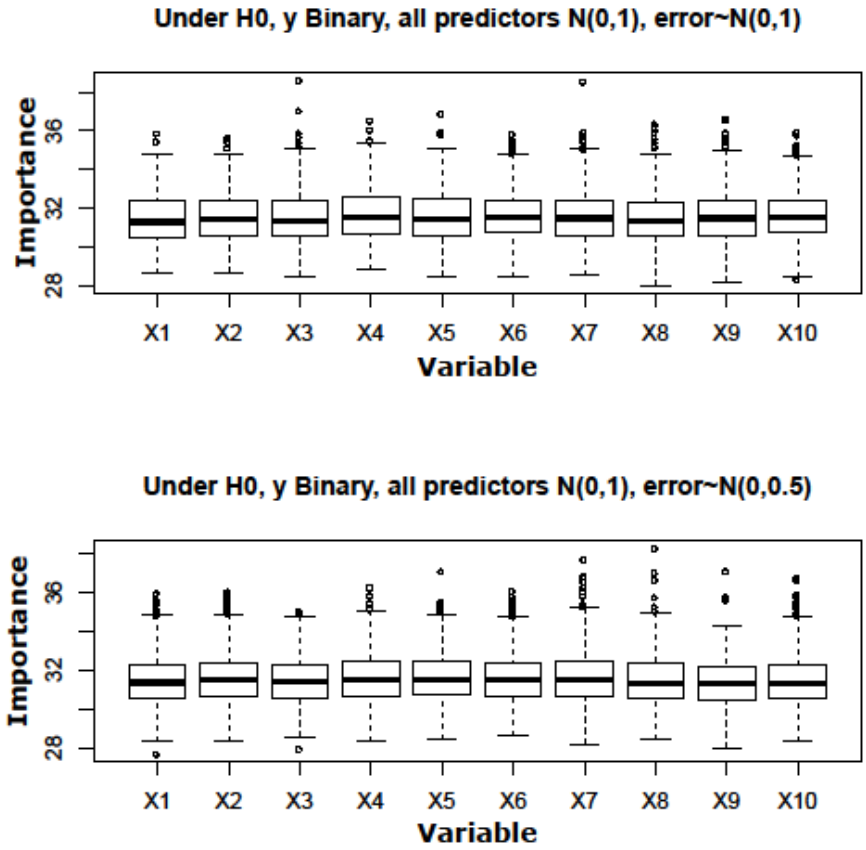

Figure 3.14. VIM_{Gini} under H_0. The top plot illustrates the VIM when the error followed a standard normal distribution. The bottom plot shows the VIM when error ~ N(0,0.5). Binary outcome.

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.

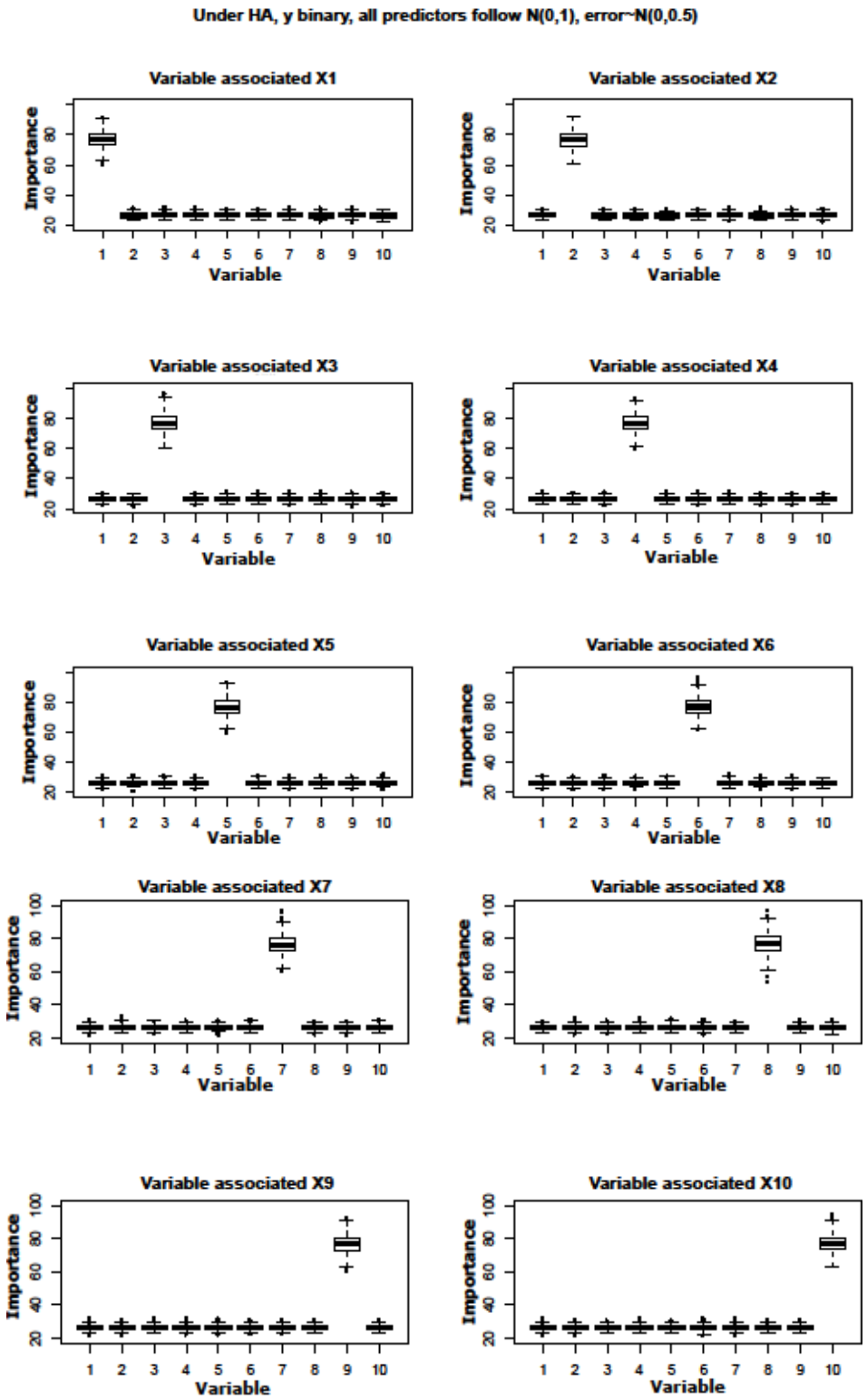

Figure 3.15. VIM_{Gini} under H_A. The figure illustrates VIM_{Gini} in the ten different single models, depending on which variable is associated, when all predictors followed a standard normal distribution, but the error ~ N(0,0.5). Binary outcome. Each number i of the X axis corresponds to the subscript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

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 VIM_{Gini} inflation observed when the error variance was lower. Therefore, VIM_{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.

UNDER H_0	**CONTINUOUS OUTCOME**	**BINARY OUTCOME**
$X_{10} \sim N(0,\sum_2)$, $e_1 \sim N(0,1)$ Variables with different variance	Unbiased	Unbiased
$X_{10} \sim N(0,I)$, $e_1 \sim N(0,1)$ Variables with different precision	**Biased**. Inflates the scores for more precise variables	**Biased**. Inflates the scores for more precise variables
$X_{10} \sim N(0,I)$, $e_1 \sim N(0,0.5)$ Less error variance	**Biased**. Inflates the scores when the error variance is higher	Unbiased

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.

UNDER H_A	**CONTINUOUS OUTCOME**	**BINARY OUTCOME**
$X_{10} \sim N(0,\sum_2)$, $e_1 \sim N(0,1)$ Variables with different variance	Inflates the scores for the predictors with more variance	Inflates the scores for the predictors with more variance
$X_{10} \sim N(0,I)$, $e_1 \sim N(0,1)$ Variables with different precision	Inflates the scores for more precise variables	Inflates the scores for more precise variables
$X_{10} \sim N(0,I)$, $e_1 \sim N(0,0.5)$ Less error variance	Inflates the scores when the error variance is higher	Inflates the scores when the error variance is lower

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.

3.4. Discussion

In this study, VIM_{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, VIM_{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, VIM_{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 VIM_{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, VIM_{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, VIM_{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, VIM_{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, VIM_{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,

VIM_{Gini} showed larger scores both for variables with more variability and for variables with more cut-points. With the binary outcome, VIM_{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 *et al.* 2006); (Pang and Zhao 2008), sometimes using data from different sources KEGG, BioCarta, and manually as Pang *et al.* (2006). Pang and Zhao (2008) applied RF based on VIM_{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_{Gini}) in gene expression data. Huynh-Thu *et al.* (2010) used VIM_{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_{Gini}. In the simulations they added random noise to the expression data measurement as well as in the dynamics of the networks. Furthermore, Petralia *et al.* (2015) also proposed a model to build genetic regulatory networks using VIM_{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_{Gini} is based on the decrease in impurity).

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_{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_{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_{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,

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 $VIM_{rawperm\text{-}RF}$ 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

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).

MGI Phenotype	N Human Genes	N Functional SNPs
Aggression	79	440
Depression	86	440
Fear/Anxiety	272	1446
Novelty	189	1067

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.

I set the number of trees to 1000 and the *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 5.55×10^{-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).

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:

2 way SNP interactions

$$\textit{Full model:}\ Y \sim \beta_1 SNP_i + \beta_2 SNP_j + \beta_3 SNP_i * SNP_j$$

$$\textit{Reduced Model:}\ Y \sim \beta_1 SNP_i + \beta_2 SNP_j$$

3 way SNP interactions

$$\textit{Full model:}\ Y \sim \beta_1 SNP_i + \beta_2 SNP_j + \beta_3 SNP_k + \beta_4 SNP_i * SNP_j + \beta_5 SNP_i * SNP_k + \beta_6 SNP_j * SNP_k + \beta_7 SNP_i * SNP_j * SNP_k$$

$$\textit{Reduced Model:}\ Y \sim \beta_1 SNP_i + \beta_2 SNP_j + \beta_3 SNP_k + \beta_4 SNP_i * SNP_j + \beta_5 SNP_i * SNP_k + \beta_6 SNP_j * SNP_k$$

Where

SNP_i , SNP_j and 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.

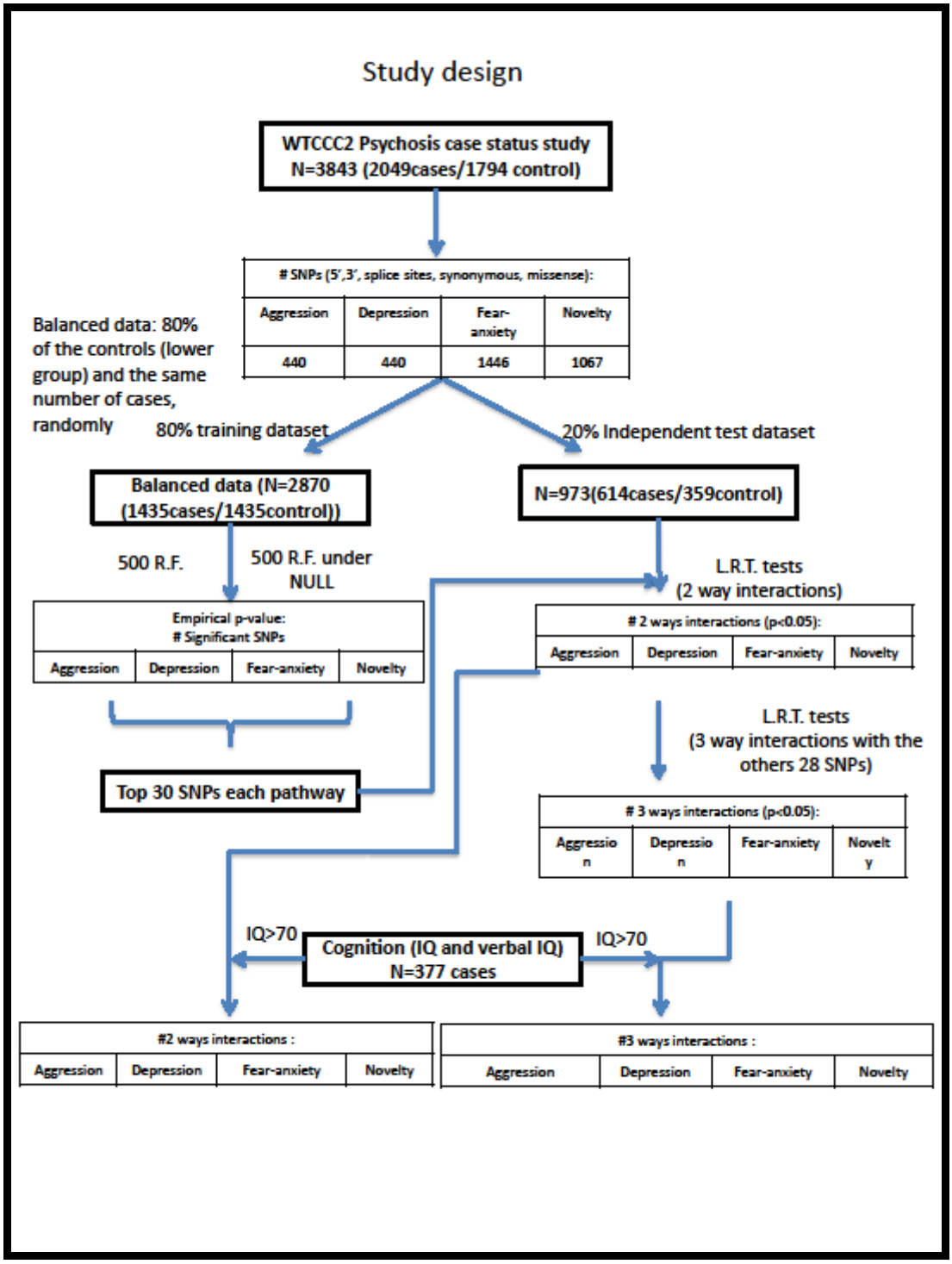

Figure 4.1. Illustration of the study design and methods.

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.

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.).

STUDY	GENES	SNPs	P-VALUE	R^2%
Case/Control	LAMA2* CYP19A1	rs3749878 * rs934633	0.0091	0.95
IQ			0.046	1.04

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

STUDY	GENES	SNPs	P-VALUE	R^2%
Case/Control	MYO5C* ESR2	rs10163109*rs8006145	0.0024	1.27
Verbal IQ			0.036	1.14

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.

STUDY	GENES	SNPs	P-VALUE	R^2%
Case/control	GRIN1*PPT1*HYDIN	rs1126442*rs3131661*rs1798532	0.035	0.61
IQ			0.006	1.92
Verbal IQ			0.021	1.36
Case/control	NTRK2*KIRREL3*HYDIN	rs1047896*rs3802815*rs1798532	0.011	0.88
IQ			0.007	1.82
Verbal IQ			0.023	1.31
Case/control	NTRK2*ESR2*HYDIN	rs1047896*rs8006145*rs1798532	0.007	0.99
IQ			0.009	1.73
Verbal IQ			0.019	1.40
Case/control	MC5R*CYP19A1*CACNA1B	rs1541276*rs934633*rs11137342	0.019	0.76
IQ			0.019	1.44
Verbal IQ			0.034	1.17

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

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*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.

STUDY	GENES	SNPs	P-VALUE	R^2%
Case/control	GAD2*GRIN2A	rs2839677* rs9806806	0.023	0.71
IQ			0.015	1.53

Table 4.5. 2-way interaction with p-*value* < 0.05 in psychosis and IQ in depression pathway.

STUDY	GENES	SNPs	P-VALUE	R^2%
Case/control	IPCEF1*SLITRK1*UBA6	rs2236259*rs9593836*rs4860853	0.044	0.56
IQ			0.040	1.08

Table 4.6. 3-way interaction with p-*value* < 0.05 in psychosis and IQ in depression pathway.

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).

STUDY	GENES	SNPs	P-VALUE	R^2%
Case/control	CRHR1*ESR1	rs16940665*rs2077647	0.003	1.17
IQ			0.043	1.06
Verbal IQ			0.032	1.21
Case/control	CRHR1*ESR1	rs16940674*rs2077647	0.006	1.03
IQ			0.039	1.11
Verbal IQ			0.033	1.19
Case/control	CRHR1*ESR1	rs4640231*rs2077647	0.003	1.17
IQ			0.044	1.06

Verbal IQ			0.032	1.21
Case/control	MAPT*ESR1	rs62063776*rs2077647	0.003	1.20
IQ			0.021	1.39
Verbal IQ			0.018	1.45
Case/control	CTNS*ABCA2	rs2873624*rs7048567	0.004	1.12
IQ			0.014	1.57
Verbal IQ			0.007	1.88

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.

STUDY	GENES	SNPs	P-VALUE	R^2%
Case/control	TOM1L2*IDUA*UBA6	rs1108648*rs4690221*rs10794537	0.019	0.75
IQ			0.031	1.19

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

STUDY	GENES	SNPs	P-VALUE	R^2%
Case/control	ALS2*GM2A*CTNS	rs3219153*rs61740602*rs2873624	0.011	1.68
Verbal IQ			0.031	1.19

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

STUDY	GENES	SNPs	P-VALUE	R^2%
Case/control	CRHR1*ESR1*TOM1L2	rs16940665*rs2077647*rs1108648	0.007	0.99
IQ			0.032	1.18
Verbal IQ			0.048	1.01
Case/control	CRHR1*ESR1*TOM1L2	rs4640231*rs2077647* rs1108648	0.007	0.99
IQ			0.031	1.18
Verbal IQ			0.047	1.01
Case/control	MAPT*ESR1*TOM1L2	rs62063776*rs2077647*rs1108648	0.010	0.91
IQ			0.020	1.38
Verbal IQ			0.036	1.13
Case/control	ALS2*GM2A*CTNS	rs3219153* rs61740602* rs222754	0.001	1.37
IQ			0.006	1.92
Verbal IQ			0.003	2.27
Case/control	NOS1*CHD6*ADCY1	rs3741475* rs3746543* rs2471267	0.036	0.61
IQ			0.031	1.18
Verbal IQ			0.009	1.73

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

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

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

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.

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-pituitaty-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).

ERS1 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 *ERS1* 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.*

(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 *TOM1L2* gene's function, my findings showed significant interactions including the gene. The *TOM1L2* 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 *ESR1* and *TOM1L2*, is the 5-prime to *MAPT* on the genome (Poorkaj *et al.* 2001).

Although molecular interaction between *CRHR1/ERS1*, *MAPT/ESR1*, *CRHR1/ERS1/TOM1L2* and *CRHR1/ESR1/TOM1L2* 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

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

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

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_{Gini} when predictors were all continuous and independent of each other. This analysis was based upon suggestions from a previous study: that VIM_{Gini} may be preferred with this type of predictor under those conditions (Boulesteix *et al.* 2012b). It was also suggested in that study that it may be preferable to apply VIM_{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_{Gini} when the error had two different variances.

In pursuit of this aim, VIM_{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 *et al.*, (2012b) in Chapter 3, considering both continuous and binary outcomes.

The results of this study showed that VIM_{Gini} is biased towards predictors with higher precision, with either continuous or binary outcomes, even under conditions of no association (H_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_{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_{Gini} was shown to be biased towards error with greater variance, both with and without association between predictor and the outcome. When predictors were

associated, VIM_{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_{Gini}, it would be sensible to standardise variables and to use the same number of decimal places for all predictors. However, as VIM_{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_{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

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.

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_{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 *et al.* 2010c); (Wu *et al.* 2012), and it has been suggested that this should be assessed empirically (Nicodemus *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

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.

Other limitation of this thesis is the lack of a real applied study in Chapter 3. A real application of VIM_{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_{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

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

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.

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

Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics* 9:291–300 . doi: 10.1038/sj.ejhg.5200619

Allen J, Balfour R, Bell R, Marmot M (2014) Social determinants of mental health. *International Review of Psychiatry* 26:392–407 . doi: 10.3109/09540261.2014.928270

American Psychiatric Association (2013) Diagnostic and Statistical Manual of Mental Disorders. *American Psychiatric Association*

Amin Z, Canli T, Epperson CN (2005) Effect of Estrogen-Serotonin Interactions on Mood and Cognition. *Behavioral and Cognitive Neuroscience Reviews* 4:43–58 . doi: 10.1177/1534582305277152

Andreasen NC, Wilcox MA, Ho B-C, Epping E, Ziebell S, Zeien E, Weiss B, Wassink T (2012) Statistical epistasis and progressive brain change in schizophrenia: an approach for examining the relationships between multiple genes. *Molecular Psychiatry* 17:1093–102 . doi: 10.1038/mp.2011.108

Archer KJ, Kimes R V. (2008) Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis* 52:2249–2260 . doi: 10.1016/j.csda.2007.08.015

Ardlie K, Liu-Cordero SN, Eberle MA, Daly M, Barrett J, Winchester E, Lander ES, Kruglyak L (2001) Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *American Journal of Human Genetics* 69:582–9 . doi: 10.1086/323251

Armando M, Nelson B, Yung AR, Ross M, Birchwood M, Girardi P, Nastro PF (2010) Psychotic-like experiences and correlation with distress and depressive symptoms in a community sample of adolescents and young adults. *Schizophrenia Research* 119:258–265 . doi: 10.1016/j.schres.2010.03.001

Ayodele TO (2010) Machine Learning Overview. In: New Advances in Machine Learning. *InTech*

Badner JA, Gershon ES (2002) Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. *Molecular Psychiatry* 7:405–11 . doi: 10.1038/sj.mp.4001012

Banf M, Rhee SY (2017) Enhancing gene regulatory network inference through data integration with markov random fields. *Scientific Reports* 7:41174 . doi: 10.1038/srep41174

Barnett JH, Smoller JW (2009) The genetics of bipolar disorder. *Neuroscience* 164:331–43 . doi: 10.1016/j.neuroscience.2009.03.080

Baum AE, Akula N, Cabanero M, Cardona I, Corona W, Klemens B, Schulze TG, Cichon S, Rietschel M, Nöthen MM, Georgi A, Schumacher J, Schwarz M, Abou Jamra R, Höfels S, Propping P, Satagopan J, Detera-Wadleigh SD, Hardy J, McMahon FJ (2008) A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Molecular Psychiatry* 13:197–207 . doi: 10.1038/sj.mp.4002012

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* 57:289–300

Benjamini Y, Yekutieli D (2001) The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics.* 29:1165–1188

Berman KF, Schmidt PJ, Rubinow DR, Danaceau MA, Van Horn JD, Esposito G, Ostrem JL, Weinberger DR (1997) Modulation of cognition-specific cortical activity by gonadal steroids: a positron-emission tomography study in women. *Proceedings of the National Academy of Sciences of the United States of America* 94:8836–41

Bhaskar H, Hoyle DC, Singh S (2006) Machine learning in bioinformatics: a brief survey and recommendations for practitioners. *Computers in Biology and Medicine* 36:1104–25 . doi: 10.1016/j.compbiomed.2005.09.002

Bi W, Yan J, Stankiewicz P, Park S-S, Walz K, Boerkoel CF, Potocki L, Shaffer LG, Devriendt K, Nowaczyk MJM, Inoue K, Lupski JR (2002) Genes in a Refined Smith-Magenis Syndrome Critical Deletion Interval on Chromosome 17p11.2 and the Syntenic Region of the Mouse. *Genome Research* 12:713–728 . doi: 10.1101/gr.73702

Bien J, Taylor J, Tibshirani R (2013) A LASSO FOR HIERARCHICAL INTERACTIONS. *Annals of Statistics* 41:1111–1141 . doi: 10.1214/13-AOS1096

Blake JA, Eppig JT, Kadin JA, Richardson JE, Smith CL, Bult CJ (2017) Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Research* 45:D723–D729 . doi: 10.1093/nar/gkw1040

Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. *BMJ (Clinical Research Ed.)* 310:170

Bogren M, Mattisson C, Isberg P-E, Nettelbladt P (2009) How common are psychotic and bipolar disorders? A 50-year follow-up of the Lundby population. *Nordic Journal of Psychiatry* 63:336–46 . doi: 10.1080/08039480903009118

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on Computational learning theory* - COLT '92. *ACM Press, New York, New York, USA*, pp 144–152

Boulesteix A-L (2006) Maximally selected chi-square statistics and binary splits of nominal variables. *Biometrical Journal.* 48:838–48

Boulesteix A-L, Bender A, Lorenzo Bermejo J, Strobl C (2012a) Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. *Briefings in Bioinformatics* 13:292–304

Boulesteix A-L, Janitza S, Hapfelmeier A, Van Steen K, Strobl C (2015) Letter to the Editor: On the term "interaction" and related phrases in the literature on Random Forests. *Briefings in Bioinformatics* 16:338–45 . doi: 10.1093/bib/bbu012

Boulesteix A-L, Janitza S, Kruppa J, König IR (2012b) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2:493–507 . doi: 10.1002/widm.1072

Breiman, Leo.; Friedman, J.H.; Olshen, R.A. and Stone CJ (1984) Classification and regression trees. *CRC Spress*

Breiman L (2001) Random Forests. *Machine Learning* 45:5–32 . doi: 10.1023/A:1010933404324

Breiman L (1996) Bagging Predictors. *Machine Learning* 24:123–140 . doi: 10.1023/A:1018054314350

Breiman L (1993) Classification and regression trees. *Chapman & Hall*

Brellier F, Chiquet-Ehrismann R (2012) How do tenascins influence the birth and life of a malignant cell? *Journal of Cellular and Molecular Medicine* 16:32–40 . doi:

10.1111/j.1582-4934.2011.01360.x

Brockschmidt A, Todt U, Ryu S, Hoischen A, Landwehr C, Birnbaum S, Frenck W, Radlwimmer B, Lichter P, Engels H, Driever W, Kubisch C, Weber RG (2007) Severe mental retardation with breathing abnormalities (Pitt-Hopkins syndrome) is caused by haploinsufficiency of the neuronal bHLH transcription factor TCF4. *Human Molecular Genetics* 16:1488–1494 . doi: 10.1093/hmg/ddm099

Broome MR, Woolley JB, Tabraham P, Johns LC, Bramon E, Murray GK, Pariante C, McGuire PK, Murray RM (2005) What causes the onset of psychosis? *Schizophrenia Research* 79:23–34 . doi: 10.1016/j.schres.2005.02.007

Bruenig D, White MJ, Young RM, Voisey J (2014) Subclinical psychotic experiences in healthy young adults: associations with stress and genetic predisposition. *Genetic Testing and Molecular Biomarkers* 18:683–9 . doi: 10.1089/gtmb.2014.0111

Bühlmann P (2012) Bagging, Boosting and Ensemble Methods. In: *Handbook of Computational Statistics*. *Springer Berlin Heidelberg, Berlin, Heidelberg*, pp 985–1022

Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P (2005) Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology* 28:171–82 . doi: 10.1002/gepi.20041

Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology* 8:e1002822 . doi: 10.1371/journal.pcbi.1002822

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14:365–376 . doi: 10.1038/nrn3475

Cai N, Chang S, Li Y, Li Q, Hu J, Liang J, Song L, Kretzschmar W, Gan X, Nicod J, Rivera M, Deng H, Du B, Li K, Sang W, Gao J, Gao S, Ha B, Ho H-Y, Hu C, Hu J, Hu Z, Huang G, Jiang G, Jiang T, Jin W, Li G, Li K, Li Y, Li Y, Li Y, Lin Y-T, Liu L, Liu T, Liu Y, Liu Y, Lu Y, Lv L, Meng H, Qian P, Sang H, Shen J, Shi J, Sun J, Tao M, Wang G, Wang G, Wang J, Wang L, Wang X, Wang X, Yang H, Yang L, Yin Y, Zhang J, Zhang K, Sun N, Zhang W, Zhang X, Zhang Z, Zhong H, Breen G, Wang J, Marchini J, Chen Y, Xu Q, Xu X, Mott R, Huang G-J, Kendler K, Flint J (2015) Molecular signatures of major depression. *Current Biology* 25:1146–56 . doi: 10.1016/j.cub.2015.03.008

Cai Q, Shu X-O, Jin F, Dai Q, Wen W, Cheng J-R, Gao Y-T, Zheng W (2003) Genetic polymorphisms in the estrogen receptor alpha gene and risk of breast cancer: results from the Shanghai Breast Cancer Study. *Cancer Epidemiology, Biomarkers & Prevention* 12:853–9

Calle ML, Urrea V (2011) Letter to the editor: Stability of Random Forest importance measures. *Briefings in Bioinformatics* 12:86–9 . doi: 10.1093/bib/bbq011

Cannon M, Jones PB, Murray RM (2002) Obstetric Complications and Schizophrenia: Historical and Meta-Analytic Review. *American Journal of Psychiatry* 159:1080–1092 . doi: 10.1176/appi.ajp.159.7.1080

Cardno AG, Owen MJ (2014) Genetic Relationships Between Schizophrenia, Bipolar Disorder, and Schizoaffective Disorder. *Schizophrenia Bulletin* 40:504–515 . doi: 10.1093/schbul/sbu016

Caspi A, Moffitt TE, Cannon M, McClay J, Murray R, Harrington H, Taylor A, Arseneault L, Williams B, Braithwaite A, Poulton R, Craig IW (2005) Moderation of the Effect of Adolescent-Onset Cannabis Use on Adult Psychosis by a Functional Polymorphism in the Catechol-O-Methyltransferase Gene: Longitudinal Evidence of a Gene X Environment Interaction. *Biological Psychiatry* 57:1117–1127 . doi: 10.1016/j.biopsych.2005.01.026

Chan MK, Krebs M-O, Cox D, Guest PC, Yolken RH, Rahmoune H, Rothermundt M, Steiner J, Leweke FM, van Beveren NJM, Niebuhr DW, Weber NS, Cowan DN, Suarez-Pinilla P, Crespo-Facorro B, Mam-Lam-Fook C, Bourgin J, Wenstrup RJ, Kaldate RR, Cooper JD, Bahn S (2015) Development of a blood-based molecular biomarker test for identification of schizophrenia before disease onset. *Translational Psychiatry* 5:e601 . doi: 10.1038/tp.2015.91

Chen C-K, Lin S-K, Sham PC, Ball D, Loh E-W, Murray RM (2005) Morbid risk for psychiatric disorder among the relatives of methamphetamine users with and without psychosis. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 136B:87–91 . doi: 10.1002/ajmg.b.30187

Chen DT, Jiang X, Akula N, Shugart YY, Wendland JR, Steele CJM, Kassem L, Park J-H, Chatterjee N, Jamain S, Cheng A, Leboyer M, Muglia P, Schulze TG, Cichon S, Nöthen MM, Rietschel M, BiGS, McMahon FJ, Farmer A, McGuffin P, Craig I, Lewis C, Hosang G, Cohen-Woods S, Vincent JB, Kennedy JL, Strauss J (2013) Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. *Molecular Psychiatry* 18:195–205 . doi: 10.1038/mp.2011.157

Cichon S, Mühleisen TW, Degenhardt FA, Mattheisen M, Miró X, Strohmaier J, Steffens M, Meesters C, Herms S, Weingarten M, Priebe L, Haenisch B, Alexander M, Vollmer J, Breuer R, Schmäl C, Tessmann P, Moebus S, Wichmann H-E, Schreiber S, Müller-Myhsok B, Lucae S, Jamain S, Leboyer M, Bellivier F, Etain B, Henry C, Kahn J-P, Heath S, Bipolar Disorder Genome Study (BiGS) Consortium, Hamshere M, O'Donovan MC, Owen MJ, Craddock N, Schwarz M, Vedder H, Kammerer-Ciernioch J, Reif A, Sasse J, Bauer M, Hautzinger M, Wright A, Mitchell PB, Schofield PR, Montgomery GW, Medland

SE, Gordon SD, Martin NG, Gustafsson O, Andreassen O, Djurovic S, Sigurdsson E, Steinberg S, Stefansson H, Stefansson K, Kapur-Pojskic L, Oruc L, Rivas F, Mayoral F, Chuchalin A, Babadjanova G, Tiganov AS, Pantelejeva G, Abramova LI, Grigoroiu-Serbanescu M, Diaconu CC, Czerski PM, Hauser J, Zimmer A, Lathrop M, Schulze TG, Wienker TF, Schumacher J, Maier W, Propping P, Rietschel M, Nöthen MM (2011) Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. *American Journal of Human Genetics* 88:372–81 . doi: 10.1016/j.ajhg.2011.01.017

Colquhoun D (2014) An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science* 1:140216–140216 . doi: 10.1098/rsos.140216

CONVERGE consortium (2015) Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* 523:588–91 . doi: 10.1038/nature14659

Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* 11:2463–8

Cordell HJ (2009) Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics* 10:392–404 . doi: 10.1038/nrg2579

Cordell HJ, Clayton DG (2005) Genetic association studies. *Lancet (London, England)* 366:1121–31 . doi: 10.1016/S0140-6736(05)67424-7

Craddock N, Khodel V, Van Eerdewegh P, Reich T (1995) Mathematical limits of multilocus models: the genetic transmission of bipolar disorder. *American Journal of Human Genetics* 57:690–702

Craddock N, Sklar P (2009) Genetics of bipolar disorder: successful start to a long journey. *Trends in Genetics* 25:99–105 . doi: 10.1016/j.tig.2008.12.002

Cristóbal-Narváez P, Sheinbaum T, Rosa A, Ballespí S, de Castro-Catala M, Peña E, Kwapil TR, Barrantes-Vidal N (2016) The Interaction between Childhood Bullying and the FKBP5 Gene on Psychotic-Like Experiences and Stress Reactivity in Real Life. *PLoS One* 11:e0158809 . doi: 10.1371/journal.pone.0158809

Cryan JF, Dinan TG (2012) Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nature Reviews Neuroscience* 13:701–12 . doi: 10.1038/nrn3346

Culverhouse R, Suarez BK, Lin J, Reich T (2002) A perspective on epistasis: limits of models displaying no main effect. *American Journal of Human Genetics* 70:461–71 . doi: 10.1086/338759

da Silva BS, Rovaris DL, Schuch JB, Mota NR, Cupertino RB, Aroche AP, Bertuzzi GP, Karam RG, Vitola ES, Tovo-Rodrigues L, Grevet EH, Bau CHD (2016) Effects of corticotropin-releasing hormone receptor 1 SNPs on major depressive disorder are influenced by sex and smoking status. *J Affect Disord* 205:282–288 . doi: 10.1016/j.jad.2016.08.008

Dachtler J, Elliott C, Rodgers RJ, Baillie GS, Clapcote SJ (2016) Missense mutation in DISC1 C-terminal coiled-coil has GSK3β signaling and sex-dependent behavioral effects in mice. *Scientific Reports* 6:18748 . doi: 10.1038/srep18748

De Suzana, Santos S, Yasumasatakahashi D, Nakata A, Fujita A (2013) A comparative study of statistical methods used to identify dependencies between gene expression signals. doi: 10.1093/bib/bbt051

Díaz-Uriarte R, Alvarez de Andrés S (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7:3 . doi: 10.1186/1471-2105-7-3

Dietterich TG (2000a) Ensemble Methods in Machine Learning. *Springer Berlin Heidelberg*, pp 1–15

Dietterich TG (2000b) An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning* 40:139–157 . doi: 10.1023/A:1007607513941

Dima D, Breen G (2015) Polygenic risk scores in imaging genetics: Usefulness and applications. *Journal of Psychopharmacology* 29:867–871 . doi: 10.1177/0269881115584470

Donoho D, Stodden V (2006) Breakdown Point of Model Selection When the Number of Variables Exceeds the Number of Observations. In: *The 2006 IEEE International Joint Conference on Neural Network Proceedings.* IEEE, pp 1916–1921

Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology* 32:227–34 . doi: 10.1002/gepi.20297

Dudoit S, Keleş S, van der Laan MJ (2008) Multiple tests of association with biological annotation metadata. In: *Probability and Statistics: Essays in Honor of David A. Freedman*. Institute of Mathematical Statistics, Beachwood, Ohio, USA, pp 153–218

Egan MF, Goldberg TE, Kolachana BS, Callicott JH, Mazzanti CM, Straub RE, Goldman D, Weinberger DR (2001) Effect of COMT Val108/158 Met genotype

on frontal lobe function and risk for schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* 98:6917–22 . doi: 10.1073/pnas.111134598

Eisenberger NI, Cole SW (2012) Social neuroscience and health: neurophysiological mechanisms linking social ties with physical health. *Nature Neuroscience* 15:669–74 . doi: 10.1038/nn.3086

Elbaz A, Ross OA, Ioannidis JPA, Soto-Ortolaza AI, Moisan F, Aasly J, Annesi G, Bozi M, Brighina L, Chartier-Harlin M-C, Destée A, Ferrarese C, Ferraris A, Gibson JM, Gispert S, Hadjigeorgiou GM, Jasinska-Myga B, Klein C, Krüger R, Lambert J-C, Lohmann K, van de Loo S, Loriot M-A, Lynch T, Mellick GD, Mutez E, Nilsson C, Opala G, Puschmann A, Quattrone A, Sharma M, Silburn PA, Stefanis L, Uitti RJ, Valente EM, Vilariño-Güell C, Wirdefeldt K, Wszolek ZK, Xiromerisiou G, Maraganore DM, Farrer MJ, Genetic Epidemiology of Parkinson's Disease (GEO-PD) Consortium (2011) Independent and joint effects of the MAPT and SNCA genes in Parkinson disease. *Annals of Neurology* 69:778–792 . doi: 10.1002/ana.22321

Ferreira MAR, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, Fan J, Kirov G, Perlis RH, Green EK, Smoller JW, Grozeva D, Stone J, Nikolov I, Chambert K, Hamshere ML, Nimgaonkar VL, Moskvina V, Thase ME, Caesar S, Sachs GS, Franklin J, Gordon-Smith K, Ardlie KG, Gabriel SB, Fraser C, Blumenstiel B, Defelice M, Breen G, Gill M, Morris DW, Elkin A, Muir WJ, McGhee KA, Williamson R, MacIntyre DJ, MacLean AW, St CD, Robinson M, Van Beck M, Pereira ACP, Kandaswamy R, McQuillin A, Collier DA, Bass NJ, Young AH, Lawrence J, Ferrier IN, Anjorin A, Farmer A, Curtis D, Scolnick EM, McGuffin P, Daly MJ, Corvin AP, Holmans PA, Blackwood DH, Gurling HM, Owen MJ, Purcell SM, Sklar P, Craddock N, Wellcome Trust Case Control Consortium (2008) Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature Genetics* 40:1056–8 . doi: 10.1038/ng.209

Fineberg NA, Haddad PM, Carpenter L, Gannon B, Sharpe R, Young AH, Joyce E, Rowe J, Wellsted D, Nutt DJ, Sahakian BJ (2013) The size, burden and cost of disorders of the brain in the UK. *Journal of Psychopharmacology* 27:761–70 . doi: 10.1177/0269881113495118

Fink M, Taylor MA (2008) Issues for DSM-V: The Medical Diagnostic Model. *American Journal of Psychiatry* 165:799–799 . doi: 10.1176/appi.ajp.2008.08020245

Fisher HL, Caspi A, Poulton R, Meier MH, Houts R, Harrington H, Arseneault L, Moffitt TE (2013) Specificity of childhood psychotic symptoms for predicting schizophrenia by 38 years of age: a birth cohort study. *Journal of Psychopharmacology* 43:2077–86 . doi: 10.1017/S0033291712003091

Friedman JH (2000) Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29:1189--1232

Funk CK, O'Dell LE, Crawford EF, Koob GF (2006) Corticotropin-Releasing Factor within the Central Nucleus of the Amygdala Mediates Enhanced Ethanol Self-Administration in Withdrawn, Ethanol-Dependent Rats. *Journal of Neuroscience* 26:11324–11332 . doi: 10.1523/JNEUROSCI.3096-06.2006

García-Magariños M, López-de-Ullibarri I, Cao R, Salas A (2009) Evaluating the Ability of Tree-Based Methods and Logistic Regression for the Detection of SNP-SNP Interaction. *Annals of Human Genetics* 73:360–369 . doi: 10.1111/j.1469-1809.2009.00511.x

Gejman P V, Sanders AR, Duan J (2010) The role of genetics in the etiology of schizophrenia. *The Psychiatric Clinics of North America* 33:35–66 . doi: 10.1016/j.psc.2009.12.003

Genz A, Bretz F (2009) Computation of multivariate normal and t probabilities. *Springer.* doi: 10.1007/978-3-642-01689-9

George PF (2014) DSM-5 and Psychotic and Mood Disorders. *Journal of the American Academy of Psychiatry and the Law* 42:182–190

Ghert MA, Qi WN, Erickson HP, Block JA, Scully SP (2001) Tenascin-C splice variant adhesive/anti-adhesive effects on chondrosarcoma cell attachment to fibronectin. *Cell Structure and Function* 26:179–87

Gibson G (2012) Rare and common variants: twenty arguments. *Nature Reviews Genetics* 13:135–145 . doi: 10.1038/nrg3118

Girirajan S, Hauck PM, Williams S, Vlangos CN, Szomju BB, Solaymani-Kohal S, Mosier PD, White KL, McCoy K, Elsea SH (2008) Tom1l2 hypomorphic mice exhibit increased incidence of infections and tumors and abnormal immunologic response. *Mammalian Genome* 19:246–262 . doi: 10.1007/s00335-008-9100-6

Gogtay N, Vyas NS, Testa R, Wood SJ, Pantelis C (2011) Age of onset of schizophrenia: perspectives from structural neuroimaging studies. *Schizophrenia Bulletin* 37:504–13 . doi: 10.1093/schbul/sbr030

Goldstein BA, Hubbard AE, Cutler A, Barcellos LF (2010) An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genetics* 11:49 . doi: 10.1186/1471-2156-11-49

Goldstein BA, Polley EC, Briggs FBS (2011) Random forests for genetic association studies. *Statistical Applications in Genetics and Molecular Biology* 10:32 . doi: 10.2202/1544-6115.1691

Gondro C, van der Werf J, Hayes B (eds) (2013) Genome-Wide Association Studies and Genomic Prediction. *Humana Press*, Totowa, NJ

Green EK, Grozeva D, Jones I, Jones L, Kirov G, Caesar S, Gordon-Smith K, Fraser C, Forty L, Russell E, Hamshere ML, Moskvina V, Nikolov I, Farmer A, McGuffin P, Wellcome Trust Case Control Consortium, Holmans PA, Owen MJ, O'Donovan MC, Craddock N (2010) The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. *Molecular Psychiatry* 15:1016–22 . doi: 10.1038/mp.2009.49

Greene GL, Gilna P, Waterfield M, Baker A, Hort Y, Shine J (1986) Sequence and expression of human estrogen receptor complementary DNA. *Science* 231:1150–4

Griffiths AJ, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM (2000) Quantifying heritability. W. H. Freeman.

Grimm S, Wirth K, Fan Y, Weigand A, Gärtner M, Feeser M, Dziobek I, Bajbouj M, Aust S (2017) The interaction of corticotropin-releasing hormone receptor gene and early life stress on emotional empathy. *Behavioural Brain Research* 329:180–185 . doi: 10.1016/j.bbr.2017.04.047

Guerreiro RJ, Gustafson DR, Hardy J (2012) The genetic architecture of Alzheimer's disease: beyond APP, PSENs and APOE. *Neurobiology of Aging* 33:437–456 . doi: 10.1016/j.neurobiolaging.2010.03.025

Guyon I, Elisseeff A (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 1157–1182

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 46:389–422 . doi: 10.1023/A:1012487302797

Hannan AJ (2013) Nature, nurture and neurobiology: Gene-environment interactions in neuropsychiatric disorders. *Neurobiology of Disease* 57:1–4 . doi: 10.1016/j.nbd.2013.01.004

Hargreaves A, Anney R, O'Dushlaine C, Nicodemus KK, Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ) M, Wellcome Trust Case Control Consortium 2 A, Gill M, Corvin A, Morris D, Donohoe G (2014) The one and the many: effects of the cell adhesion molecule pathway on neuropsychological function in psychosis. *Journal of Psychopharmacology* 44:2177–87 . doi: 10.1017/S0033291713002663

Harrison PJ, Owen MJ (2003) Genes for schizophrenia? Recent findings and their pathophysiological implications. *Lancet (London, England)* 361:417–9 . doi:

10.1016/S0140-6736(03)12379-3

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. *Springer New York*, New York, NY. doi: 10.1007/978-0-387-84858-7.

Hawkins DM (2004) The problem of overfitting. *Journal of Chemical Information and Computer Sciences* 44:1–12 . doi: 10.1021/ci0342472

Heckers S, Tandon R, Bustillo J (2010) Catatonia in the DSM--shall we move or not? *Schizophrenia Bulletin* 36:205–7 . doi: 10.1093/schbul/sbp136

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* 6:95–108 . doi: 10.1038/nrg1521

Hothorn T, Hornik K, Strobl C, Zeileis A (2015) party package. http://party.r-forge.r-project.org. Accessed 6 Oct 2016

Hothorn T, Hornik K, Zeileis A (2006) Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15:651–674 . doi: 10.1198/106186006X133933

Hou L, Bergen SE, Akula N, Song J, Hultman CM, Landen M, Adli M, Alda M, Ardau R, Arias B, Aubry J-M, Backlund L, Badner JA, Barrett TB, Bauer M, Baune BT, Bellivier F, Benabarre A, Bengesser S, Berrettini WH, Bhattacharjee AK, Biernacka JM, Birner A, Bloss CS, Brichant-Petitjean C, Bui ET, Byerley W, Cervantes P, Chillotti C, Cichon S, Colom F, Coryell W, Craig DW, Cruceanu C, Czerski PM, Davis T, Dayer A, Degenhardt F, Del Zompo M, DePaulo JR, Edenberg HJ, Etain B, Falkai P, Foroud T, Forstner AJ, Frisen L, Frye MA, Fullerton JM, Gard S, Garnham JS, Gershon ES, Goes FS, Greenwood TA, Grigoroiu-Serbanescu M, Hauser J, Heilbronner U, Heilmann-Heimbach S, Herms S, Hipolito M, Hitturlingappa S, Hoffmann P, Hofmann A, Jamain S, Jimenez E, Kahn J-P, Kassem L, Kelsoe JR, Kittel-Schneider S, Kliwicki S, Koller DL, Konig B, Lackner N, Laje G, Lang M, Lavebratt C, Lawson WB, Leboyer M, Leckband SG, Liu C, Maaser A, Mahon PB, Maier W, Maj M, Manchia M, Martinsson L, McCarthy MJ, McElroy SL, McInnis MG, McKinney R, Mitchell PB, Mitjans M, Mondimore FM, Monteleone P, Muhleisen TW, Nievergelt CM, Nothen MM, Novak T, Nurnberger JI, Nwulia EA, Osby U, Pfennig A, Potash JB, Propping P, Reif A, Reininghaus E, Rice J, Rietschel M, Rouleau GA, Rybakowski JK, Schalling M, Scheftner WA, Schofield PR, Schork NJ, Schulze TG, Schumacher J, Schweizer BW, Severino G, Shekhtman T, Shilling PD, Simhandl C, Slaney CM, Smith EN, Squassina A, Stamm T, Stopkova P, Streit F, Strohmaier J, Szelinger S, Tighe SK, Tortorella A, Turecki G, Vieta E, Volkert J, Witt SH, Wright A, Zandi PP, Zhang P, Zollner S, McMahon FJ (2016a) Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. *bioRxiv*

Hou L, Heilbronner U, Degenhardt F, Adli M, Akiyama K, Akula N, Ardau R, Arias B, Backlund L, Banzato CEM, Benabarre A, Bengesser S, Bhattacharjee AK, Biernacka JM, Birner A, Brichant-Petitjean C, Bui ET, Cervantes P, Chen G-B, Chen H-C, Chillotti C, Cichon S, Clark SR, Colom F, Cousins DA, Cruceanu C, Czerski PM, Dantas CR, Dayer A, Étain B, Falkai P, Forstner AJ, Frisén L, Fullerton JM, Gard S, Garnham JS, Goes FS, Grof P, Gruber O, Hashimoto R, Hauser J, Herms S, Hoffmann P, Hofmann A, Jamain S, Jiménez E, Kahn J-P, Kassem L, Kittel-Schneider S, Kliwicki S, König B, Kusumi I, Lackner N, Laje G, Landén M, Lavebratt C, Leboyer M, Leckband SG, Jaramillo CAL, MacQueen G, Manchia M, Martinsson L, Mattheisen M, McCarthy MJ, McElroy SL, Mitjans M, Mondimore FM, Monteleone P, Nievergelt CM, Nöthen MM, Ösby U, Ozaki N, Perlis RH, Pfennig A, Reich-Erkelenz D, Rouleau GA, Schofield PR, Schubert KO, Schweizer BW, Seemüller F, Severino G, Shekhtman T, Shilling PD, Shimoda K, Simhandl C, Slaney CM, Smoller JW, Squassina A, Stamm T, Stopkova P, Tighe SK, Tortorella A, Turecki G, Volkert J, Witt S, Wright A, Young LT, Zandi PP, Potash JB, DePaulo JR, Bauer M, Reininghaus EZ, Novák T, Aubry J-M, Maj M, Baune BT, Mitchell PB, Vieta E, Frye MA, Rybakowski JK, Kuo P-H, Kato T, Grigoroiu-Serbanescu M, Reif A, Del Zompo M, Bellivier F, Schalling M, Wray NR, Kelsoe JR, Alda M, Rietschel M, McMahon FJ, Schulze TG (2016b) Genetic variants associated with response to lithium treatment in bipolar disorder: a genome-wide association study. *Lancet (London, England)* 387:1085–93 . doi: 10.1016/S0140-6736(16)00143-4

Howie B, Marchini J, Stephens M (2011) Genotype Imputation with Thousands of Genomes. *G3 Genes|Genomes|Genetics* 1:457–470 . doi: 10.1534/g3.111.001198

Huang J, Perlis RH, Lee PH, Rush AJ, Fava M, Sachs GS, Lieberman J, Hamilton SP, Sullivan P, Sklar P, Purcell S, Smoller JW (2010) Cross-Disorder Genomewide Analysis of Schizophrenia, Bipolar Disorder, and Depression. *American Journal of Psychiatry* 167:1254–1263 . doi: 10.1176/appi.ajp.2010.09091335

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS One* 5:e12776 . doi: 10.1371/journal.pone.0012776

Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR, Tung JY, Hinds DA, Perlis RH, Winslow AR (2016) Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nature Genetics* 48:1031–6 . doi: 10.1038/ng.3623

Iniesta R, Stahl D, McGuffin P (2016) Machine learning, statistical learning and the future of biological research in psychiatry. *Journal of Psychopharmacology* 46:2455–65 . doi: 10.1017/S0033291716001367

Insel T (2012) Director's Blog: Research Domain Criteria -- RDoC. https://www.nimh.nih.gov/about/director/2012/research-domain-criteria-rdoc.shtml

Insel TR (2014) The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry. *American Journal of Psychiatry* 171:395–397 . doi: 10.1176/appi.ajp.2014.14020138

International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460:748–52 . doi: 10.1038/nature08185

Ioannidis JPA, Ralston SH, Bennett ST, Brandi ML, Grinberg D, Karassa FB, Langdahl B, van Meurs JBJ, Mosekilde L, Scollen S, Albagha OME, Bustamante M, Carey AH, Dunning AM, Enjuanes A, van Leeuwen JPTM, Mavilia C, Masi L, McGuigan FEA, Nogues X, Pols HAP, Reid DM, Schuit SCE, Sherlock RE, Uitterlinden AG, GENOMOS Study (2004) Differential genetic effects of ESR1 gene polymorphisms on osteoporosis outcomes. *JAMA* 292:2105 . doi: 10.1001/jama.292.17.2105

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

Ishwaran, Hemant, Kogalur, Udaya B., Blackstone, Eugene H., Lauer MS (2008) RANDOM SURVIVAL FORESTS. *The Annals of Applied Statistics* 2:841–860

Ishwaran H, Kogalur UB (2007) Random Survival Forests for R. *R news* 7:25--31

Ishwaran H, Kogalur UB, Chen X, Minn AJ (2011) Random Survival Forests for High-Dimensional Data. doi: 10.1002/sam.10103

Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS (2010) High-Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association* 105:205–217 . doi: 10.1198/jasa.2009.tm08622

Janitza S, Strobl C, Boulesteix A-L (2013) An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics* 14:119 . doi: 10.1186/1471-2105-14-119

Jansen R, Penninx BWJH, Madar V, Xia K, Milaneschi Y, Hottenga JJ, Hammerschlag AR, Beekman A, van der Wee N, Smit JH, Brooks AI, Tischfield J, Posthuma D, Schoevers R, van Grootheest G, Willemsen G, de Geus EJ, Boomsma DI, Wright FA, Zou F, Sun W, Sullivan PF (2016) Gene expression in major depressive

disorder. *Molecular Psychiatry* 21:339–347 . doi: 10.1038/mp.2015.57

Johns LC, Cannon M, Singleton N, Murray RM, Farrell M, Brugha T, Bebbington P, Jenkins R, Meltzer H (2004) Prevalence and correlates of self-reported psychotic symptoms in the British population. *The British Journal of Psychiatry* 185:298–305 . doi: 10.1192/bjp.185.4.298

Johnson SL (2005) Life events in bipolar disorder: Towards more specific models. *Clinical Psychology Review* 25:1008–1027 . doi: 10.1016/j.cpr.2005.06.004

Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science* 349: 255-260

Joyce C (2014) Transforming Our Approach to Translational Neuroscience: The Role and Impact of Charitable Nonprofits in Research. *Neuron* 84:526–532 . doi: 10.1016/j.neuron.2014.10.030

Judy JT, Seifuddin F, Pirooznia M, Mahon PB, Bipolar Genome Study Consortium TBGS, Jancic D, Goes FS, Schulze T, Cichon S, Noethen M, Rietschel M, Depaulo JR, Potash JB, Zandi PP, Zandi PP (2013) Converging Evidence for Epistasis between ANK3 and Potassium Channel Gene KCNQ2 in Bipolar Disorder. *Frontiers in Genetics* 4:87 . doi: 10.3389/fgene.2013.00087

Kahn RS, Keefe RSE (2013) Schizophrenia is a cognitive illness: time for a change in focus. *JAMA psychiatry* 70:1107–12 . doi: 10.1001/jamapsychiatry.2013.155

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28:27–30

Kavanagh DH, Tansey KE, O'Donovan MC, Owen MJ (2015) Schizophrenia genetics: emerging themes for a complex disorder. *Molecular Psychiatry* 20:72–76 . doi: 10.1038/mp.2014.148

Kelleher I, Corcoran P, Keeley H, Wigman JTW, Devlin N, Ramsay H, Wasserman C, Carli V, Sarchiapone M, Hoven C, Wasserman D, Cannon M (2013) Psychotic symptoms and population risk for suicide attempt: a prospective cohort study. *JAMA psychiatry* 70:940–8 . doi: 10.1001/jamapsychiatry.2013.140

Kelleher I, Devlin N, Wigman JTW, Kehoe A, Murtagh A, Fitzpatrick C, Cannon M (2014) Psychotic experiences in a mental health clinic sample: implications for suicidality, multimorbidity and functioning. *Journal of Psychopharmacology* 44:1615–24 . doi: 10.1017/S0033291713002122

Keller J, Flores B, Gomez RG, Solvason HB, Kenna H, Williams GH, Schatzberg AF (2006) Cortisol Circadian Rhythm Alterations in Psychotic Major Depression. *Biological Psychiatry* 60:275–281 . doi: 10.1016/j.biopsych.2005.10.014

Ketter TA (2010) Diagnostic features, prevalence, and impact of bipolar disorder. *The Journal of Clinical Psychiatry* 71:e14 . doi: 10.4088/JCP.8125tx11c

Kohavi R, John GH (1997) Wrappers for feature subset selection. *Artificial Intelligence* 97:273–324 . doi: 10.1016/S0004-3702(97)00043-X

Koo CL, Liew MJ, Mohamad MS, Salleh AHM (2013) A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *BioMed Research International* 2013:432375 . doi: 10.1155/2013/432375

Kooperberg C, LeBlanc M, Obenchain V (2010) Risk prediction using genome-wide association studies. *Genetic Epidemiology* 34:643–52 . doi: 10.1002/gepi.20509

Korkeila M, Kaprio J, Rissanen A, Koskenvuo M (1991) Effects of gender and age on the heritability of body mass index. *International Journal of Obesity* 15:647–54

Krug A, Krach S, Jansen A, Nieratschker V, Witt SH, Shah NJ, Nothen MM, Rietschel M, Kircher T (2013) The Effect of Neurogranin on Neural Correlates of Episodic Memory Encoding and Retrieval. *Schizophrenia Bulletin* 39:141–150 . doi: 10.1093/schbul/sbr076

Kruppa J, Ziegler A, König IR (2012) Risk estimation and risk prediction using machine-learning methods. *Human Genetics* 131:1639–54 . doi: 10.1007/s00439-012-1194-y

Kwok JBJ, Loy CT, Hamilton G, Lau E, Hallupp M, Williams J, Owen MJ, Broe GA, Tang N, Lam L, Powell JF, Lovestone S, Schofield PR (2008) Glycogen synthase kinase-3β and tau genes interact in Alzheimer's disease. *Annals of Neurology* 64:446–454 . doi: 10.1002/ana.21476

Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics* 7:385–394 . doi: 10.1038/nrg1839

Larson NB, Schaid DJ (2013) A Kernel Regression Approach to Gene-Gene Interaction Detection for Case-Control Studies. *Genetic Epidemiology* 37:695–703 . doi: 10.1002/gepi.21749

Laruelle M, Abi-Dargham A, van Dyck CH, Gil R, D'Souza CD, Erdos J, McCance E, Rosenblatt W, Fingado C, Zoghbi SS, Baldwin RM, Seibyl JP, Krystal JH, Charney DS, Innis RB (1996) Single photon emission computerized tomography imaging of amphetamine-induced dopamine release in drug-free schizophrenic subjects. *Proceedings of the National Academy of Sciences of the United States of America* 93:9235–40

Lawrie SM, O'Donovan MC, Saks E, Burns T, Lieberman JA (2016) Improving classification of psychoses. *The Lancet Psychiatry* 3:367–374 . doi:

10.1016/S2215-0366(15)00577-5

Lee SH, Ripke S, Neale BM, Faraone S V, Purcell SM, Perlis RH, Mowry BJ, Thapar A, Goddard ME, Witte JS, Absher D, Agartz I, Akil H, Amin F, Andreassen OA, Anjorin A, Anney R, Anttila V, Arking DE, Asherson P, Azevedo MH, Backlund L, Badner JA, Bailey AJ, Banaschewski T, Barchas JD, Barnes MR, Barrett TB, Bass N, Battaglia A, Bauer M, Bayés M, Bellivier F, Bergen SE, Berrettini W, Betancur C, Bettecken T, Biederman J, Binder EB, Black DW, Blackwood DHR, Bloss CS, Boehnke M, Boomsma DI, Breen G, Breuer R, Bruggeman R, Cormican P, Buccola NG, Buitelaar JK, Bunney WE, Buxbaum JD, Byerley WF, Byrne EM, Caesar S, Cahn W, Cantor RM, Casas M, Chakravarti A, Chambert K, Choudhury K, Cichon S, Cloninger CR, Collier DA, Cook EH, Coon H, Cormand B, Corvin A, Coryell WH, Craig DW, Craig IW, Crosbie J, Cuccaro ML, Curtis D, Czamara D, Datta S, Dawson G, Day R, De Geus EJ, Degenhardt F, Djurovic S, Donohoe GJ, Doyle AE, Duan J, Dudbridge F, Duketis E, Ebstein RP, Edenberg HJ, Elia J, Ennis S, Etain B, Fanous A, Farmer AE, Ferrier IN, Flickinger M, Fombonne E, Foroud T, Frank J, Franke B, Fraser C, Freedman R, Freimer NB, Freitag CM, Friedl M, Frisén L, Gallagher L, Gejman P V, Georgieva L, Gershon ES, Geschwind DH, Giegling I, Gill M, Gordon SD, Gordon-Smith K, Green EK, Greenwood TA, Grice DE, Gross M, Grozeva D, Guan W, Gurling H, De Haan L, Haines JL, Hakonarson H, Hallmayer J, Hamilton SP, Hamshere ML, Hansen TF, Hartmann AM, Hautzinger M, Heath AC, Henders AK, Herms S, Hickie IB, Hipolito M, Hoefels S, Holmans PA, Holsboer F, Hoogendijk WJ, Hottenga J-J, Hultman CM, Hus V, Ingason A, Ising M, Jamain S, Jones EG, Jones I, Jones L, Tzeng J-Y, Kähler AK, Kahn RS, Kandaswamy R, Keller MC, Kennedy JL, Kenny E, Kent L, Kim Y, Kirov GK, Klauck SM, Klei L, Knowles JA, Kohli MA, Koller DL, Konte B, Korszun A, Krabbendam L, Krasucki R, Kuntsi J, Kwan P, Landén M, Långström N, Lathrop M, Lawrence J, Lawson WB, Leboyer M, Ledbetter DH, Lee PH, Lencz T, Lesch K-P, Levinson DF, Lewis CM, Li J, Lichtenstein P, Lieberman JA, Lin D-Y, Linszen DH, Liu C, Lohoff FW, Loo SK, Lord C, Lowe JK, Lucae S, MacIntyre DJ, Madden PAF, Maestrini E, Magnusson PKE, Mahon PB, Maier W, Malhotra AK, Mane SM, Martin CL, Martin NG, Mattheisen M, Matthews K, Mattingsdal M, McCarroll SA, McGhee KA, McGough JJ, McGrath PJ, McGuffin P, McInnis MG, McIntosh A, McKinney R, McLean AW, McMahon FJ, McMahon WM, McQuillin A, Medeiros H, Medland SE, Meier S, Melle I, Meng F, Meyer J, Middeldorp CM, Middleton L, Milanova V, Miranda A, Monaco AP, Montgomery GW, Moran JL, Moreno-De-Luca D, Morken G, Morris DW, Morrow EM, Moskvina V, Muglia P, Mühleisen TW, Muir WJ, Müller-Myhsok B, Murtha M, Myers RM, Myin-Germeys I, Neale MC, Nelson SF, Nievergelt CM, Nikolov I, Nimgaonkar V, Nolen WA, Nöthen MM, Nurnberger JI, Nwulia EA, Nyholt DR, O'Dushlaine C, Oades RD, Olincy A, Oliveira G, Olsen L, Ophoff RA, Osby U, Owen MJ, Palotie A, Parr JR, Paterson AD, Pato CN, Pato MT, Penninx BW, Pergadia ML, Pericak-Vance MA, Pickard BS, Pimm J, Piven J, Posthuma D, Potash JB, Poustka F, Propping P, Puri V, Quested DJ, Quinn EM, Ramos-Quiroga JA, Rasmussen HB, Raychaudhuri S, Rehnström K, Reif A, Ribasés M, Rice JP, Rietschel M, Roeder K, Roeyers H, Rossin L, Rothenberger

A, Rouleau G, Ruderfer D, Rujescu D, Sanders AR, Sanders SJ, Santangelo SL, Sergeant JA, Schachar R, Schalling M, Schatzberg AF, Scheftner WA, Schellenberg GD, Scherer SW, Schork NJ, Schulze TG, Schumacher J, Schwarz M, Scolnick E, Scott LJ, Shi J, Shilling PD, Shyn SI, Silverman JM, Slager SL, Smalley SL, Smit JH, Smith EN, Sonuga-Barke EJS, St. Clair D, State M, Steffens M, Steinhausen H-C, Strauss JS, Strohmaier J, Stroup TS, Sutcliffe JS, Szatmari P, Szelinger S, Thirumalai S, Thompson RC, Todorov AA, Tozzi F, Treutlein J, Uhr M, van den Oord EJCG, Van Grootheest G, Van Os J, Vicente AM, Vieland VJ, Vincent JB, Visscher PM, Walsh CA, Wassink TH, Watson SJ, Weissman MM, Werge T, Wienker TF, Wijsman EM, Willemsen G, Williams N, Willsey AJ, Witt SH, Xu W, Young AH, Yu TW, Zammit S, Zandi PP, Zhang P, Zitman FG, Zöllner S, Devlin B, Kelsoe JR, Sklar P, Daly MJ, O'Donovan MC, Craddock N, Sullivan PF, Smoller JW, Kendler KS, Wray NR (2013) Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics* 45:984–994 . doi: 10.1038/ng.2711

Leeson VC, Robbins TW, Matheson E, Hutton SB, Ron MA, Barnes TRE, Joyce EM (2009) Discrimination learning, reversal, and set-shifting in first-episode schizophrenia: stability over six years and specific associations with medication type and disorganization syndrome. *Biological Psychiatry* 66:586–93 . doi: 10.1016/j.biopsych.2009.05.016

Leisch F, Weingessel A, Maintainer KH (2015) Package "bindata" Title Generation of Artificial Binary Data

Lembke A, Gomez R, Tenakoon L, Keller J, Cohen G, Williams GH, Kraemer FB, Schatzberg AF (2013) The mineralocorticoid receptor agonist, fludrocortisone, differentially inhibits pituitary–adrenal activity in humans with psychotic major depression. *Psychoneuroendocrinology* 38:115–121 . doi: 10.1016/j.psyneuen.2012.05.006

Leszczyńska-Rodziewicz A, Maciukiewicz M, Szczepankiewicz A, Pogłodziński A, Hauser J (2013) Association between OPCRIT dimensions and polymorphisms of HPA axis genes in bipolar disorder. *Journal of Affective Disorders* 151:744–747 . doi: 10.1016/j.jad.2013.08.012

Lewis CM, Knight J (2012) Introduction to Genetic Association Studies. *Cold Spring Harbor Protocols* 2012:pdb.top068163-top068163 . doi: 10.1101/pdb.top068163

Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R news* 2:18--22

Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. *Nature Reviews Genetics* 16:321–332 . doi: 10.1038/nrg3920

Lu AT-H, Austin E, Bonner A, Huang H-H, Cantor RM (2014) Applications of

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

Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P (2004) Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics* 5:32 . doi: 10.1186/1471-2156-5-32

Ma SL, Tang NLS, Leung GTY, Fung AWT, Lam LCW (2014) Estrogen Receptor α Polymorphisms and the Risk of Cognitive Decline: A 2-Year Follow-Up Study. *The American Journal of Geriatric Psychiatry* 22:489–498 . doi: 10.1016/j.jagp.2012.08.006

Malhotra AK, Kestler LJ, Mazzanti C, Bates JA, Goldberg T, Goldman D (2002) A functional polymorphism in the COMT gene and performance on a test of prefrontal cognition. *American Journal of Psychiatry* 159:652–4 . doi: 10.1176/appi.ajp.159.4.652

Mancuso F, Horan WP, Kern RS, Green MF (2011) Social cognition in psychosis: multidimensional structure, clinical correlates, and relationship with functional outcome. *Schizophrenia Research* 125:143–51 . doi: 10.1016/j.schres.2010.11.007

Mardis ER (2011) A decade's perspective on DNA sequencing technology. *Nature* 470:198–203 . doi: 10.1038/nature09796

Martins-de-Souza D (2014) Proteomics, metabolomics, and protein interactomics in the characterization of the molecular features of major depressive disorder. *Dialogues in Clinical Neuroscience* 16:63–73

McGrath J, Saha S, Chant D, Welham J (2008) Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiologic Reviews* 30:67–76 . doi: 10.1093/epirev/mxn001

McQueen MB, Devlin B, Faraone S V, Nimgaonkar VL, Sklar P, Smoller JW, Abou Jamra R, Albus M, Bacanu S-A, Baron M, Barrett TB, Berrettini W, Blacker D, Byerley W, Cichon S, Coryell W, Craddock N, Daly MJ, Depaulo JR, Edenberg HJ, Foroud T, Gill M, Gilliam TC, Hamshere M, Jones I, Jones L, Juo S-H, Kelsoe JR, Lambert D, Lange C, Lerer B, Liu J, Maier W, Mackinnon JD, McInnis MG, McMahon FJ, Murphy DL, Nothen MM, Nurnberger JI, Pato CN, Pato MT, Potash JB, Propping P, Pulver AE, Rice JP, Rietschel M, Scheftner W, Schumacher J, Segurado R, Van Steen K, Xie W, Zandi PP, Laird NM (2005) Combined analysis from eleven linkage studies of bipolar disorder provides strong evidence of susceptibility loci on chromosomes 6q and 8q. *American Journal of Human Genetics* 77:582–95 . doi: 10.1086/491603

Meinshausen N, Bühlmann P (2010) Stability selection. *Journal of the Royal Statistical Society B (Statistical Methodology)* 72:417–473 . doi: 10.1111/j.1467-9868.2010.00740.x

Meng YA, Yu Y, Cupples LA, Farrer LA, Lunetta KL (2009) Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinformatics* 10:78 . doi: 10.1186/1471-2105-10-78

Michie D, Spiegelhalter DJ, Taylor CC (1994) Machine Learning, Neural and Statistical Classification

Mokhtari R, Lachman HM (2016) The Major Histocompatibility Complex (MHC) in Schizophrenia: A Review. *Journal of Clinical & Cellular Immunology* 7: . doi: 10.4172/2155-9899.1000479

Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26:445–55 . doi: 10.1093/bioinformatics/btp713

Moran PM, O'Tuathaigh CMP, Papaleo F, Waddington JL (2014) Dopaminergic function in relation to genes associated with risk for schizophrenia: translational mutant mouse models. *Progress in Brain Research* 211:79–112 . doi: 10.1016/B978-0-444-63425-2.00004-0

Mostafavi S, Battle A, Zhu X, Potash JB, Weissman MM, Shi J, Beckman K, Haudenschild C, McCormick C, Mei R, Gameroff MJ, Gindes H, Adams P, Goes FS, Mondimore FM, MacKinnon DF, Notes L, Schweizer B, Furman D, Montgomery SB, Urban AE, Koller D, Levinson DF (2014) Type I interferon signaling genes in recurrent major depression: increased expression detected by whole-blood RNA sequencing. *Molecular Psychiatry* 19:1267–1274 . doi: 10.1038/mp.2013.161

Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD (2008) Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genetic Epidemiology* 32:325–40 . doi: 10.1002/gepi.20307

MQ T mental health (2015) MQ lanscape analysis - UK Mental Health Research Funding

Müller MB, Zimmermann S, Sillaber I, Hagemeyer TP, Deussing JM, Timpl P, Kormann MSD, Droste SK, Kühn R, Reul JMHM, Holsboer F, Wurst W (2003) Limbic corticotropin-releasing hormone receptor 1 mediates anxiety-related behavior and hormonal adaptation to stress. *Nature Neuroscience* 6:1100–1107 . doi: 10.1038/nn1123

Murawiec S, Krysta K (2015) One of many lessons from the European Mental Health Integration Index. *Psychiatria Danubina* 27 Suppl 1:S97-102

National Cancer Research Institute (2013) Cancer research spend in the UK 2002-2011: An overview of the research funded by NCRI Partners. London

Nery FG, Borba EF, Hatch JP, Soares JC, Bonfá E, Neto FL (2007) Major depressive disorder and disease activity in systemic lupus erythematosus. *Comprehensive Psychiatry* 48:14–9 . doi: 10.1016/j.comppsych.2006.04.002

Neve RL, Harris P, Kosik KS, Kurnit DM, Donlon TA (1986) Identification of cDNA clones for the human microtubule-associated protein tau and chromosomal localization of the genes for tau and microtubule-associated protein 2. *Brain Research* 387:271–80

Nicodemus KK (2011) Letter to the editor: on the stability and ranking of predictors from random forest variable importance measures. *Briefings in Bioinformatics* 12:369–73 . doi: 10.1093/bib/bbr016

Nicodemus KK, Callicott JH, Higier RG, Luna A, Nixon DC, Lipska BK, Vakkalanka R, Giegling I, Rujescu D, St Clair D, Muglia P, Shugart YY, Weinberger DR (2010a) Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: biological validation with functional neuroimaging. *Human Genetics* 127:441–52 . doi: 10.1007/s00439-009-0782-y

Nicodemus KK, Law AJ, Radulescu E, Luna A, Kolachana B, Vakkalanka R, Rujescu D, Giegling I, Straub RE, McGee K, Gold B, Dean M, Muglia P, Callicott JH, Tan H-Y, Weinberger DR (2010b) Biological validation of increased schizophrenia risk with NRG1, ERBB4, and AKT1 epistasis via functional neuroimaging in healthy controls. *Archives of General Psychiatry* 67:991–1001 . doi: 10.1001/archgenpsychiatry.2010.117

Nicodemus KK, Malley JD (2009) Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics* 25:1884–90 . doi: 10.1093/bioinformatics/btp331

Nicodemus KK, Malley JD, Strobl C, Ziegler A (2010c) The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics* 11:110 . doi: 10.1186/1471-2105-11-110

Nicodemus KK, Marenco S, Batten AJ, Vakkalanka R, Egan MF, Straub RE, Weinberger DR (2008) Serious obstetric complications interact with hypoxia-regulated/vascular-expression genes to influence schizophrenia risk. *Molecular Psychiatry* 13:873–7 . doi: 10.1038/sj.mp.4002153

Niel C, Sinoquet C, Dina C, Rocheleau G (2015) A survey about methods dedicated to epistasis detection. *Frontiers in Genetics* 6:285 . doi: 10.3389/fgene.2015.00285

Nuevo R, Chatterji S, Verdes E, Naidoo N, Arango C, Ayuso-Mateos JL (2012) The continuum of psychotic symptoms in the general population: a cross-national study. *Schizophrenia Bulletin* 38:475–85 . doi: 10.1093/schbul/sbq099

O'Donoghue B, Lyne J, Madigan K, Lane A, Turner N, O'Callaghan E, Clarke M (2015) Environmental factors and the age at onset in first episode psychosis. *Schizophrenia Research* 168:106–12 . doi: 10.1016/j.schres.2015.07.004

O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskvina V, Nikolov I, Hamshere M, Carroll L, Georgieva L, Dwyer S, Holmans P, Marchini JL, Spencer CCA, Howie B, Leung H-T, Hartmann AM, Möller H-J, Morris DW, Shi Y, Feng G, Hoffmann P, Propping P, Vasilescu C, Maier W, Rietschel M, Zammit S, Schumacher J, Quinn EM, Schulze TG, Williams NM, Giegling I, Iwata N, Ikeda M, Darvasi A, Shifman S, He L, Duan J, Sanders AR, Levinson DF, Gejman P V, Gejman P V, Sanders AR, Duan J, Levinson DF, Buccola NG, Mowry BJ, Freedman R, Amin F, Black DW, Silverman JM, Byerley WF, Cloninger CR, Cichon S, Nöthen MM, Gill M, Corvin A, Rujescu D, Kirov G, Owen MJ (2008) Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nature Genetics* 40:1053–1055 . doi: 10.1038/ng.201

O'Rourke DH, Gottesman II, Suarez BK, Rice J, Reich T (1982) Refutation of the general single-locus model for the etiology of schizophrenia. *American Journal of Human Genetics* 34:630–49

Ohayon MM, Schatzberg AF (2002) Prevalence of depressive episodes with psychotic features in the general population. *American Journal of Psychiatry* 159:1855–61 . doi: 10.1176/appi.ajp.159.11.1855

Osterlund MK, Grandien K, Keller E, Hurd YL (2000) The human brain has distinct regional expression patterns of estrogen receptor alpha mRNA isoforms derived from alternative promoters. *Journal of Neurochemistry* 75:1390–7

Osterlund MK, Hurd YL (2001) Estrogen receptors in the human forebrain and the relation to neuropsychiatric disorders. *Progress in Neurobiology* 64:251–67

Owen MJ, Williams NM, O'Donovan MC (2004) The molecular genetics of schizophrenia: new findings promise new insights. *Molecular Psychiatry* 9:14–27 . doi: 10.1038/sj.mp.4001444

Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, Floyd E, Zhao H (2006) Pathway analysis using random forests classification and regression. *Bioinformatics* 22:2028–2036 . doi: 10.1093/bioinformatics/btl344

Pang H, Zhao H (2008) Building pathway clusters from Random Forests classification using class votes. *BMC Bioinformatics* 9:87 . doi: 10.1186/1471-2105-9-87

Papaleo F, Burdick MC, Callicott JH, Weinberger DR (2014) Epistatic interaction between COMT and DTNBP1 modulates prefrontal function in mice and in humans. *Molecular Psychiatry* 19:311–316 . doi: 10.1038/mp.2013.133

Park MY, Hastie T (2008) Penalized logistic regression for detecting gene interactions. *Biostatistics* 9:30–50 . doi: 10.1093/biostatistics/kxm010

Park S-C, Kim J-M, Jun T-Y, Lee M-S, Kim J-B, Yim H-W, Park YC (2016) How many different symptom combinations fulfil the diagnostic criteria for major depressive disorder? Results from the CRESCEND study. *Nordic Journal of Psychiatry* 1–6 . doi: 10.1080/08039488.2016.1265584

Parsons MJ, Mata I, Beperet M, Iribarren-Iriso F, Arroyo B, Sainz R, Arranz MJ, Kerwin R (2007) A dopamine D2 receptor gene-related polymorphism is associated with schizophrenia in a Spanish population isolate. *Psychiatric Genetics* 17:159–163 . doi: 10.1097/YPG.0b013e328017f8a4

Patel J, Shah S, Thakkar P, Kotecha K (2015) Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Systems with Applications* 42:259–268 . doi: 10.1016/j.eswa.2014.07.040

Pedregosa F, Alexandre Gramfort N, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Pedregosa F, Varoquaux G, Gramfort A, Thirion B, Prettenhofer P, Vanderplas J, Brucher M, Perrot an Edouard Duchesnay M, Matthieu Brucher A, Perrot M, Edouard Duchesnay CF (2011) Scikit-learn: Machine Learning in Python Gaël Varoquaux. *Journal of Machine Learning Research* 12:2825–2830

Perlman WR, Tomaskovic-Crook E, Montague DM, Webster MJ, Rubinow DR, Kleinman JE, Weickert CS (2005) Alteration in Estrogen Receptor α mRNA Levels in Frontal Cortex and Hippocampus of Patients with Major Mental Illness. *Biological Psychiatry* 58:812–824 . doi: 10.1016/j.biopsych.2005.04.047

Perlman WR, Webster MJ, Kleinman JE, Weickert CS (2004) Reduced glucocorticoid and estrogen receptor alpha messenger ribonucleic acid levels in the amygdala of patients with major mental illness. *Biological Psychiatry* 56:844–852 . doi: 10.1016/j.biopsych.2004.09.006

Petralia F, Wang P, Yang J, Tu Z (2015) Integrative random forest for gene regulatory network inference. *Bioinformatics* 31:i197-205. doi: 10.1093/bioinformatics/btv268

Pettersson E, Larsson H, Lichtenstein P (2016) Common psychiatric disorders share the same genetic origin: a multivariate sibling study of the Swedish population. *Molecular Psychiatry* 21:717–21 . doi: 10.1038/mp.2015.116

Polanczyk G, Caspi A, Williams B, Price TS, Danese A, Sugden K, Uher R, Poulton R, Moffitt TE (2009) Protective Effect of CRHR1 Gene Variants on the Development of Adult Depression Following Childhood Maltreatment. *Archives of General Psychiatry* 66:978 . doi: 10.1001/archgenpsychiatry.2009.114

Polanczyk G, Moffitt TE, Arseneault L, Cannon M, Ambler A, Keefe RSE, Houts R, Odgers CL, Caspi A (2010) Etiological and clinical features of childhood psychotic symptoms: results from a birth cohort. *Archives of General Psychiatry* 67:328–38 . doi: 10.1001/archgenpsychiatry.2010.14

Polderman TJC, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM, Posthuma D (2015) Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics* 47:702–9 . doi: 10.1038/ng.3285

Poorkaj P, Kas A, D'Souza I, Zhou Y, Pham Q, Stone M, Olson M V, Schellenberg GD (2001) A genomic sequence analysis of the mouse and human microtubule-associated protein tau. *Mammalian Genome* 12:700–12

Porteous DJ, Thomson PA, Millar JK, Evans KL, Hennah W, Soares DC, McCarthy S, McCombie WR, Clapcote SJ, Korth C, Brandon NJ, Sawa A, Kamiya A, Roder JC, Lawrie SM, McIntosh AM, St Clair D, Blackwood DH (2014) DISC1 as a genetic risk factor for schizophrenia and related major mental illness: response to Sullivan. *Molecular Psychiatry* 19:141–143 . doi: 10.1038/mp.2013.160

Power RA, Tansey KE, Buttenschøn HN, Cohen-Woods S, Bigdeli T, Hall LS, Kutalik Z, Lee SH, Ripke S, Steinberg S, Teumer A, Viktorin A, Wray NR, Arolt V, Baune BT, Boomsma DI, Børglum AD, Byrne EM, Castelao E, Craddock N, Craig IW, Dannlowski U, Deary IJ, Degenhardt F, Forstner AJ, Gordon SD, Grabe HJ, Grove J, Hamilton SP, Hayward C, Heath AC, Hocking LJ, Homuth G, Hottenga JJ, Kloiber S, Krogh J, Landén M, Lang M, Levinson DF, Lichtenstein P, Lucae S, MacIntyre DJ, Madden P, Magnusson PKE, Martin NG, McIntosh AM, Middeldorp CM, Milaneschi Y, Montgomery GW, Mors O, Müller-Myhsok B, Nyholt DR, Oskarsson H, Owen MJ, Padmanabhan S, Penninx BWJH, Pergadia ML, Porteous DJ, Potash JB, Preisig M, Rivera M, Shi J, Shyn SI, Sigurdsson E, Smit JH, Smith BH, Stefansson H, Stefansson K, Strohmaier J, Sullivan PF, Thomson P, Thorgeirsson TE, Van der Auwera S, Weissman MM, Breen G, Lewis CM (2017) Genome-wide Association for Major Depression Through Age at Onset Stratification: Major Depressive Disorder

Working Group of the Psychiatric Genomics Consortium. *Biological Psychiatry* 81:325–335 . doi: 10.1016/j.biopsych.2016.05.010

Prabakaran S, Swatton JE, Ryan MM, Huffaker SJ, Huang J-J, Griffin JL, Wayland M, Freeman T, Dudbridge F, Lilley KS, Karp NA, Hester S, Tkachev D, Mimmack ML, Yolken RH, Webster MJ, Torrey EF, Bahn S (2004) Mitochondrial dysfunction in Schizophrenia: evidence for compromised brain metabolism and oxidative stress. *Molecular Psychiatry* 9:684–97, 643 . doi: 10.1038/sj.mp.4001511

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38:904–909 . doi: 10.1038/ng1847

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

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P, Purcell (Leader) SM, Stone JL, Sullivan PF, Ruderfer DM, McQuillin A, Morris DW, O'Dushlaine CT, Corvin A, Holmans PA, O'Donovan MC, Sklar P, Wray NR, Macgregor S, Sklar P, Sullivan PF, O'Donovan MC, Visscher PM, Gurling H, Blackwood DHR, Corvin A, Craddock NJ, Gill M, Hultman CM, Kirov GK, Lichtenstein P, McQuillin A, Muir WJ, O'Donovan MC, Owen MJ, Pato CN, Purcell SM, Scolnick EM, St Clair D, Stone JL, Sullivan PF, Sklar (Leader) P, O'Donovan MC, Kirov GK, Craddock NJ, Holmans PA, Williams NM, Georgieva L, Nikolov I, Norton N, Williams H, Toncheva D, Milanova V, Owen MJ, Hultman CM, Lichtenstein P, Thelander EF, Sullivan P, Morris DW, O'Dushlaine CT, Kenny E, Quinn EM, Gill M, Corvin A, McQuillin A, Choudhury K, Datta S, Pimm J, Thirumalai S, Puri V, Krasucki R, Lawrence J, Quested D, Bass N, Gurling H, Crombie C, Fraser G, Leh Kuan S, Walker N, St Clair D, Blackwood DHR, Muir WJ, McGhee KA, Pickard B, Malloy P, Maclean AW, Van Beck M, Wray NR, Macgregor S, Visscher PM, Pato MT, Medeiros H, Middleton F, Carvalho C, Morley C, Fanous A, Conti D, Knowles JA, Paz Ferreira C, Macedo A, Helena Azevedo M, Pato CN, Stone JL, Ruderfer DM, Kirby AN, Ferreira MAR, Daly MJ, Purcell SM, Sklar P, Purcell SM, Stone JL, Chambert K, Ruderfer DM, Kuruvilla F, Gabriel SB, Ardlie K, Moran JL, Daly MJ, Scolnick EM, Sklar P (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460:748 . doi: 10.1038/nature08185

Refojo D, Schweizer M, Kuehne C, Ehrenberg S, Thoeringer C, Vogl AM, Dedic N, Schumacher M, von Wolff G, Avrabos C, Touma C, Engblom D, Schutz G, Nave K-A, Eder M, Wotjak CT, Sillaber I, Holsboer F, Wurst W, Deussing JM (2011) Glutamatergic and Dopaminergic Neurons Mediate Anxiogenic and Anxiolytic

Effects of CRHR1. *Science* 333:1903–1907 . doi: 10.1126/science.1202107

Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, Kupfer DJ (2013) DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses. *American Journal of Psychiatry* 170:59–70 . doi: 10.1176/appi.ajp.2012.12070999

Rettig WJ, Triche TJ, Garin-Chesa P (1989) Stimulation of human neuronectin secretion by brain-derived growth factors. *Brain Research* 487:171–7

Riley B, Kendler KS (2006) Molecular genetic studies of schizophrenia. *European Journal of Human Genetics* 14:669–680 . doi: 10.1038/sj.ejhg.5201571

Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H, Pers TH, Agartz I, Agerbo E, Albus M, Alexander M, Amin F, Bacanu SA, Begemann M, Belliveau Jr RA, Bene J, Bergen SE, Bevilacqua E, Bigdeli TB, Black DW, Bruggeman R, Buccola NG, Buckner RL, Byerley W, Cahn W, Cai G, Campion D, Cantor RM, Carr VJ, Carrera N, Catts S V., Chambert KD, Chan RCK, Chen RYL, Chen EYH, Cheng W, Cheung EFC, Ann Chong S, Robert Cloninger C, Cohen D, Cohen N, Cormican P, Craddock N, Crowley JJ, Curtis D, Davidson M, Davis KL, Degenhardt F, Del Favero J, Demontis D, Dikeos D, Dinan T, Djurovic S, Donohoe G, Drapeau E, Duan J, Dudbridge F, Durmishi N, Eichhammer P, Eriksson J, Escott-Price V, Essioux L, Fanous AH, Farrell MS, Frank J, Franke L, Freedman R, Freimer NB, Friedl M, Friedman JI, Fromer M, Genovese G, Georgieva L, Giegling I, Giusti-Rodríguez P, Godard S, Goldstein JI, Golimbet V, Gopal S, Gratten J, de Haan L, Hammer C, Hamshere ML, Hansen M, Hansen T, Haroutunian V, Hartmann AM, Henskens FA, Herms S, Hirschhorn JN, Hoffmann P, Hofman A, Hollegaard M V., Hougaard DM, Ikeda M, Joa I, Julià A, Kahn RS, Kalaydjieva L, Karachanak-Yankova S, Karjalainen J, Kavanagh D, Keller MC, Kennedy JL, Khrunin A, Kim Y, Klovins J, Knowles JA, Konte B, Kucinskas V, Ausrele Kucinskiene Z, Kuzelova-Ptackova H, Kähler AK, Laurent C, Lee Chee Keong J, Hong Lee S, Legge SE, Lerer B, Li M, Li T, Liang K-Y, Lieberman J, Limborska S, Loughland CM, Lubinski J, Lönnqvist J, Macek Jr M, Magnusson PKE, Maher BS, Maier W, Mallet J, Marsal S, Mattheisen M, Mattingsdal M, McCarley RW, McDonald C, McIntosh AM, Meier S, Meijer CJ, Melegh B, Melle I, Mesholam-Gately RI, Metspalu A, Michie PT, Milani L, Milanova V, Mokrab Y, Morris DW, Mors O, Murphy KC, Murray RM, Myin-Germeys I, Müller-Myhsok B, Nelis M, Nenadic I, Nertney DA, Nestadt G, Nicodemus KK, Nikitina-Zake L, Nisenbaum L, Nordin A, O'Callaghan E, O'Dushlaine C, O'Neill FA, Oh S-Y, Olincy A, Olsen L, Van Os J, Endophenotypes International Consortium P, Pantelis C, Papadimitriou GN, Papiol S, Parkhomenko E, Pato MT, Paunio T, Pejovic-Milovancevic M, Perkins DO, Pietiläinen O, Pimm J, Pocklington AJ, Powell J, Price A, Pulver AE, Purcell SM, Quested D, Rasmussen HB, Reichenberg A, Reimers MA, Richards AL, Roffman JL, Roussos P, Ruderfer DM, Salomaa V, Sanders AR, Schall U,

Schubert CR, Schulze TG, Schwab SG, Scolnick EM, Scott RJ, Seidman LJ, Shi J, Sigurdsson E, Silagadze T, Silverman JM, Sim K, Slominsky P, Smoller JW, So H-C, Spencer CA, Stahl EA, Stefansson H, Steinberg S, Stogmann E, Straub RE, Strengman E, Strohmaier J, Scott Stroup T, Subramaniam M, Suvisaari J, Svrakic DM, Szatkiewicz JP, Söderman E, Thirumalai S, Toncheva D, Tosato S, Veijola J, Waddington J, Walsh D, Wang D, Wang Q, Webb BT, Weiser M, Wildenauer DB, Williams NM, Williams S, Witt SH, Wolen AR, Wong EHM, Wormley BK, Simon Xi H, Zai CC, Zheng X, Zimprich F, Wray NR, Stefansson K, Visscher PM, Trust Case-Control Consortium W, Adolfsson R, Andreassen OA, Blackwood DHR, Bramon E, Buxbaum JD, Børglum AD, Cichon S, Darvasi A, Domenici E, Ehrenreich H, Esko T, Gejman P V., Gill M, Gurling H, Hultman CM, Iwata N, Jablensky A V., Jönsson EG, Kendler KS, Kirov G, Knight J, Lencz T, Levinson DF, Li QS, Liu J, Malhotra AK, McCarroll SA, McQuillin A, Moran JL, Mortensen PB, Mowry BJ, Nöthen MM, Ophoff RA, Owen MJ, Palotie A, Pato CN, Petryshen TL, Posthuma D, Rietschel M, Riley BP, Rujescu D, Sham PC, Sklar P, St Clair D, Weinberger DR, Wendland JR, Werge T, Daly MJ, Sullivan PF, O'Donovan MC (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511:421–427 . doi: 10.1038/nature13595

Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA, Lin D-Y, Duan J, Ophoff RA, Andreassen OA, Scolnick E, Cichon S, St. Clair D, Corvin A, Gurling H, Werge T, Rujescu D, Blackwood DHR, Pato CN, Malhotra AK, Purcell S, Dudbridge F, Neale BM, Rossin L, Visscher PM, Posthuma D, Ruderfer DM, Fanous A, Stefansson H, Steinberg S, Mowry BJ, Golimbet V, De Hert M, Jönsson EG, Bitter I, Pietiläinen OPH, Collier DA, Tosato S, Agartz I, Albus M, Alexander M, Amdur RL, Amin F, Bass N, Bergen SE, Black DW, Børglum AD, Brown MA, Bruggeman R, Buccola NG, Byerley WF, Cahn W, Cantor RM, Carr VJ, Catts S V, Choudhury K, Cloninger CR, Cormican P, Craddock N, Danoy PA, Datta S, de Haan L, Demontis D, Dikeos D, Djurovic S, Donnelly P, Donohoe G, Duong L, Dwyer S, Fink-Jensen A, Freedman R, Freimer NB, Friedl M, Georgieva L, Giegling I, Gill M, Glenthøj B, Godard S, Hamshere M, Hansen M, Hansen T, Hartmann AM, Henskens FA, Hougaard DM, Hultman CM, Ingason A, Jablensky A V, Jakobsen KD, Jay M, Jürgens G, Kahn RS, Keller MC, Kenis G, Kenny E, Kim Y, Kirov GK, Konnerth H, Konte B, Krabbendam L, Krasucki R, Lasseter VK, Laurent C, Lawrence J, Lencz T, Lerer FB, Liang K-Y, Lichtenstein P, Lieberman JA, Linszen DH, Lönnqvist J, Loughland CM, Maclean AW, Maher BS, Maier W, Mallet J, Malloy P, Mattheisen M, Mattingsdal M, McGhee KA, McGrath JJ, McIntosh A, McLean DE, McQuillin A, Melle I, Michie PT, Milanova V, Morris DW, Mors O, Mortensen PB, Moskvina V, Muglia P, Myin-Germeys I, Nertney DA, Nestadt G, Nielsen J, Nikolov I, Nordentoft M, Norton N, Nöthen MM, O'Dushlaine CT, Olincy A, Olsen L, O'Neill FA, Ørntoft TF, Owen MJ, Pantelis C, Papadimitriou G, Pato MT, Peltonen L, Petursson H, Pickard B, Pimm J, Pulver AE, Puri V, Quested D, Quinn EM, Rasmussen HB, Réthelyi JM, Ribble R, Rietschel M, Riley BP, Ruggeri M, Schall U, Schulze TG, Schwab SG, Scott RJ, Shi J, Sigurdsson E, Silverman JM, Spencer CCA, Stefansson K, Strange A, Strengman E, Stroup TS, Suvisaari J, Terenius L, Thirumalai S, Thygesen JH, Timm S,

Toncheva D, van den Oord E, van Os J, van Winkel R, Veldink J, Walsh D, Wang AG, Wiersma D, Wildenauer DB, Williams HJ, Williams NM, Wormley B, Zammit S, Sullivan PF, O'Donovan MC, Daly MJ, Gejman P V (2011) Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics* 43:969–976 . doi: 10.1038/ng.940

Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM, Breen G, Byrne EM, Blackwood DHR, Boomsma DI, Cichon S, Heath AC, Holsboer F, Lucae S, Madden PAF, Martin NG, McGuffin P, Muglia P, Noethen MM, Penninx BP, Pergadia ML, Potash JB, Rietschel M, Lin D, Müller-Myhsok B, Shi J, Steinberg S, Grabe HJ, Lichtenstein P, Magnusson P, Perlis RH, Preisig M, Smoller JW, Stefansson K, Uher R, Kutalik Z, Tansey KE, Teumer A, Viktorin A, Barnes MR, Bettecken T, Binder EB, Breuer R, Castro VM, Churchill SE, Coryell WH, Craddock N, Craig IW, Czamara D, De Geus EJ, Degenhardt F, Farmer AE, Fava M, Frank J, Gainer VS, Gallagher PJ, Gordon SD, Goryachev S, Gross M, Guipponi M, Henders AK, Herms S, Hickie IB, Hoefels S, Hoogendijk W, Hottenga JJ, Iosifescu D V, Ising M, Jones I, Jones L, Jung-Ying T, Knowles JA, Kohane IS, Kohli MA, Korszun A, Landen M, Lawson WB, Lewis G, MacIntyre D, Maier W, Mattheisen M, McGrath PJ, McIntosh A, McLean A, Middeldorp CM, Middleton L, Montgomery GM, Murphy SN, Nauck M, Nolen WA, Nyholt DR, O'Donovan M, Oskarsson H, Pedersen N, Scheftner WA, Schulz A, Schulze TG, Shyn SI, Sigurdsson E, Slager SL, Smit JH, Stefansson H, Steffens M, Thorgeirsson T, Tozzi F, Treutlein J, Uhr M, van den Oord EJCG, Van Grootheest G, Völzke H, Weilburg JB, Willemsen G, Zitman FG, Neale B, Daly M, Levinson DF, Sullivan PF (2013) A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry* 18:497–511 . doi: 10.1038/mp.2012.21

Risch N (1990) Genetic linkage and complex diseases, with special reference to psychiatric disorders. *Genetic Epidemiology* 7:3-16-45 . doi: 10.1002/gepi.1370070103

Ritchie MD, Motsinger AA, Bush WS, Coffey CS, Moore JH (2007) Genetic Programming Neural Networks: A Powerful Bioinformatics Tool for Human Genetics. *Applied Soft Computing* 7:471–479 . doi: 10.1016/j.asoc.2006.01.013

Rosa A, Peralta V, Cuesta MJ, Zarzuela A, Serrano F, Martínez-Larrea A, Fañanás L (2004) New evidence of association between COMT gene and prefrontal neurocognitive function in healthy individuals from sibling pairs discordant for psychosis. *American Journal of Psychiatry* 161:1110–2 . doi: 10.1176/appi.ajp.161.6.1110

Rose EJ, Morris DW, Fahey C, Robertson IH, Greene C, O'Doherty J, Newell FN, Garavan H, McGrath J, Bokde A, Tropea D, Gill M, Corvin AP, Donohoe G (2012) The Effect of the Neurogranin Schizophrenia Risk Variant rs12807809 on Brain Structure and Function. *Twin Research and Human Genetics* 15:296–303 .

doi: 10.1017/thg.2012.7

Ross CA, Margolis RL, Reading SAJ, Pletnikov M, Coyle JT (2006) Neurobiology of schizophrenia. *Neuron* 52:139–53 . doi: 10.1016/j.neuron.2006.09.015

Rothschild AJ (2013) Challenges in the treatment of major depressive disorder with psychotic features. *Schizophrenia Bulletin* 39:787–96 . doi: 10.1093/schbul/sbt046

Różycka A, Słopień R, Słopień A, Dorszewska J, Seremak-Mrozikiewicz A, Lianeri M, Maciukiewicz M, Warenik-Szymankiewicz A, Grzelak T, Kurzawińska G, Drews K, Klejewski A, Jagodziński PP (2016) The MAOA, COMT, MTHFR and ESR1 gene polymorphisms are associated with the risk of depression in menopausal women. *Maturitas* 84:42–54 . doi: 10.1016/j.maturitas.2015.10.011

Ruderfer DM, Fanous AH, Ripke S, McQuillin A, Amdur RL, Schizophrenia Working Group of Psychiatric Genomics Consortium, Bipolar Disorder Working Group of Psychiatric Genomics Consortium, Cross-Disorder Working Group of Psychiatric Genomics Consortium, Gejman P V, O'Donovan MC, Andreassen OA, Djurovic S, Hultman CM, Kelsoe JR, Jamain S, Landén M, Leboyer M, Nimgaonkar V, Nurnberger J, Smoller JW, Craddock N, Corvin A, Sullivan PF, Holmans P, Sklar P, Kendler KS (2014) Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Molecular Psychiatry* 19:1017–24 . doi: 10.1038/mp.2013.138

Ryan J, Ancelin M-L (2012) Polymorphisms of Estrogen Receptors and Risk of Depression. *Drugs* 72:1725–1738 . doi: 10.2165/11635960-000000000-00000

Ryan J, Scali J, Carrière I, Peres K, Rouaud O, Scarabin P-Y, Ritchie K, Ancelin M-L (2012) Estrogen receptor alpha gene variants and major depressive episodes. *Journal of Affective Disorders* 136:1222–1226 . doi: 10.1016/j.jad.2011.10.010

Saeys Y, Abeel T, Peer Y Van de (2008) Robust Feature Selection Using Ensemble Feature Selection Techniques. *Springer* 313–325

Schatzberg AF, Keller J, Tennakoon L, Lembke A, Williams G, Kraemer FB, Sarginson JE, Lazzeroni LC, Murphy GM (2014) HPA axis genetic variation, cortisol and psychosis in major depression. *Molecular Psychiatry* 19:220–227 . doi: 10.1038/mp.2013.129

Schobel SA, Chaudhury NH, Khan UA, Paniagua B, Styner MA, Asllani I, Inbar BP, Corcoran CM, Lieberman JA, Moore H, Small SA (2013) Imaging Patients with Psychosis and a Mouse Model Establishes a Spreading Pattern of Hippocampal Dysfunction and Implicates Glutamate as a Driver. *Neuron* 78:81–93 . doi: 10.1016/j.neuron.2013.02.011

Schwarz DF, König IR, Ziegler A (2010) On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics* 26:1752–8 . doi: 10.1093/bioinformatics/btq257

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

Segal MR (2004) Machine Learning Benchmarks and Random Forest Regression. *CBMB Working Paper*

Sennvik K, Boekhoorn K, Lasrado R, Terwel D, Verhaeghe S, Korr H, Schmitz C, 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 Journal* 21:2149–2161 . doi: 10.1096/fj.06-7735com

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

Shih RA, Belmonte PL, Zandi PP (2004) A review of the evidence from family, twin 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,

Lawson WB, Nwulia EA, Hipolito M, Coryell W, Rice J, Byerley W, McMahon FJ, Schulze TG, Berrettini W, Lohoff FW, Potash JB, Mahon PB, McInnis MG, Zöllner S, Zhang P, Craig DW, Szelinger S, Barrett TB, Breuer R, Meier S, Strohmaier J, Witt SH, Tozzi F, Farmer A, McGuffin P, Strauss J, Xu W, Kennedy JL, Vincent JB, Matthews K, Day R, Ferreira MA, O'Dushlaine C, Perlis R, Raychaudhuri S, Ruderfer D, Hyoun PL, Smoller JW, Li J, Absher D, Thompson RC, Meng FG, Schatzberg AF, Bunney WE, Barchas JD, Jones EG, Watson SJ, Myers RM, Akil H, Boehnke M, Chambert K, Moran J, Scolnick E, Djurovic S, Melle I, Morken G, Gill M, Morris D, Quinn E, Mühleisen TW, Degenhardt FA, Mattheisen M, Schumacher J, Maier W, Steffens M, Propping P, Nöthen MM, Anjorin A, Bass N, Gurling H, Kandaswamy R, Lawrence J, McGhee K, McIntosh A, McLean AW, Muir WJ, Pickard BS, Breen G, St. Clair D, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Jones IR, Kirov G, Moskvina V, Nikolov I, O'Donovan MC, Owen MJ, Collier DA, Elkin A, Williamson R, Young AH, Ferrier IN, Stefansson K, Stefansson H, Þorgeirsson Þ, Steinberg S, Gustafsson Ó, Bergen SE, Nimgaonkar V, Hultman C, Landén M, Lichtenstein P, Sullivan P, Schalling M, Osby U, Backlund L, Frisén L, Langstrom N, Jamain S, Leboyer M, Etain B, Bellivier F, Petursson H, Sigur∂sson E, Müller-Mysok B, Lucae S, Schwarz M, Schofield PR, Martin N, Montgomery GW, Lathrop M, Óskarsson H, Bauer M, Wright A, Mitchell PB, Hautzinger M, Reif A, Kelsoe JR, Purcell SM (2011) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics* 43:977–983 . doi: 10.1038/ng.943

Smith DJ, Nicholl BI, Cullen B, Martin D, Ul-Haq Z, Evans J, Gill JMR, Roberts B, Gallacher J, Mackay D, Hotopf M, Deary I, Craddock N, Pell JP (2013) Prevalence and Characteristics of Probable Major Depression and Bipolar Disorder within UK Biobank: Cross-Sectional Study of 172,751 Participants. *PLoS One* 8:e75362 . doi: 10.1371/journal.pone.0075362

Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietiläinen OPH, Mors O, Mortensen PB, Sigurdsson E, Gustafsson O, Nyegaard M, Tuulio-Henriksson A, Ingason A, Hansen T, Suvisaari J, Lonnqvist J, Paunio T, Børglum AD, Hartmann A, Fink-Jensen A, Nordentoft M, Hougaard D, Norgaard-Pedersen B, Böttcher Y, Olesen J, Breuer R, Möller H-J, Giegling I, Rasmussen HB, Timm S, Mattheisen M, Bitter I, Réthelyi JM, Magnusdottir BB, Sigmundsson T, Olason P, Masson G, Gulcher JR, Haraldsson M, Fossdal R, Thorgeirsson TE, Thorsteinsdottir U, Ruggeri M, Tosato S, Franke B, Strengman E, Kiemeney LA, GROUP† RS, Melle I, Djurovic S, Abramova L, Kaleda V, Sanjuan J, de Frutos R, Bramon E, Vassos E, Fraser G, Ettinger U, Picchioni M, Walker N, Toulopoulou T, Need AC, Ge D, Lim Yoon J, Shianna K V., Freimer NB, Cantor RM, Murray R, Kong A, Golimbet V, Carracedo A, Arango C, Costas J, Jönsson EG, Terenius L, Agartz I, Petursson H, Nöthen MM, Rietschel M, Matthews PM, Muglia P, Peltonen L, St Clair D, Goldstein DB, Stefansson K, Collier DA, Kahn RS, Linszen DH, van Os J, Wiersma D, Bruggeman R, Cahn W, de Haan L, Krabbendam L, Myin-Germeys I (2009) Common variants conferring risk of schizophrenia. *Nature* 460:744 . doi:

10.1038/nature08186

Steinberg S, de Jong S, Irish Schizophrenia Genomics Consortium, Andreassen OA, Werge T, Børglum AD, Mors O, Mortensen PB, Gustafsson O, Costas J, Pietiläinen OPH, Demontis D, Papiol S, Huttenlocher J, Mattheisen M, Breuer R, Vassos E, Giegling I, Fraser G, Walker N, Tuulio-Henriksson A, Suvisaari J, Lönnqvist J, Paunio T, Agartz I, Melle I, Djurovic S, Strengman E, GROUP, Jürgens G, Glenthøj B, Terenius L, Hougaard DM, Ørntoft T, Wiuf C, Didriksen M, Hollegaard M V, Nordentoft M, van Winkel R, Kenis G, Abramova L, Kaleda V, Arrojo M, Sanjuán J, Arango C, Sperling S, Rossner M, Ribolsi M, Magni V, Siracusano A, Christiansen C, Kiemeney LA, Veldink J, van den Berg L, Ingason A, Muglia P, Murray R, Nöthen MM, Sigurdsson E, Petursson H, Thorsteinsdottir U, Kong A, Rubino IA, De Hert M, Réthelyi JM, Bitter I, Jönsson EG, Golimbet V, Carracedo A, Ehrenreich H, Craddock N, Owen MJ, O'Donovan MC, Wellcome Trust Case Control Consortium 2, Ruggeri M, Tosato S, Peltonen L, Ophoff RA, Collier DA, St Clair D, Rietschel M, Cichon S, Stefansson H, Rujescu D, Stefansson K (2011) Common variants at VRK2 and TCF4 conferring risk of schizophrenia. *Human Molecular Genetics* 20:4076–81 . doi: 10.1093/hmg/ddr325

Stepniak B, Papiol S, Hammer C, Ramin A, Everts S, Hennig L, Begemann M, Ehrenreich H (2014) Accumulated environmental risk determining age at schizophrenia onset: a deep phenotyping-based study. *The lancet Psychiatry* 1:444–53 . doi: 10.1016/S2215-0366(14)70379-7

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* 100:9440–5 . doi: 10.1073/pnas.1530509100

Straub RE, Weinberger DR (2006) Schizophrenia genes - famine to feast. *Biological Psychiatry* 60:81–3 . doi: 10.1016/j.biopsych.2006.06.002

Strobl C, Boulesteix A-L, Augustin T (2007a) Unbiased split selection for classification trees based on the Gini Index. *Computational Statistics & Data Analysis* 52:483–501 . doi: 10.1016/j.csda.2006.12.030

Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. *BMC Bioinformatics* 9:307 . doi: 10.1186/1471-2105-9-307

Strobl C, Boulesteix A-L, Zeileis A, Hothorn T (2007b) Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 8:25 . doi: 10.1186/1471-2105-8-25

Strobl C, Malley J, Tutz G (2009) An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging,

and random forests. *Psychological Methods* 14:323–48 . doi: 10.1037/a0016973

Sullivan PF, Daly MJ, O'Donovan M (2012) Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics* 13:537–51 . doi: 10.1038/nrg3240

Sutton CD (2005) Classification and Regression Trees, Bagging, and Boosting. doi: 10.1016/S0169-7161(04)24011-1

Sweilam NH, Tharwat AA, Abdel Moniem NK (2010) Support vector machine for diagnosis cancer disease: A comparative study. *Egyptian Informatics Journal* 11:81–92 . doi: 10.1016/j.eij.2010.10.005

Tandon R, Carpenter WT, Jr (2012) DSM-5 status of psychotic disorders: 1 year prepublication. *Schizophrenia Bulletin* 38:369–70 . doi: 10.1093/schbul/sbs048

Tandon R, Gaebel W, Barch DM, Bustillo J, Gur RE, Heckers S, Malaspina D, Owen MJ, Schultz S, Tsuang M, Van Os J, Carpenter W (2013) Definition and description of schizophrenia in the DSM-5. *Schizophrenia Research* 150:3–10 . doi: 10.1016/j.schres.2013.05.028

Teare MD, Heighway J, Santibáñez Koref MF (2006) An expectation-maximization algorithm for the analysis of allelic expression imbalance. *American Journal of Human Genetics* 79:539–43 . doi: 10.1086/506968

Tenesa A, Haley CS (2013) The heritability of human disease: estimation, uses and abuses. *Nature Reviews Genetics* 14:139–149 . doi: 10.1038/nrg3377

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526:68–74 . doi: 10.1038/nature15393

The National Academies Collection: Reports funded by National Institutes of Health. (2015) Enabling Discovery, Development, and Translation of Treatments for Cognitive Dysfunction in Depression. *National Academies Press (US)*

Timpl P, Spanagel R, Sillaber I, Kresse A, Reul JMHM, Stalla GK, Blanquet V, Steckler T, Holsboer F, Wurst W (1998) Impaired stress response and reduced anxiety in mice lacking a functional corticotropin-releasing hormone receptor 1. *Nature Genetics* 19:162–166 . doi: 10.1038/520

Tin Kam Ho (1998) The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20:832–844 . doi: 10.1109/34.709601

Tse S, Davidson L, Chung K-F, Yu CH, Ng KL, Tsoi E (2015) Logistic regression analysis of psychosocial correlates associated with recovery from schizophrenia

in a Chinese community. *The International Journal of Social Psychiatry* 61:50–7 . doi: 10.1177/0020764014535756

van Os J, Hanssen M, Bijl R V, Vollebergh W (2001) Prevalence of psychotic disorder and community level of psychotic symptoms: an urban-rural comparison. *Archives of General Psychiatry* 58:663–8

Verpillat P, Camuzat A, Hannequin D, Thomas-Anterion C, Puel M, Belliard S, Dubois B, Didic M, Michel B-F, Lacomblez L, Moreaud O, Sellal F, Golfier V, Campion D, Clerget-Darpoux F, Brice A (2002) Association between the extended tau haplotype and frontotemporal dementia. *Archives of Neurology* 59:935–9

Vickers AJ (2005) Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Medical Research Methodology* 5:35 . doi: 10.1186/1471-2288-5-35

Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics* 9:255–266 . doi: 10.1038/nrg2322

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research* 31:258–61

Walesby K, Harrison J, Russ T (2017) What big data could achieve in Scotland. *The Journal of the Royal College of Physicians of Edinburgh* 474997:114–9 . doi: 10.4997/JrCPe.2017.201

Wang X, Xing EP, Schaid DJ (2015) Kernel methods for large-scale genomic data analysis. *Briefings in Bioinformatics* 16:183–92 . doi: 10.1093/bib/bbu024

Watson S, Gallagher P, Dougall D, Porter R, Moncrieff J, Ferrier IN, Young AH (2014) Childhood trauma in bipolar disorder. *The Australian and New Zealand Journal of Psychiatry* 48:564–70 . doi: 10.1177/0004867413516681

Weatherall DJ (2000) Single gene disorders or complex traits: lessons from the thalassaemias and other monogenic diseases. *BMJ* 321:1117–20

Weinberger DR (1987) Implications of normal brain development for the pathogenesis of schizophrenia. *Archives of General Psychiatry* 44:660–9

Whitlock MC (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology* 18:1368–1373 . doi: 10.1111/j.1420-9101.2005.00917.x

Widiger TA, American Psychiatric Association. Task Force on DSM-IV. (1994) *DSM-IV sourcebook*. 1st ed. Published by the American Psychiatric Association, Washington DC

Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, Huebner M, Biernacka JM (2012) SNP interaction detection with Random Forests in high-dimensional genetic data. *BMC Bioinformatics* 13:164 . doi: 10.1186/1471-2105-13-164

Wright MN, Ziegler A, König IR (2016) Do little interactions get lost in dark random forests? *BMC Bioinformatics* 17:145 . doi: 10.1186/s12859-016-0995-8

Wu Q, Ye Y, Liu Y, Ng MK (2012) SNP Selection and Classification of Genome-Wide SNP Data Using Stratified Sampling Random Forests. *IEEE Transactions on NanoBioscience* 11:216–227 . doi: 10.1109/TNB.2012.2214232

Wykes T, Haro JM, Belli SR, Obradors-Tarragó C, Arango C, Ayuso-Mateos JL, Bitter I, Brunn M, Chevreul K, Demotes-Mainard J, Elfeddali I, Evans-Lacko S, Fiorillo A, Forsman AK, Hazo J-B, Kuepper R, Knappe S, Leboyer M, Lewis SW, Linszen D, Luciano M, Maj M, McDaid D, Miret M, Papp S, Park A-L, Schumann G, Thornicroft G, van der Feltz-Cornelis C, van Os J, Wahlbeck K, Walker-Tilley T, Wittchen H-U (2015) Mental health research priorities for Europe. *The Lancet Psychiatry* 2:1036–1042 . doi: 10.1016/S2215-0366(15)00332-6

Yang P, Hwa Yang Y, B. Zhou B, Y. Zomaya A (2010) A Review of Ensemble Methods in Bioinformatics. *Current Bioinformatics* 5:296–308 . doi: 10.2174/157489310794072508

Yang Z-T, Yeo S-Y, Yin Y-X, Lin Z-H, Lee H-M, Xuan Y-H, Cui Y, Kim S-H (2016) Tenascin-C, a Prognostic Determinant of Esophageal Squamous Cell Carcinoma. *PLoS One* 11:e0145807 . doi: 10.1371/journal.pone.0145807

Yu L, Huang J, Zhai D, Liu L, Guo K, Long X, Xiong J, Zhang Z, Wang Y, Zhao Y, Wu P, Wang D, Lin Z, Wu J, Xiong N, Wang T (2014) MAPT rs242562 and GSK3B rs334558 are associated with Parkinson's Disease in central China. *BMC Neuroscience* 15:54 . doi: 10.1186/1471-2202-15-54

Yu L, Liu H (2004) Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research* 5:1205–1224

Zanello A, Curtis L, Badan Bâ M, Merlo MCG (2009) Working memory impairments in first-episode psychosis and chronic schizophrenia. *Psychiatry Research* 165:10–18 . doi: 10.1016/j.psychres.2007.10.006

Zeileis A, Hothorn T (2002) Diagnostic Checking in Regression Relationships. *R News* 2:7–10

Zhang H, Singer B (1999) Recursive Partitioning in the Health Sciences. *Springer New York,* New York, NY

Zhang N, Yu J-T, Yang Y, Yang J, Zhang W, Tan L (2011) Association analysis of GSK3B and MAPT polymorphisms with Alzheimer's disease in Han Chinese. *Brain Research* 1391:147–153 . doi: 10.1016/j.brainres.2011.03.052

Zhao Y, Chen F, Zhai R, Lin X, Wang Z, Su L, Christiani DC (2012) Correction for population stratification in random forest analysis. *International Journal of Epidemiology* 41:1798–1806 . doi: 10.1093/ije/dys183

Zimmerman M, Ellison W, Young D, Chelminski I, Dalrymple K (2015) How many different ways do patients meet the diagnostic criteria for major depressive disorder? *Comprehensive Psychiatry* 56:29–34 . doi: 10.1016/j.comppsych.2014.09.007

(2016) The MQ manifesto for young people's mental health

(1946) WHO | World Health Organization. In: WHO

Appendix A

Tables and Figures chapter 2

WAC	r=0.80		r=0.40		R=0.10	
N	BIAS	COV%	BIAS	COV%	BIAS	COV%
5	-0.0007	94.4	-0.0005	97.4	-0.0011	93.8
20	0.00005	95.6	0.00039	93.4	-0.0005	94.6
40	0.00107	93.2	0.0011	93.6	0.00094	9.42

Table A.1. Bias and coverage of V_2 under the null hypothesis.

Uncorrelated	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
SAC single	0.00019	-0.0001	0.00012	0.00028	-0.0002	-0.0004	0.00024	0.000246	-0.0001
WAC single	-0.00001	-0.00004	-0.00007	-0.00001	-0.00004	-0.00007	-0.00001	-0.00004	-0.0001
SAC interaction	0.00005	0.00006	-0.0001	0.000003	0.00006	-0.0005	-0.00003	-0.00007	-0.0001
WAC interaction	-0.0002	0.00015	0.00031	-0.00011	-0.00001	-0.0004	0.00025	-0.000008	-0.0002

Table A.2. Medians of the observed correlation between the correlated predictors.

Correlated	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
SAC single	0.795	0.800	0.799	0.385	0.397	0.400	0.081	0.097	0.099
WAC single	0.795	0.799	0.800	0.387	0.398	0.399	0.081	0.097	0.099
SAC interaction	0.794	0.799	0.800	0.384	0.398	0.399	0.079	0.098	0.099
WAC interaction	0.795	0.799	0.800	0.385	0.399	0.399	0.082	0.098	0.099

Table A.3. Medians of the observed correlation between the uncorrelated predictors.

	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
gini	5.00	5.1	5.15	5.06	5.11	5.13	5.08	5.14	5.08
rawpermRF	4.95	5.82	6.29	4.92	5.09	5.24	4.97	5.05	4.99
Breiman	5.05	5.58	5.99	4.99	5.29	5.42	5.07	5.13	5.10
Liaw	5.05	5.58	5.99	4.99	5.29	5.42	5.07	5.13	5.10
rawpermCF	4.89	4.92	4.94	4.87	4.89	4.97	0	0	0
Party	4.96	5.00	5.05	4.85	4.98	4.95	4.88	5.01	4.98
AUC	4.96	5.00	5.05	4.85	4.99	4.96	4.88	5.00	4.98
mindepth 39	4.96	5.00	5.05	4.85	4.99	4.95	4.89	5.00	4.98
mindepth 27	4.98	5.11	5.11	4.93	5.03	5.13	4.97	4.97	4.98

Table A.4. Percentage of importance scores greater than or equal to the cut-off across all 500 null VIMs or minimal depth.

vim median	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
gini	1.25	1.01	1.027	1.44	1.37	1.37	1.5	1.48	1.48
rawpermRF	0.0003	0.0012	0.002	0.00006	0.00036	0.00068	-0.00007	-0.00006	-0.00002
Breiman	1.13	3.57	4.93	0.18	1.12	1.98	-0.24	-0.18	-0.06
Liaw	0.059	0.182	0.254	0.009	0.058	0.103	-0.012	-0.009	-0.003
rawpermCF	-0.00002	0.00000	0.00000	-0.00003	0	0	0	0	0
Party	-0.00003	0.00000	0.00003	-0.00005	-0.00003	-0.00002	-0.00007	-0.00007	-0.00006
AUC	-0.00003	0.00001	0.00003	-0.00006	-0.00003	-0.00002	-0.00007	-0.00007	-0.00007
mindepth 39	6.88	7.3	7.35	6.67	6.77	6.79	6.63	6.67	6.63
mindepth 27	6.93	7.39	7.45	6.68	6.8	6.81	6.615	6.63	6.63

Table A.5. Median of VIM and minimal depth medians for the correlated variables under H0.

vim median	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
gini	1.51	1.63	1.84	1.50	1.54	1.59	1.5	1.5	1.51
rawpermRF	-0.00008	-0.00009	-0.00010	-0.00008	-0.00008	-0.00009	-0.00008	-0.00008	-0.00008
Breiman	-0.25	-0.27	-0.29	-0.25	-0.24	-0.26	-0.26	-0.23	-0.23
Liaw	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
rawpermCF	-0.00008	-0.00009	-0.00011	-0.00008	-0.00008	-0.00008	0.00000	0.00000	0.00000
Party	-0.00007	-0.00008	-0.00009	-0.00007	-0.00007	-0.00008	-0.00007	-0.00007	-0.00007
AUC	-0.00007	-0.00008	-0.00009	-0.00007	-0.00007	-0.00008	-0.00007	-0.00007	-0.00007
mindepth 39	6.6	6.49	6.32	6.61	6.59	6.54	6.62	6.61	6.61
mindepth 27	6.6	6.48	6.27	6.6	6.58	6.52	6.61	6.61	6.6

Table A.6. Median of VIM medians and minimal depth medians for the uncorrelated variables under H0.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	349.05	326.55	320.78	471.57	451.05	445.34	530.37	519.48	516.70
rawpermRF	0.450	0.412	0.405	0.604	0.577	0.567	0.656	0.646	0.641
BREIMAN	91.26	81.73	79.65	185.60	167.80	160.93	237.69	226.14	221.59
Liaw	1.56	1.54	1.53	1.63	1.62	1.62	1.64	1.63	1.63
rawpermCF	0.0807	0.0010	0.0001	0.22	0.0003	0	0	0	0
Party	0.961	0.870	0.853	1.439	1.366	1.343	1.649	1.626	1.619
AUC	0.961	0.871	0.854	1.439	1.366	1.342	1.648	1.627	1.619
mindepth39	0.87	0.87	0.87	0.87	0.87	0.87	0.88	0.88	0.87
mindepth27	1.25	1.24	1.24	1.25	1.24	1.24	1.27	1.26	1.25

Table A.7. Median of VIM and minimal for V_2 (associated variable) under H_A. Strong single study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	58.74	11.16	4.61	20.44	5.06	2.56	2.42	1.59	1.59
rawpermRF	0.0160	0.0024	0.0010	0.0039	0.0008	0.0004	0.00013	0.00007	0.00007
BREIMAN	17.82	9.42	6.86	9.22	4.16	3.06	0.83	0.58	0.58
Liaw	0.81	0.47	0.35	0.46	0.21	0.16	0.04	0.03	0.03
rawpermCF	0.00010	0.000002	0.00000004	0.00013	0	0	0	0	0
Party	0.0507	0.0067	0.0024	0.0142	0.0025	0.0011	0.0007	0.0002	0.0002
AUC	0.0507	0.0067	0.0024	0.0142	0.0025	0.0011	0.0007	0.0002	0.0002
mindepth39	2.39	4.89	5.79	3.52	5.51	6.08	6.10	6.36	6.36
mindepth27	2.28	4.49	5.50	3.23	5.16	5.85	5.94	6.23	6.23

Table A.8. Median of VIM medians and minimal depth medians for the correlated variables under H_A. Strong single study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	0.35	0.26	0.24	0.74	0.63	0.60	0.60	0.84	0.84
rawpermRF	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.00000	-0.00001	-0.00001
BREIMAN	-0.05	-0.05	-0.05	-0.04	-0.05	-0.03	-0.03	-0.07	-0.07
Liaw	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
rawpermCF	-0.000001	-0.000001	-0.0000007	-0.000003	-0.000003	-0.000001	-0.000001	0	0
Party	-0.000009	-0.000004	-0.000003	-0.00002	-0.00001	-0.00001	-0.00001	-0.00003	-0.00003
AUC	-0.000009	-0.000004	-0.000003	-0.00003	-0.00001	-0.00001	-0.00001	-0.00003	-0.00003
mindepth39	6.70	6.74	6.73	6.66	6.68	6.68	6.68	6.70	6.70
mindepth27	6.68	6.72	6.73	6.64	6.65	6.65	6.65	6.67	6.67

Table A.9. Median of VIM medians and minimal depth medians for the uncorrelated variables under H_A. Strong single study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	1.92	1.41	1.25	2.78	2.24	2.12	2.81	2.91	2.89
rawpermRF	0.0022	0.0034	0.0037	0.0025	0.0022	0.0026	0.0018	0.0023	0.0025
BREIMAN	3.75	5.63	6.06	3.55	3.70	4.42	2.79	3.29	3.45
Liaw	0.19	0.29	0.31	0.18	0.19	0.23	0.14	0.17	0.18
rawpermCF	0.0001	-0.000006	-0.0000002	0.0008	0	0	0	0	0
Party	0.0009	0.0007	0.0004	0.0018	0.0012	0.0011	0.0017	0.0019	0.0021
AUC	0.0010	0.0007	0.0004	0.0019	0.0012	0.0011	0.0017	0.0020	0.0020
mindepth39	5.92	6.69	6.97	5.05	5.61	5.83	5.08	4.94	4.97
mindepth27	5.92	6.62	6.90	5.17	5.72	5.82	5.19	5.06	5.14

Table A.10. Median of VIM and minimal depth for V_2 (associated variable) under H_A. Weak single study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	1.53	1.15	1.12	1.52	1.40	1.39	1.50	1.49	1.49
rawpermRF	0.0014	0.0021	0.0027	0.0002	0.0005	0.0007	-0.00004	-0.00004	-0.00004
BREIMAN	3.05	4.55	5.41	0.65	1.35	2.05	-0.13	-0.13	-0.13
Liaw	0.16	0.23	0.28	0.03	0.07	0.11	-0.01	-0.01	-0.01
rawpermCF	-0.00003	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Party	0.0002	0.0001	0.00004	-0.00003	-0.00002	-0.00002	-0.00007	-0.00006	-0.00006
AUC	0.0002	0.0001	0.00005	-0.00003	-0.00003	-0.00002	-0.0001	-0.0001	-0.0001
mindepth39	6.47	7.08	7.20	6.56	6.72	6.75	6.60	6.60	6.60
mindepth27	6.42	7.00	7.11	6.56	6.70	6.73	6.61	6.61	6.61

Table A.11. Median of VIM medians and minimal depth medians for the correlated non-associated variables under H_A. Weak single study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	1.51	1.61	1.78	1.50	1.53	1.59	1.50	1.50	1.51
rawpermRF	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.00007	-0.00007	-0.00008
BREIMAN	-0.24	-0.26	-0.27	-0.22	-0.22	-0.26	-0.22	-0.22	-0.24
Liaw	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
rawpermCF	-0.00008	-0.000088	-0.0000919	-0.00007	0	0	0	0	0
Party	-0.0001	-0.0001	-0.00009	-0.00007	-0.00007	-0.00007	-0.00007	-0.00007	-0.00007
AUC	-0.0001	-0.0001	-0.00009	-0.00007	-0.00007	-0.00008	-0.00007	-0.00007	-0.00007
mindepth39	6.58	6.47	6.29	6.59	6.56	6.51	6.60	6.59	6.59
mindepth27	6.59	6.48	6.33	6.60	6.57	6.52	6.61	6.60	6.59

Table A.12. Median of VIM medians and minimal depth medians for the uncorrelated variables under H_A. Weak single study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	48393.30	43774.15	42254.90	65046.25	61749.85	60544.60	73138.40	71885.10	71163.25
rawpermRF	24.41	21.46	20.97	33.02	31.23	30.75	36.59	36.16	35.84
BREIMAN	7265.62	6453.20	6170.27	12448.70	11371.75	11066.90	15459.00	14945.30	14682.70
Liaw	151.26	148.13	146.80	159.90	158.93	158.62	161.62	161.39	161.27
rawpermCF	4.32	0.03	0.002	12.67	0.01	0	0	0	0
Party	1.65	1.49	1.45	2.33	2.21	2.17	2.51	2.55	2.54
AUC	1.65	1.50	1.45	2.33	2.22	2.18	2.50	2.55	2.54
mindepth39	1.15	1.18	1.20	1.13	1.13	1.13	1.15	1.15	1.15
mindepth27	1.52	1.55	1.57	1.49	1.47	1.48	1.52	1.51	1.51

Table A.13. Median of VIM and minimal depth for V_2 (associated correlated variable) under H_A. Strong interaction study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	75889.75	75661.45	74736.55	73061.45	72518.90	71915.90	72545.80	72610.55	72611.55
rawpermRF	36.75	36.53	36.49	35.69	35.40	35.23	36.42	36.31	36.26
BREIMAN	15923.70	15749.35	15828.55	15224.95	15058.40	14957.20	15508.60	15401.20	15301.90
Liaw	161.82	161.75	161.77	161.53	161.45	161.41	161.63	161.61	161.56
rawpermCF	36.67	36.53	36.42	35.68	35.41	35.19	0.00	0.00	0.00
Party	2.74	2.75	2.76	2.62	2.63	2.62	2.50	2.58	2.60
AUC	2.74	2.74	2.76	2.62	2.64	2.62	2.50	2.59	2.59
mindepth39	1.18	1.20	1.22	1.13	1.12	1.13	1.17	1.16	1.14
mindepth27	1.53	1.57	1.59	1.48	1.46	1.47	1.52	1.51	1.50

Table A.14. Median of VIM and minimal for V_{90} (associated uncorrelated variable) under H_A. Strong interaction study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	8003.21	2016.86	1037.13	2835.74	1045.01	680.75	493.28	459.72	432.86
rawpermRF	1.33	0.28	0.16	0.27	0.07	0.04	0.002	0.002	0.002
BREIMAN	1300.48	608.60	466.39	536.04	251.29	201.58	16.89	14.89	11.63
Liaw	62.63	31.17	24.11	27.51	13.02	10.48	0.88	0.77	0.61
rawpermCF	0.0113	-0.00002	0	0.00955	0	0	0	0	0
Party	0.121	0.019	0.0074	0.0289	0.0053	0.0021	0.0003	0.0002	0.0001
AUC	0.120	0.019	0.0073	0.0291	0.0052	0.0022	0.0004	0.0002	0.0001
mindepth39	3.49	5.62	6.32	4.86	6.16	6.54	6.73	6.79	6.83
mindepth27	3.31	5.42	6.19	4.55	5.95	6.40	6.67	6.73	6.78

Table A.15. Median of VIM medians and minimal depth medians for the correlated variables under H_A. Strong interaction study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	305.09	287.81	295.00	373.36	358.86	354.36	378.92	376.95	375.40
rawpermRF	-0.0009	-0.0011	-0.0009	-0.0010	-0.0008	-0.0011	-0.0015	-0.0015	-0.0012
BREIMAN	-9.33	-11.37	-9.63	-8.43	-7.03	-9.97	-12.25	-12.65	-9.66
Liaw	-0.49	-0.59	-0.50	-0.44	-0.37	-0.52	-0.64	-0.66	-0.50
rawpermCF	-0.00091	-0.00091	-0.00086	-0.0011	0	0	0	0	0
Party	-0.00011	-0.00011	-0.00012	-0.00014	-0.00012	-0.00012	-0.0002	-0.0002	-0.0002
AUC	-0.00011	-0.00011	-0.00012	-0.00015	-0.00012	-0.00012	-0.0002	-0.0002	-0.0002
mindepth39	6.90	6.93	6.90	6.90	6.91	6.89	6.93	6.93	6.93
mindepth27	6.87	6.88	6.86	6.87	6.87	6.86	6.91	6.91	6.91

Table A.16. Median of VIM medians and minimal depth medians for the uncorrelated variables under H_A. Strong interaction study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	184.08	136.29	126.10	225.60	208.50	203.49	243.24	244.30	247.51
rawpermRF	0.16	0.22	0.29	0.13	0.16	0.19	0.12	0.14	0.14
BREIMAN	333.49	467.16	565.88	264.57	337.84	401.27	248.12	262.61	274.08
Liaw	17.29	24.09	29.05	13.75	17.53	20.77	12.90	13.65	14.23
rawpermCF	0.01290	0.00008	0.00004	0.0484	0	0	0	0	0
Party	0.00088	0.00054	0.00037	0.00124	0.00110	0.00101	0.00126	0.00133	0.00135
AUC	0.00086	0.00052	0.00037	0.00125	0.00109	0.00106	0.00132	0.00140	0.00135
mindepth39	6.31	6.98	7.18	5.89	6.08	6.18	5.72	5.76	5.77
mindepth27	6.27	6.90	7.07	5.97	6.11	6.17	5.82	5.84	5.84

Table A.17. Median of VIM and minimal for V_2 (associated correlated variable) under H_A. Weak interaction study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	259.06	298.01	344.61	247.21	254.67	270.93	249.28	247.39	259.63
rawpermRF	0.20	0.28	0.39	0.12	0.17	0.19	0.14	0.13	0.16
BREIMAN	353.45	472.18	584.85	248.45	322.53	353.12	270.01	256.79	286.88
Liaw	18.31	24.36	29.97	12.92	16.73	18.27	14.02	13.33	14.89
rawpermCF	0.19	0.28	0.38	0.13	0.17	0.18	0	0	0
Party	0.00193	0.00261	0.00372	0.00145	0.00158	0.00194	0.00139	0.00132	0.00152
AUC	0.00200	0.00262	0.00360	0.00138	0.00164	0.00189	0.00147	0.00133	0.00157
mindepth39	5.58	5.17	4.73	5.69	5.66	5.43	5.70	5.77	5.60
mindepth27	5.65	5.28	4.91	5.81	5.76	5.54	5.79	5.85	5.73

Table A.18. Median of VIM and minimal for V_{90} (associated uncorrelated variable) under H_A. Weak interaction study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	148.69	115.08	112.63	154.72	144.28	142.72	155.45	154.73	153.22
rawpermRF	0.0918	0.1524	0.2246	0.0123	0.0391	0.0649	-0.0068	-0.0039	-0.0014
BREIMAN	227.55	388.21	510.53	37.17	118.27	190.91	-21.42	-11.17	-3.92
Liaw	11.85	20.10	26.23	1.94	6.14	9.93	-1.12	-0.58	-0.20
rawpermCF	-0.00042	-0.00030	-0.00002	-0.0036	0	0	0	0	0
Party	0.00016	0.00007	0.00006	-0.00002	-0.00002	-0.00001	-0.0001	-0.0001	-0.0001
AUC	0.00016	0.00007	0.00006	-0.00002	-0.00002	-0.00002	-0.0001	-0.0001	-0.0001
mindepth39	6.65	7.20	7.34	6.60	6.75	6.79	6.62	6.62	6.63
mindepth27	6.60	7.12	7.24	6.60	6.74	6.77	6.63	6.62	6.63

Table A.19. Median of VIM medians and minimal depth medians for the correlated variables under H_A. Weak interaction study.

VIM median	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	155.17	165.43	183.54	154.82	157.99	163.25	154.73	154.68	155.45
rawpermRF	-0.0071	-0.0078	-0.0096	-0.0071	-0.0074	-0.0086	-0.0079	-0.0076	-0.0081
BREIMAN	-22.19	-23.69	-27.46	-22.07	-21.45	-26.07	-24.09	-24.46	-24.12
Liaw	-1.16	-1.24	-1.43	-1.15	-1.12	-1.36	-1.26	-1.27	-1.26
rawpermCF	-0.0069	-0.0077	-0.0089	-0.0075	-0.0074	-0.0085	0	0	0
Party	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
AUC	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
mindepth39	6.60	6.50	6.33	6.61	6.59	6.53	6.61	6.61	6.60
mindepth27	6.61	6.52	6.38	6.62	6.60	6.55	6.61	6.62	6.61

Table A.20. Median of VIM medians and minimal depth medians for the uncorrelated variables under H_A. Weak interaction study.

WAC	r = 0.80			r = 0.40			r = 0.10		
N	**5**	**20**	**40**	**5**	**20**	**40**	**5**	**20**	**40**
GINI	34.29	9.97	4.73	67.06	48.77	42.81	69.84	73.06	72.36
rawpermRF	76.06	75.12	52.07	74.17	66.86	61.85	65.46	71.29	72.49
BREIMAN	72.61	68.08	40.48	66.04	61.29	52.69	56.53	63.20	65.44
Liaw	72.60	68.06	40.48	66.03	61.30	52.71	56.54	63.20	65.43
rawpermCF	13.34	0.00	0	45.01	0.0028	0	0	0	0
Party	65.45	59.80	45.58	80.65	68.57	66.31	79.59	81.79	81.74
AUC	65.58	60.90	45.55	80.92	70.13	67.63	79.81	82.41	81.90
mindepth39	49.93	13.15	4.44	79.16	60.22	53.62	81.02	82.34	81.85
mindepth27	51.78	14.26	5.08	79.26	61.22	56.07	80.36	82.20	81.25

Table A.21. Power of VIMs and minimal depth detecting V_2, permuting the outcome. Weak single study (WAC).

SAC	r = 0.80			r = 0.40			r = 0.10		
N	**5**	**20**	**40**	**5**	**20**	**40**	**5**	**20**	**40**
GINI	100	100	100	100	100	100	100	100	100
rawpermRF	100	100	100	100	100	100	100	100	100
BREIMAN	100	100	100	100	100	100	100	100	100
Liaw	100	100	100	100	100	100	100	100	100
rawpermCF	100	55.3404	0.0172	100	32.7868	5.0472	0	0	0
Party	100	100	100	100	100	100	100	100	100
AUC	100	100	100	100	100	100	100	100	100
mindepth39	100	100	100	100	100	100	100	100	100
mindepth27	100	100	100	100	100	100	100	100	100

Table A.22. Power of VIMs and minimal depth detecting V_2, permuting the outcome. Strong single study (SAC).

V2 WAC	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	23.81	3.83	0.95	45.80	35.75	32.96	53.32	54.74	55.42
rawpermRF	64.65	56.97	36.38	58.34	58.92	53.37	56.58	59.46	60.03
BREIMAN	66.27	53.50	30.43	55.59	56.63	48.15	53.74	56.25	56.74
Liaw	66.26	53.51	30.41	55.59	56.62	48.15	53.73	56.25	56.74
rawpermCF	8.65	0.02	0.00	31.18	0.18	0	0	0	0
Party	66.13	52.23	41.78	76.10	73.56	69.79	75.65	76.88	76.66
AUC	64.63	51.53	42.08	74.75	72.14	70.11	75.59	77.89	76.45
mindepth39	27.94	4.91	0.46	51.25	42.42	36.37	59.15	58.40	57.34
mindepth27	29.95	4.63	0.51	49.96	42.08	36.80	56.83	58.27	56.89

Table A.23. Power of VIMs and minimal depth detecting V_2, permuting the outcome. Weak interaction study (WAC).

V90 WAC	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	61.35	72.15	83.05	55.30	59.98	67.32	55.98	55.01	61.40
rawpermRF	65.76	62.82	51.08	57.66	59.94	52.41	61.67	57.03	63.14
BREIMAN	63.77	52.55	41.16	53.95	55.05	41.57	57.87	54.58	59.03
Liaw	63.78	52.58	41.20	53.96	55.05	41.59	57.87	54.58	59.04
rawpermCF	66.97	81.48	89.49	59.67	68.21	77.74	0	0	0
Party	82.01	91.49	94.18	77.79	84.97	84.67	77.39	76.04	78.37
AUC	82.10	91.47	94.40	78.61	84.86	86.61	76.97	75.39	79.18
mindepth39	67.02	79.84	88.25	60.42	66.64	74.98	61.28	58.59	65.61
mindepth27	65.83	79.99	89.23	58.60	65.83	74.26	60.04	57.69	63.56

Table A.24. Power of VIMs and minimal depth detecting V_{90}, permuting the outcome. Weak interaction study (WAC).

V2 SAC	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	100	100	100	100	100	100	100	100	100
rawpermRF	100	100	100	100	100	100	100	100	100
BREIMAN	100	100	100	100	100	100	100	100	100
Liaw	100	100	100	100	100	100	100	100	100
rawpermCF	100	2.81	0.44	100	21.08	2.13	0	0	0
Party	100	100	100	100	100	100	100	100	100
AUC	100	100	100	100	100	100	100	100	100
mindepth39	100	100	100	100	100	100	100	100	100
mindepth27	100	100	100	100	100	100	100	100	100

Table A.25. Power of VIMs and minimal depth detecting V_2, permuting the outcome. Strong interaction study (SAC).

V90 SAC	r = 0.80			r = 0.40			r = 0.10		
N	5	20	40	5	20	40	5	20	40
GINI	100	100	100	100	100	100	100	100	100
rawpermRF	100	100	100	100	100	100	100	100	100
BREIMAN	100	100	100	100	100	100	100	100	100
Liaw	100	100	100	100	100	100	100	100	100
rawpermCF	100	100	100	100	100	100	0	0	0
Party	100	100	100	100	100	100	100	100	100
AUC	100	100	100	100	100	100	100	100	100
mindepth39	100	100	100	100	100	100	100	100	100
mindepth27	100	100	100	100	100	100	100	100	100

Table A.26. Power of VIMs and minimal depth detecting V_{90}, permuting the outcome. Strong interaction study (SAC).

Median Depth threshold	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
Null hypothesis	9.920	9.975	10.078	9.910	9.939	9.970	9.913	9.914	9.915

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.

Median Depth threshold	r=0.80			r=0.40			r=0.10		
N	5	20	40	5	20	40	5	20	40
SAC single	8.984	8.790	8.885	9.079	8.945	8.968	9.287	9.215	9.142
WAC single	8.971	8.679	8.701	9.013	8.753	8.889	9.023	9.133	9.133
SAC interaction	9.938	10.193	10.417	10.131	10.085	10.037	10.259	10.223	10.223
WAC interaction	9.953	10.037	10.187	9.931	9.963	10.003	9.915	9.929	9.929

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-association studies. For both mtry values the median was the same.

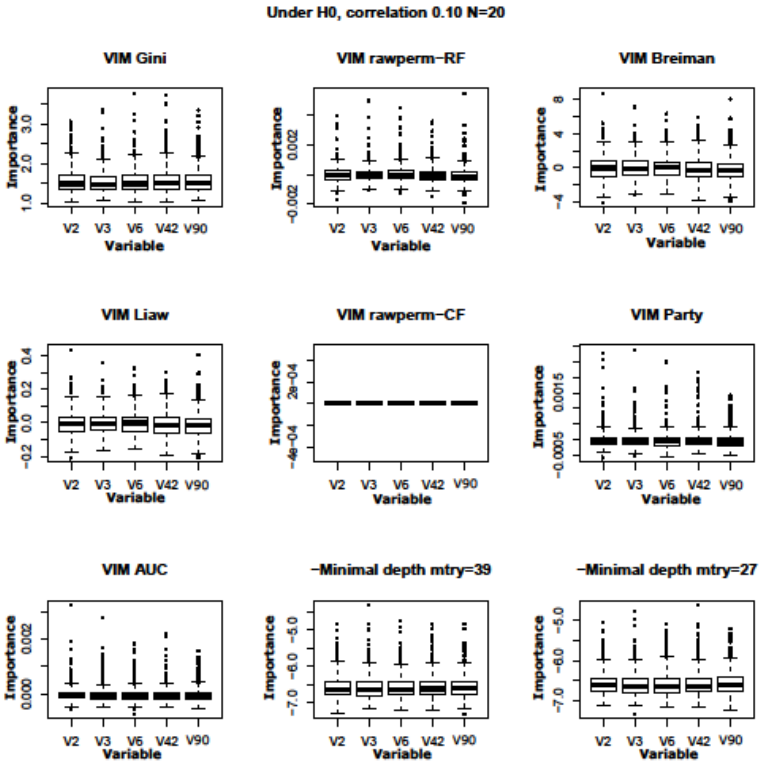

Figure A.1. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_0 for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.10 and N = 20.

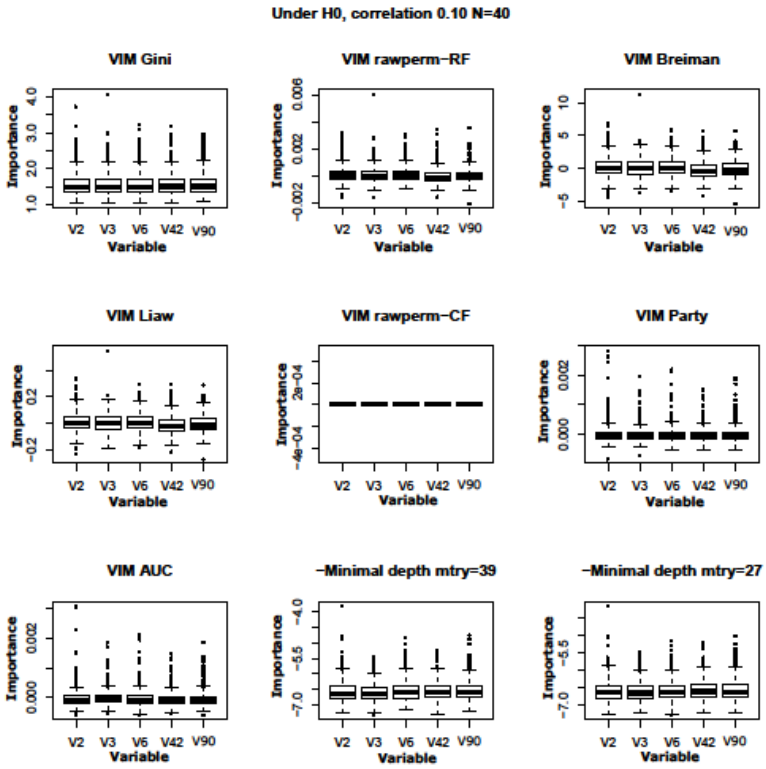

Figure A.2. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_0 for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.10 and N = 40.

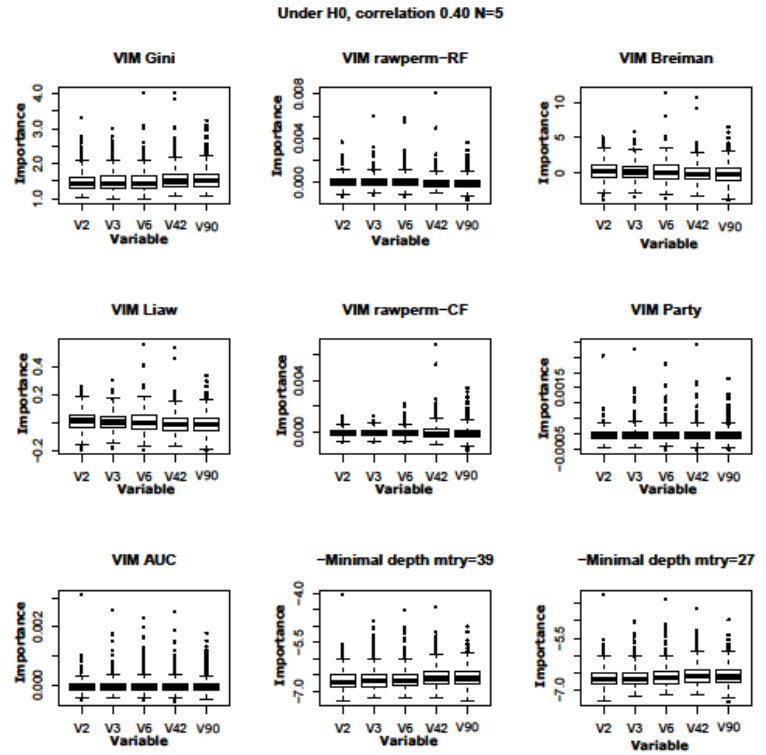

Figure A.3. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_0 for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.40 and N = 5.

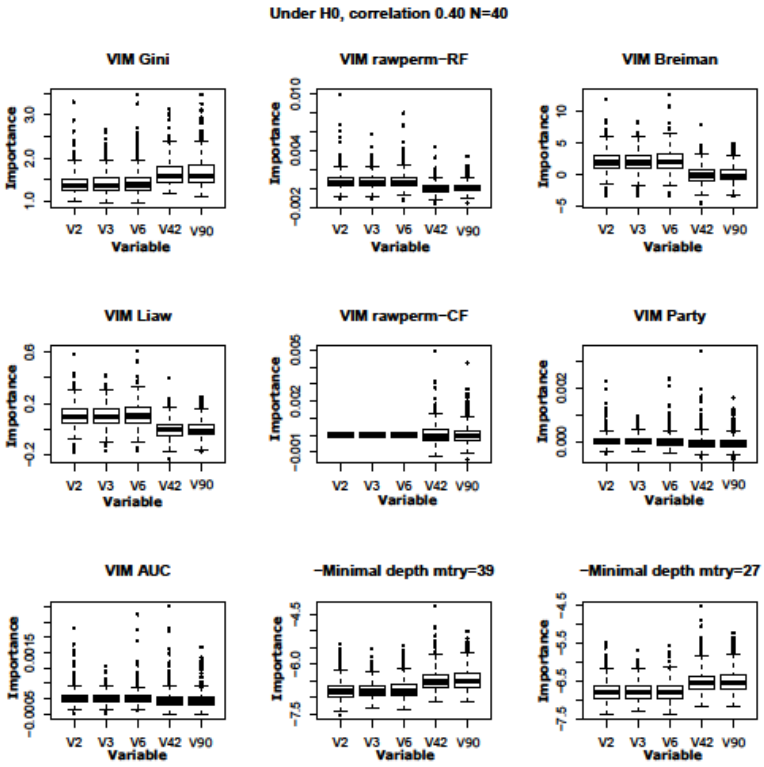

Figure A.4. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_0 for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.40 and N = 40.

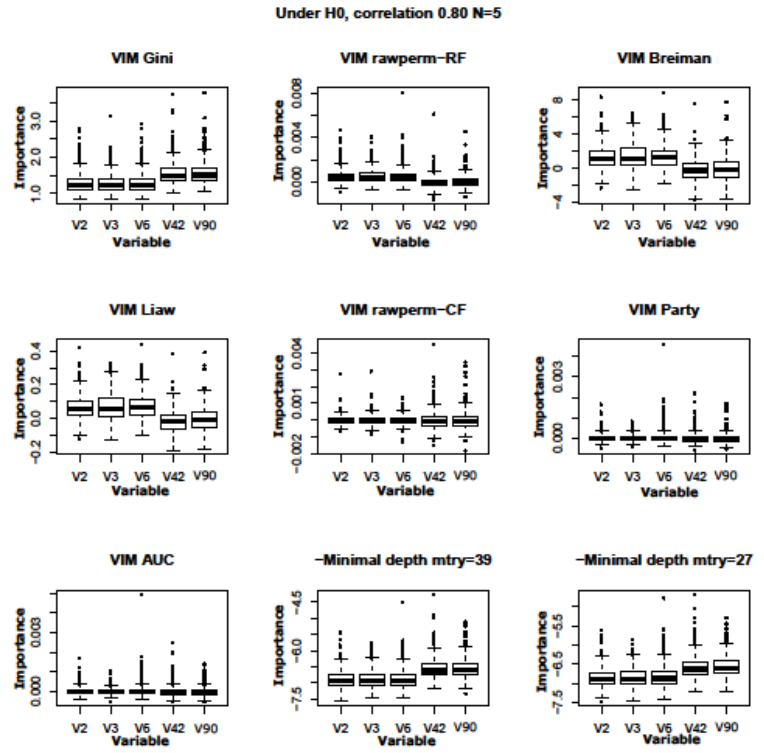

Figure A.5. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_0 for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.80 and N = 5.

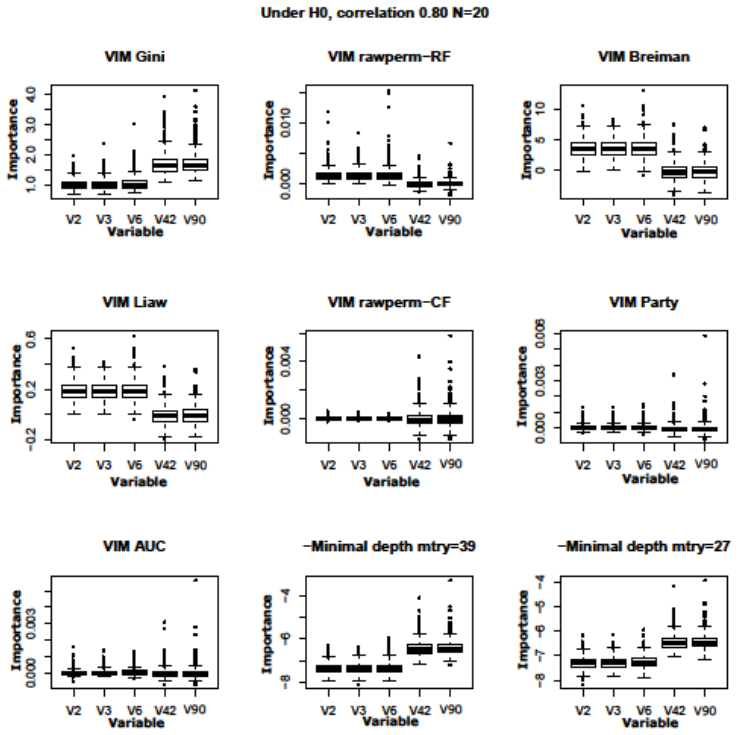

Figure A.6. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_0 for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.80 and N = 20.

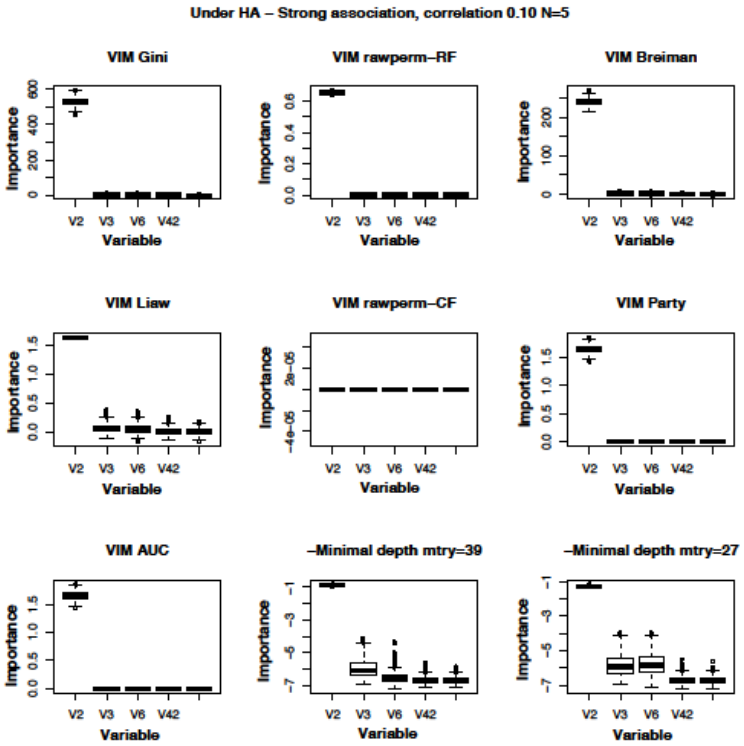

Figure A.7. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.10 and N = 5. Strong single study.

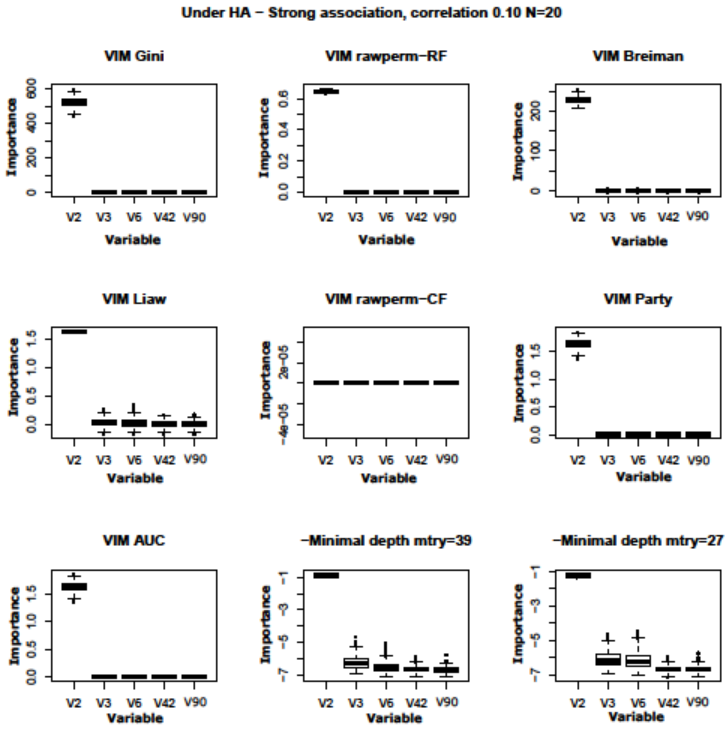

Figure A.8. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.10 and N = 20. Strong single study.

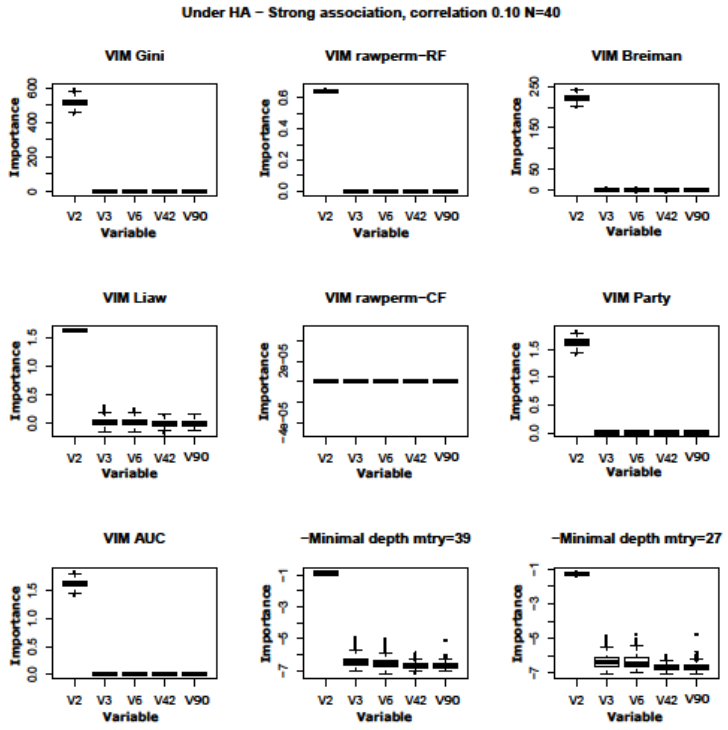

Figure A.9. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.10 and N = 40. Strong single study.

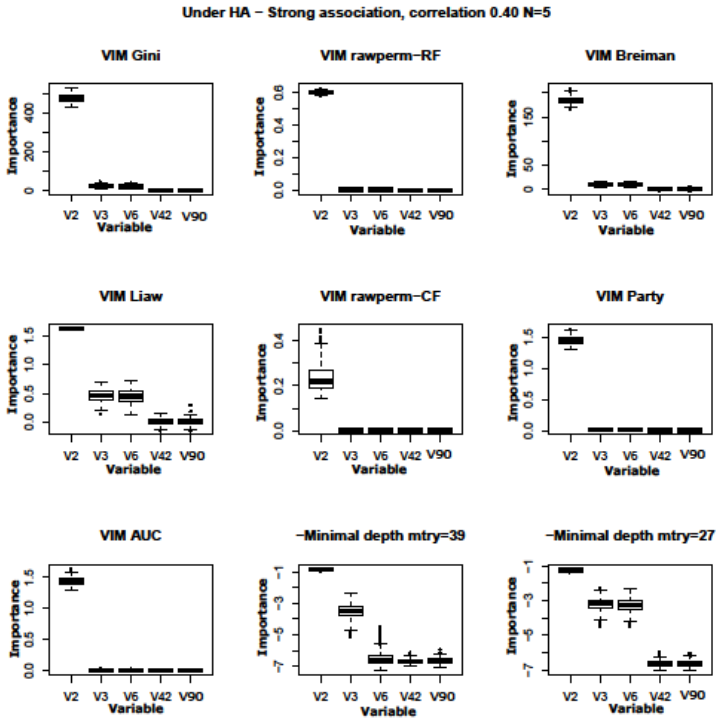

Figure A.10. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.40 and N = 5. Strong single study.

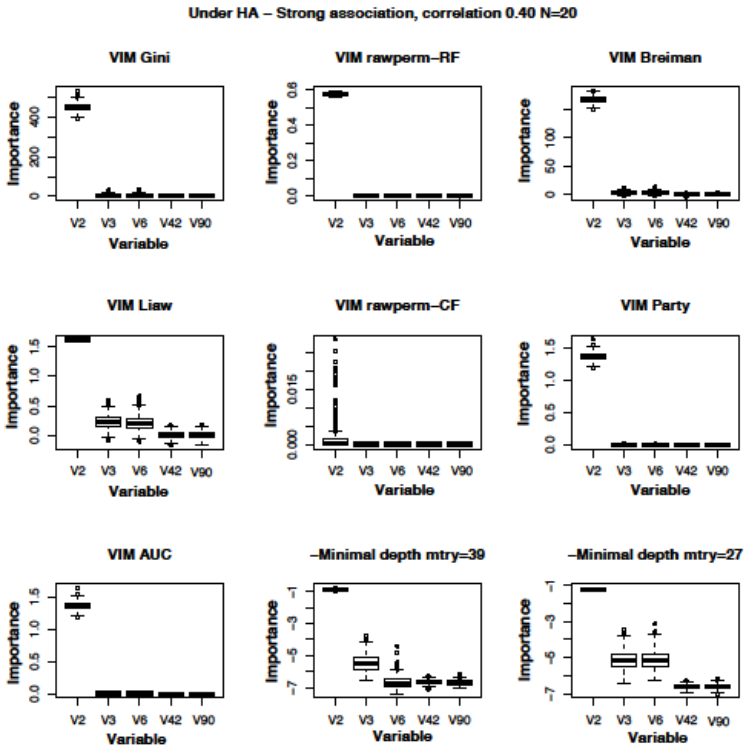

Figure A.11. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.40 and N = 20. Strong single study.

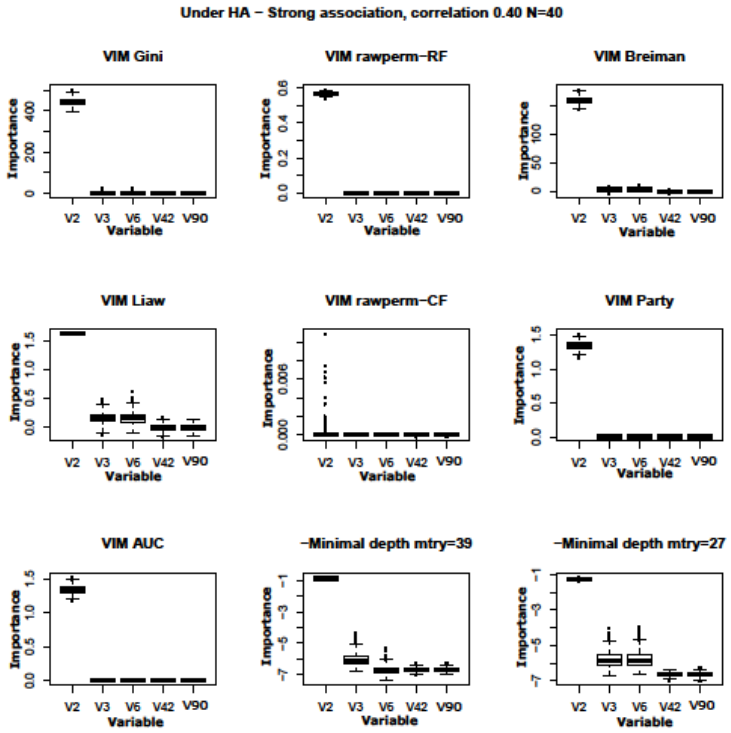

Figure A.12. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.40 and N = 40. Strong single study.

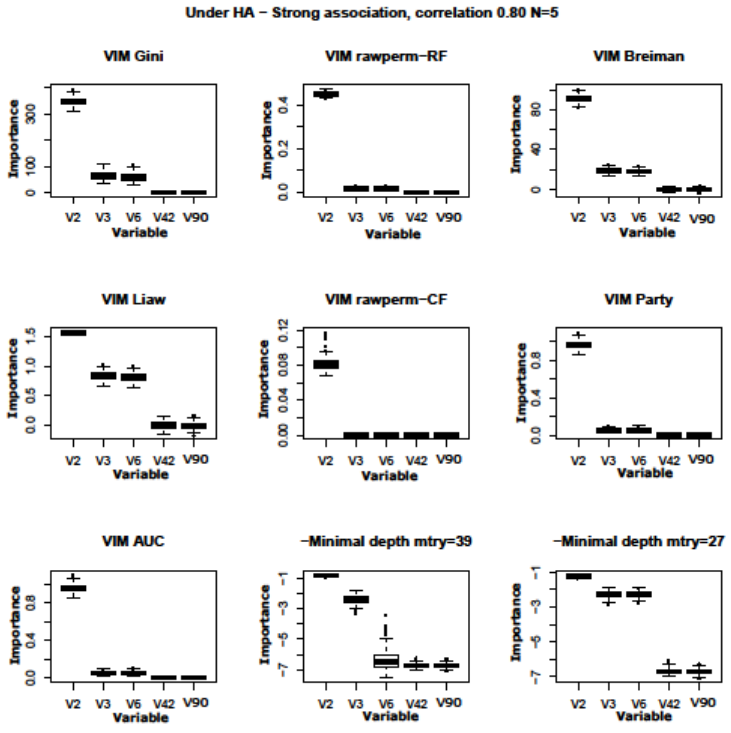

Figure A.13. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.80 and N = 5. Strong single study.

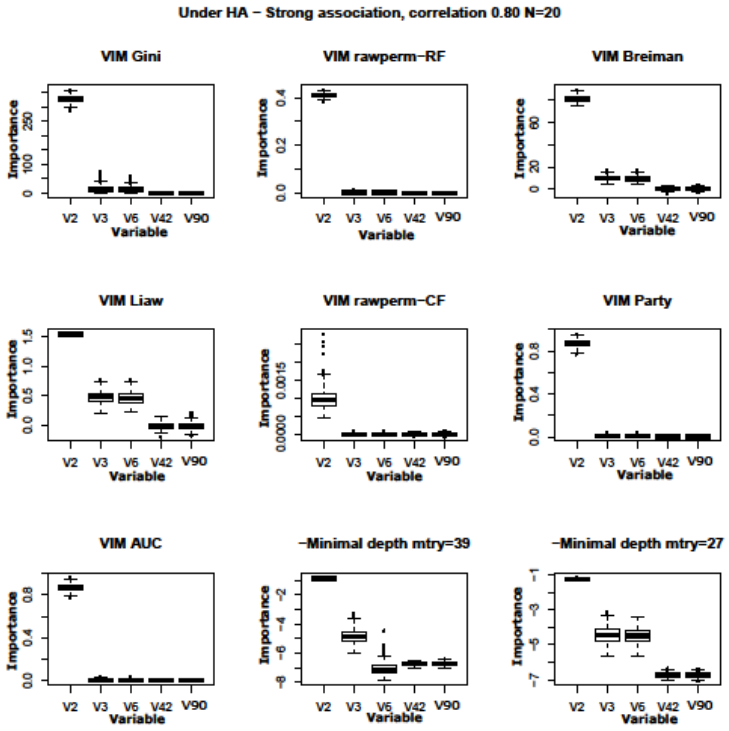

Figure A.14. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.80 and N = 20. Strong single study.

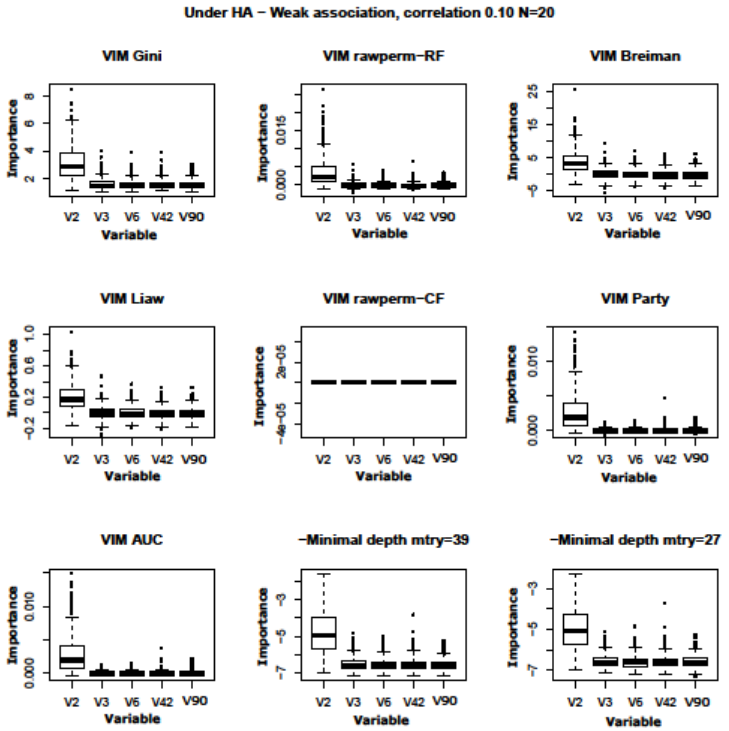

Figure A.15. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.10 and N = 20. Weak single study.

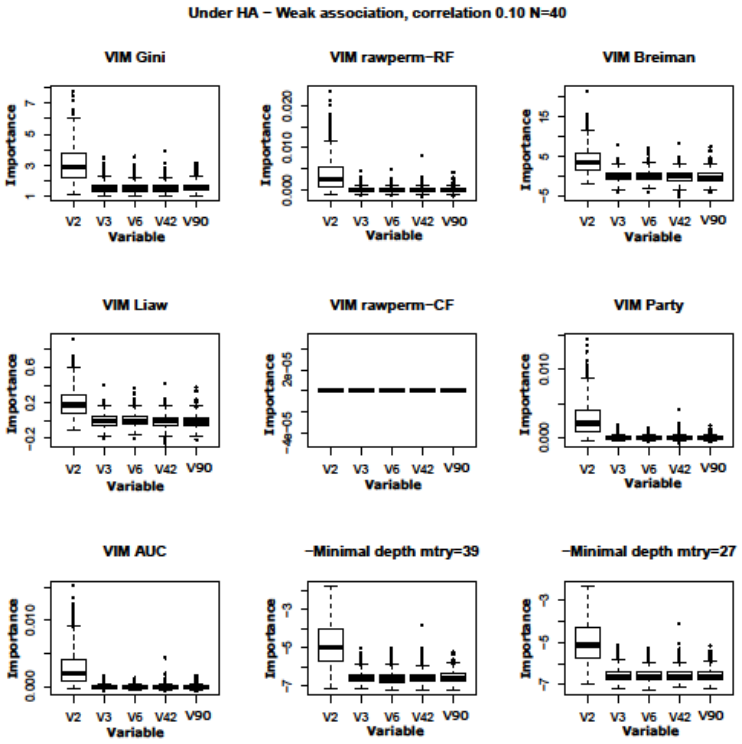

Figure A.16. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.10 and N = 40. Weak single study.

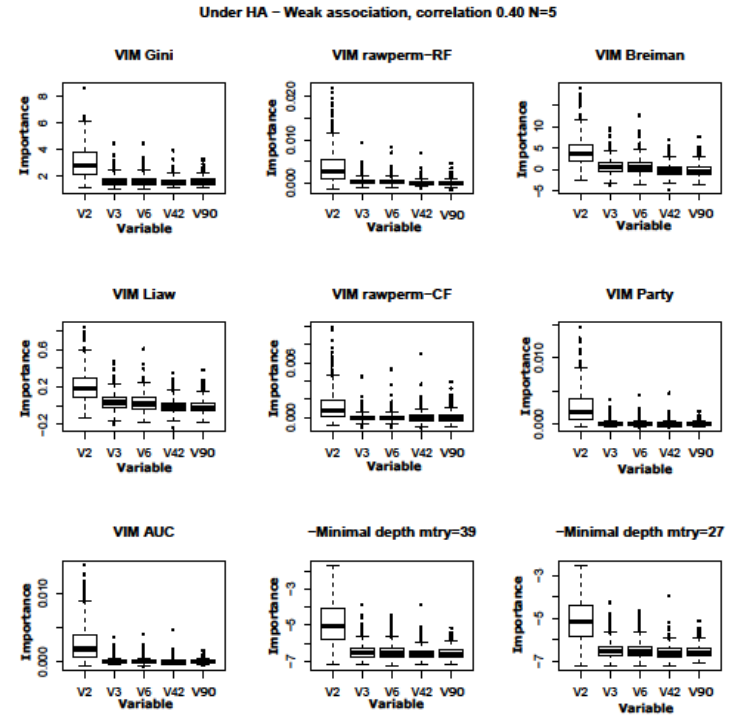

Figure A.17. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.40 and N = 5. Weak single study.

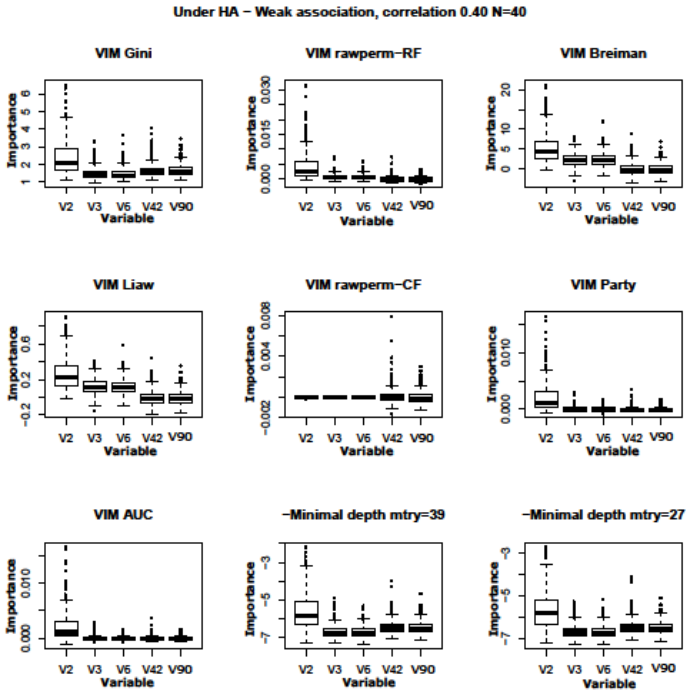

Figure A.18. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.40 and N = 40. Weak single study.

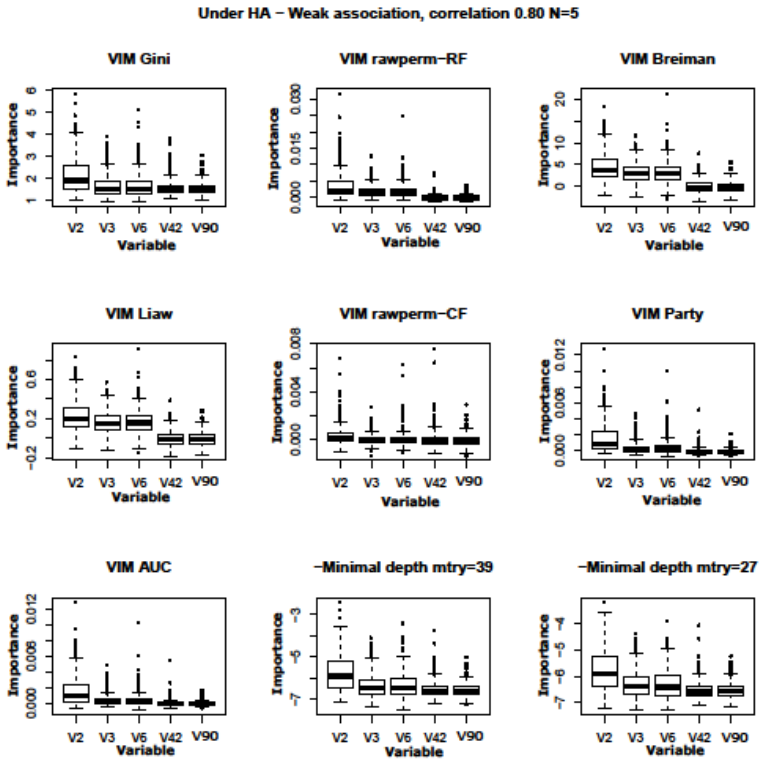

Figure A.19. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.80 and N = 5. Weak single study.

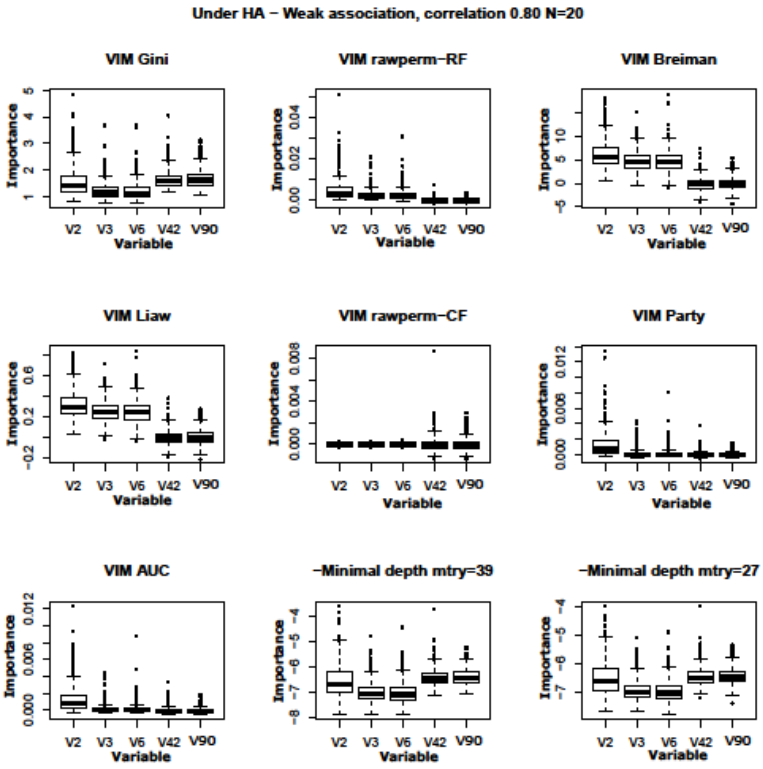

Figure A.20. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2, for two variable correlated V_3 and V_6, and for two independent variables V_{42} and V_{90} when r = 0.80 and N = 20. Weak single study.

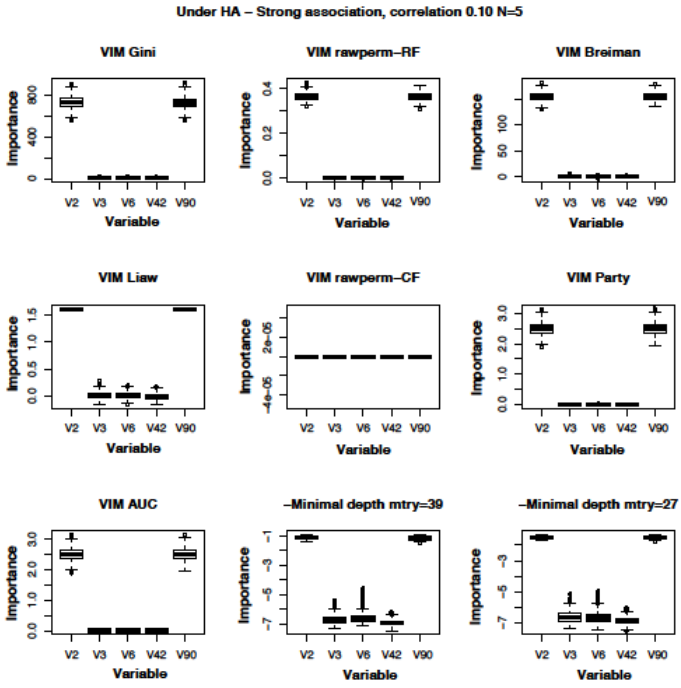

Figure A.21. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.10 and N = 5. Strong interaction study.

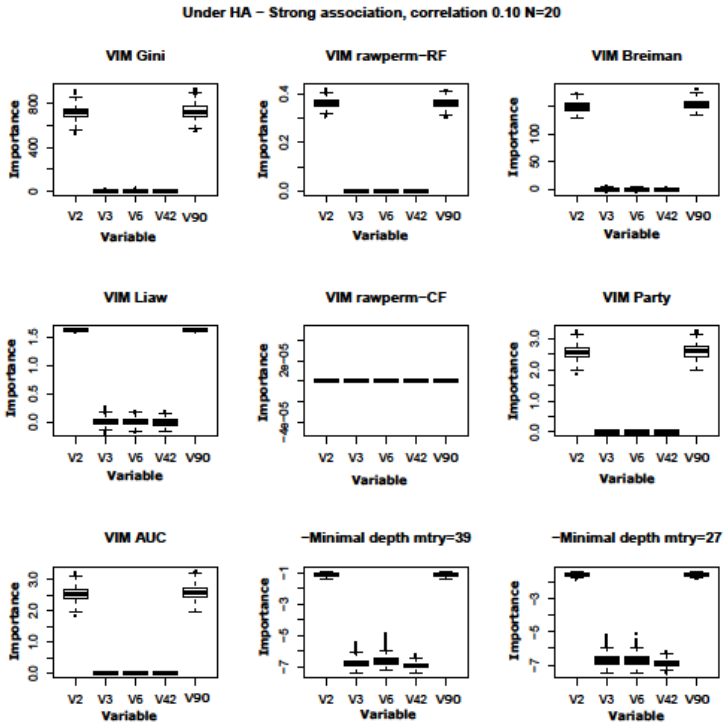

Figure A.22. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.10 and N = 20. Strong interaction study.

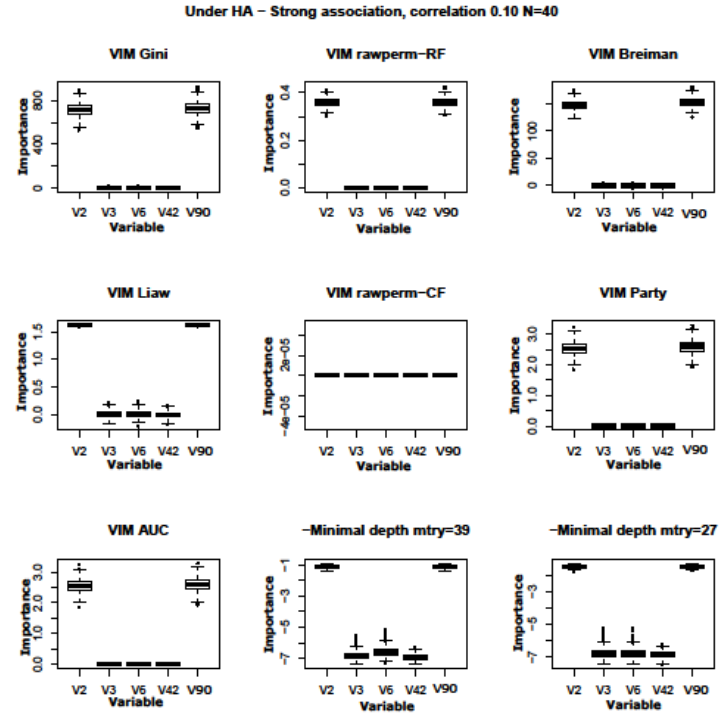

Figure A.23. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.10 and N = 40. Strong interaction study.

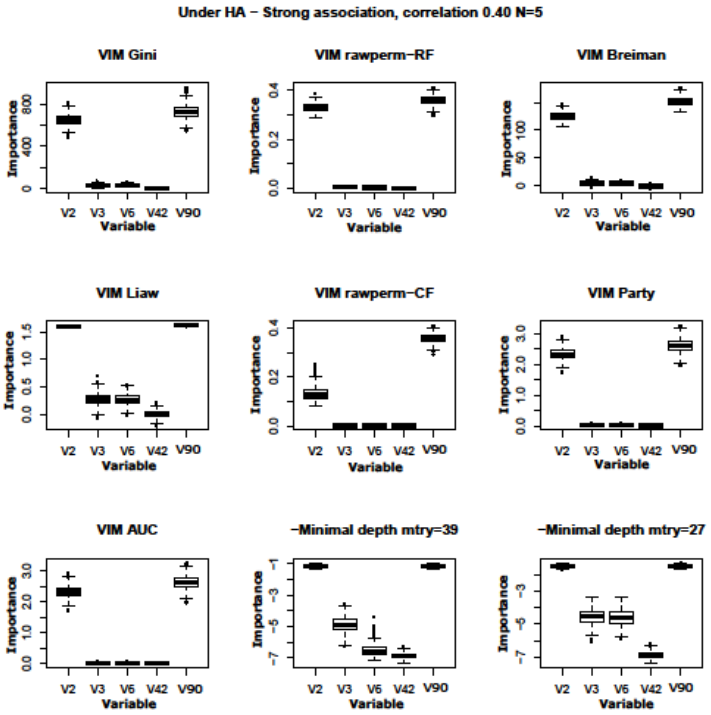

Figure A.24. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.40 and N = 5. Strong interaction study.

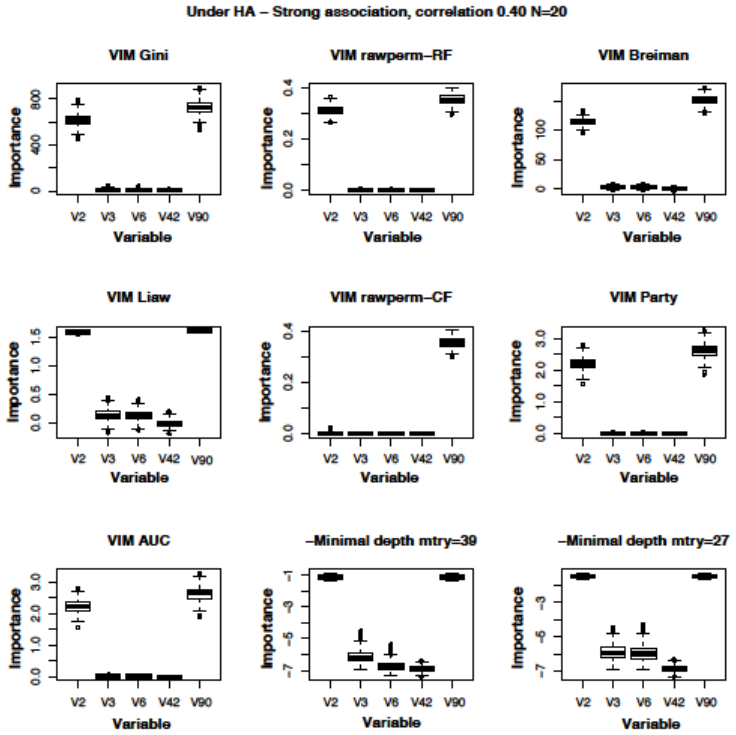

Figure A.25. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.40 and N = 20. Strong interaction study.

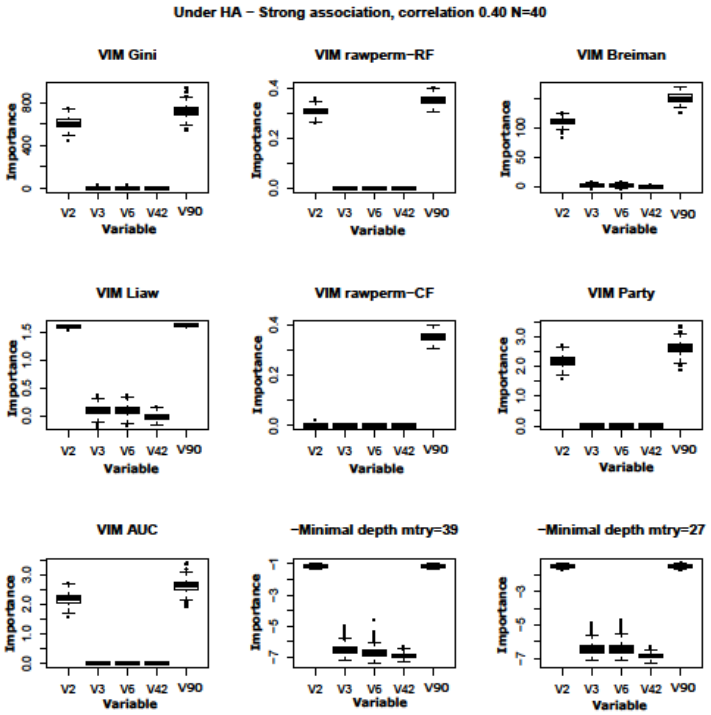

Figure A.26. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.40 and N = 5. Strong interaction study.

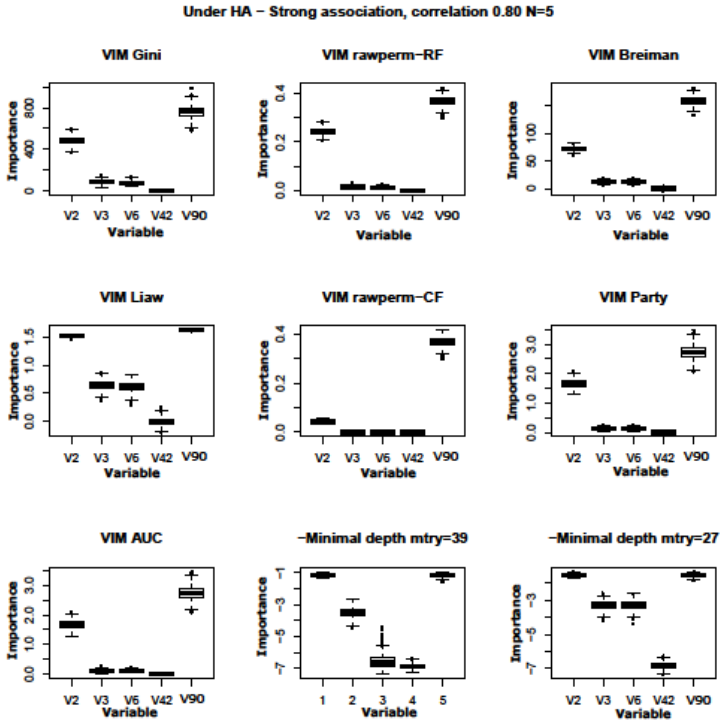

Figure A.27. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.80 and N = 5. Strong interaction study.

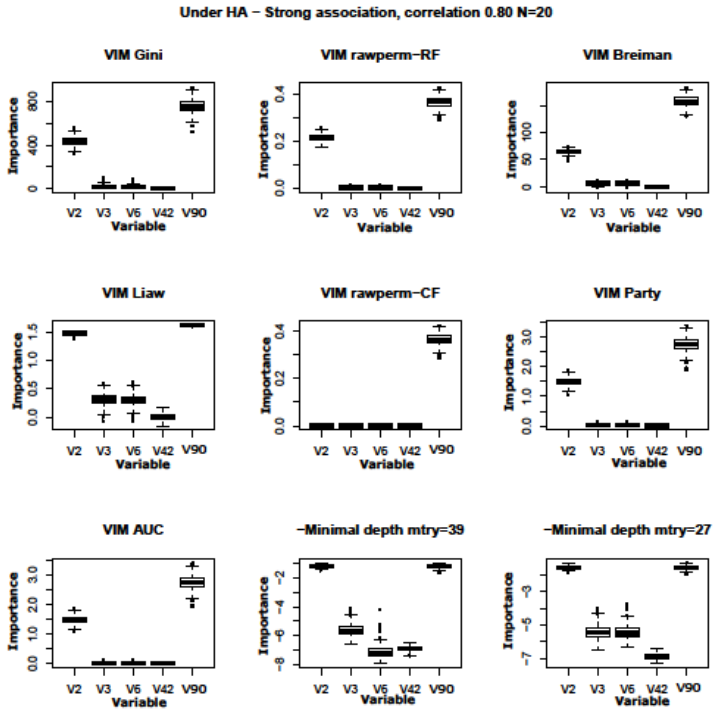

Figure A.28. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.80 and N = 20. Strong interaction study.

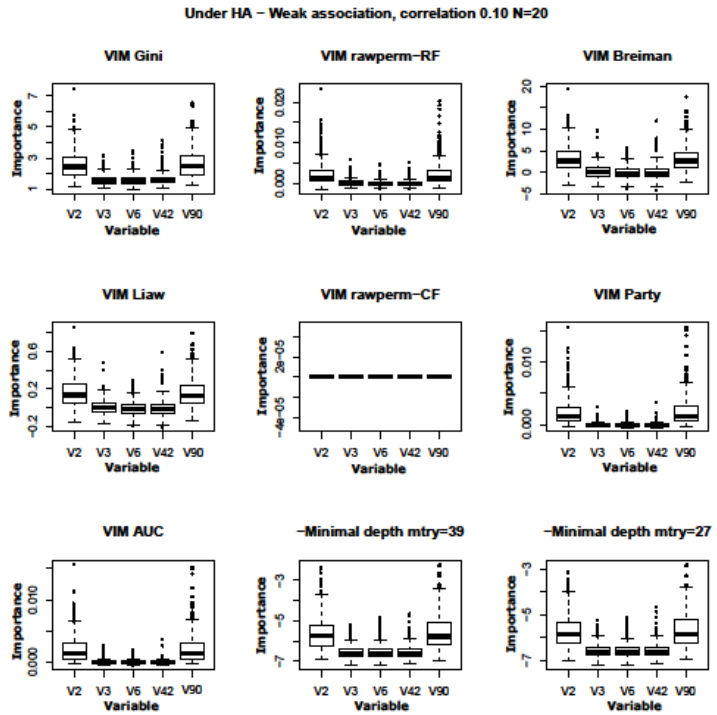

Figure A.29. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.10 and N = 20. Weak interaction study.

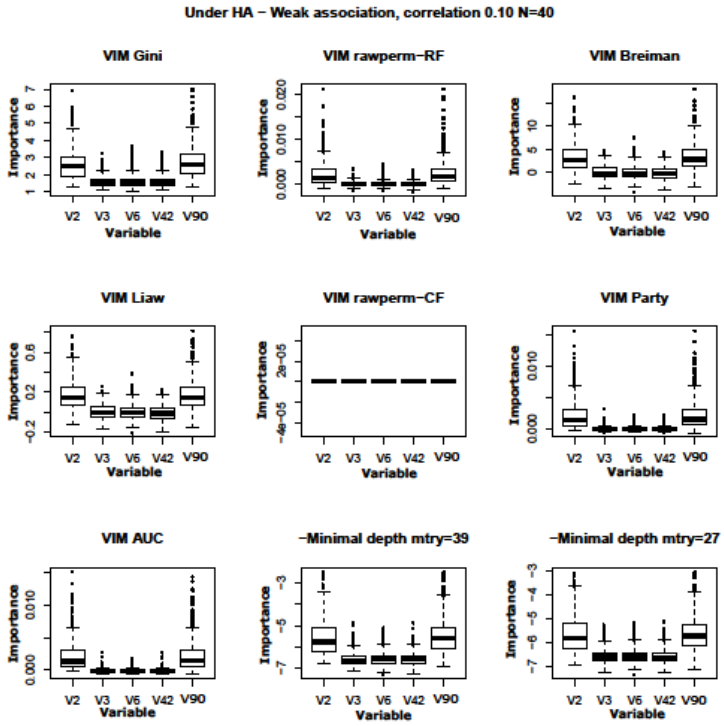

Figure A.30. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.10 and N = 40. Weak interaction study.

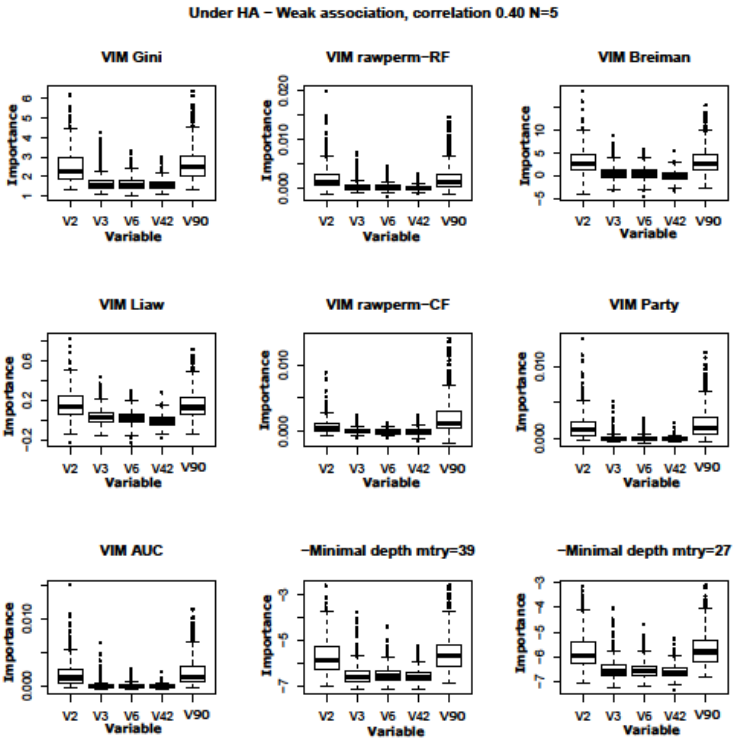

Figure A.31. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.40 and N = 5. Weak interaction study.

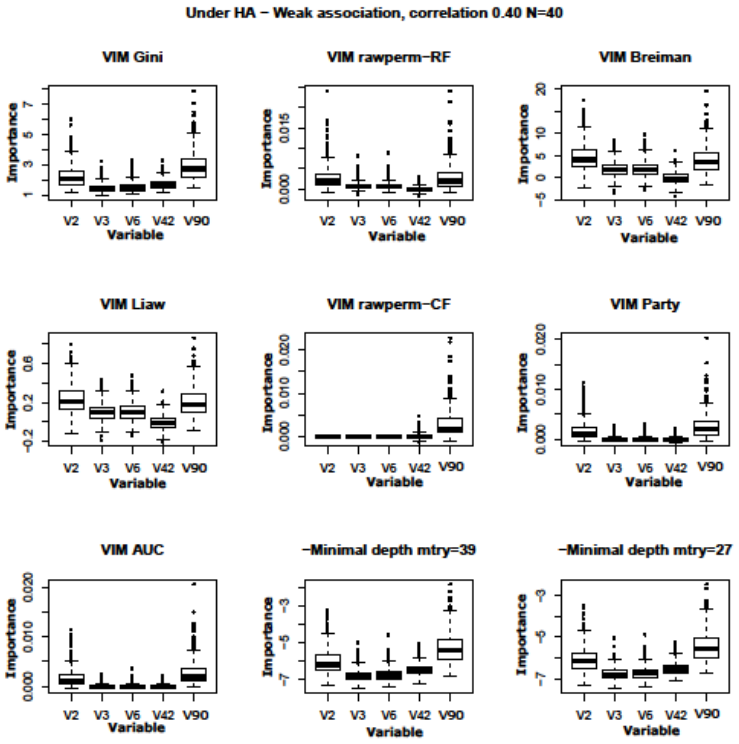

Figure A.32. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.40 and N = 40. Weak interaction study.

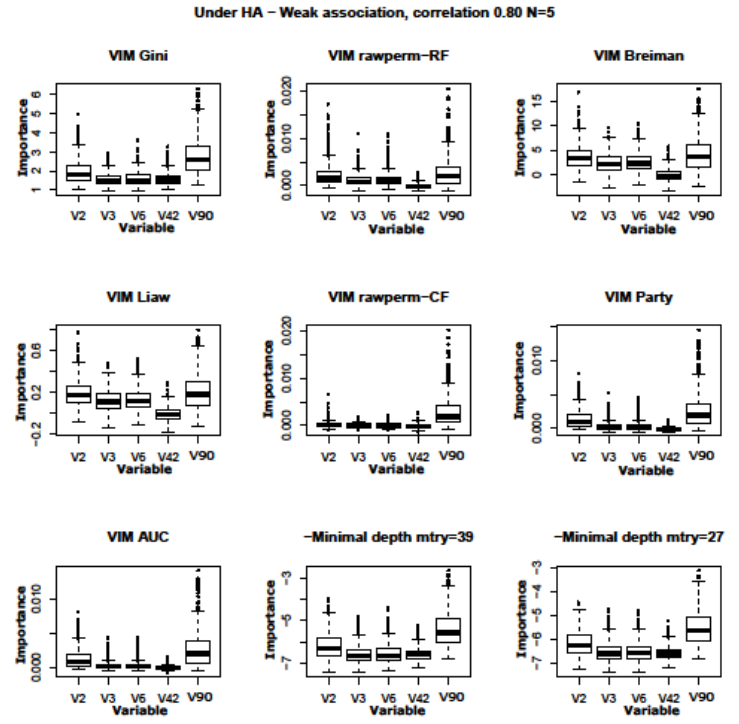

Figure A.33. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.80 and N = 5. Weak interaction study.

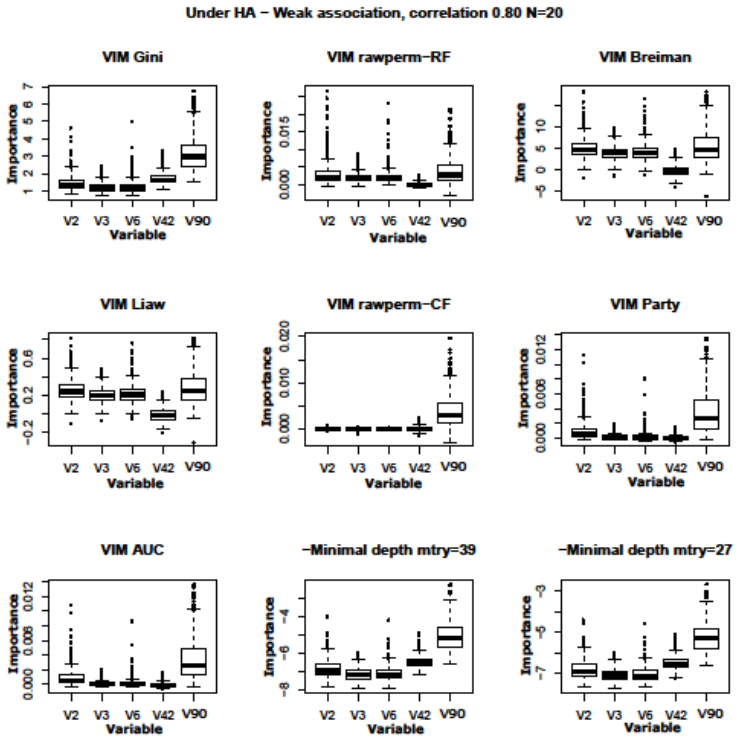

Figure A.34. RF VIMs, minimal depth, VIM_{AUC} and VIM_{party} under H_A for V_2 and V_{90}, for two variables correlated V_3 and V_6, and for two independent ones V_{42} when r = 0.80 and N = 20. Weak interaction study.

Codes Chapter 2

```
#Function to generate multivariate normal random variables it depends on the number of #observations,
on the mean, on the variance, on the correlation, mehthod for the matrix #decomposition
rmvnormc<-function (n, mean = rep(0, nrow(sigma)), sigma = diag(length(mean)),
    corr=diag(length(mean)), method = c("eigen", "svd", "chol"), pre0.9_9994 = FALSE)
{
  if (!isSymmetric(sigma, tol = sqrt(.Machine$double.eps), #if sigma is not simmetric, then stop
    check.attributes = FALSE)) {
    stop("sigma must be a symmetric matrix")
  }
  if (!isSymmetric(corr, tol = sqrt(.Machine$double.eps), #if corr is not simmetric, then stop
    check.attributes = FALSE)) {
    stop("corr must be a symmetric matrix")
  }
#if the length of the mean vector is different than the number of variables with sigma
  if (length(mean) != nrow(sigma)) {
    stop("mean and sigma have non-conforming size")
  }
#if the length of the vector is different than the number of variables with corr
  if (length(mean) != nrow(corr)) {
    stop("mean and corr have non-conforming size")
  }
#sigma taking into account the correlation
  sigmb<-matrix(0,nrow=nrow(sigma),ncol=ncol(sigma))
    for(i in 1:nrow(sigma)){
    for(j in 1:nrow(sigma)){
        sigmb[i,j]<-sqrt(diag(sigma)[i])*sqrt(diag(sigma)[j])*corr[i,j]
    }
        }
#see if the new sigma is simmetric
  dimnames(sigmb) <- NULL
  if (!isTRUE(all.equal(sigmb, t(sigmb)))) {
    warning("sigma is numerically not symmetric")
  }
#method for decomposition, the default is eigen
  method <- match.arg(method) #method for decomposition, the default is eigen
  if (method == "eigen") {
    ev <- eigen(sigmb, symmetric = TRUE)
    if (!all(ev$values >= -sqrt(.Machine$double.eps) * abs(ev$values[1]))) {
      warning("sigma is numerically not positive definite")
    }
    retval <- ev$vectors %*% diag(sqrt(ev$values), length(ev$values)) %*%
      t(ev$vectors)
  }
  else if (method == "svd") {
    sigsvd <- svd(sigmb)
    if (!all(sigsvd$d >= -sqrt(.Machine$double.eps) * abs(sigsvd$d[1]))) {
      warning("sigma is numerically not positive definite")
    }
    retval <- t(sigsvd$v %*% (t(sigsvd$u) * sqrt(sigsvd$d)))
  }
  else if (method == "chol") {
    retval <- chol(sigmb, pivot = TRUE)
    o <- order(attr(retval, "pivot"))
    retval <- retval[, o]
  }
  retval <- matrix(rnorm(n * ncol(sigmb)), nrow = n, byrow = !pre0.9_9994) %*%
    retval
  retval <- sweep(retval, 2, mean, "+")
  colnames(retval) <- names(mean)
  retval
}
###############################################################################
```

Code A.1. Function rmvnormc to generate multivariate normal predictors with correlation.

```
#Vector of the 100 means, zero means
media<-c(rep(0,100))
#matrix of the variance, identity matrix with 100x100 dimension
sigma<-diag(length(media))

#Number of variables correlated
N=40
#Strength of correlation
correlation=0.80

#Correlation matrix
#First a matrix for the correlated variables
corr1<-matrix(0,nrow=N,ncol=N)
for(i in 1:N){
        for(j in 1:N){
                if(i==j){
                        corr1[i,j]=1 # 1 for the correlation of each variable with itself
                }
                else if(i!=j){
                        if(i==3 | j==3){
                                # negative correlation of the 3rd variable with others
                                corr1[i,j]=-correlation
#the value of the correlation between the variable i and j is the correlation between j and i
                                corr1[j,i]=corr1[i,j]
                        }
                        else if(i!=3 & j!=3){
                                corr1[i,j]=correlation #all other variables are positive correlated
                                corr1[j,i]=corr1[i,j] #the same between j and i, and i and j
                        }
                }
        }
}
#the rest of the correlation matrix has 0 values between different variables
ceros1<-matrix(0,nrow=(100-N),ncol=N)
corr1<-rbind(corr1,ceros1)
ceros2<-matrix(0,nrow=N,ncol=(100-N))
corr2<-diag(100-N)
corr2<-rbind(ceros2,corr2)
corr<-cbind(corr1,corr2)

#number of databases
n=500

#loop to create the databases
for(i in 1:n){
        #generate the 100 standard normal variables with 1000 observations
        #with the mean, sigma and corr fix above
        x<-rmvnormc(1000,mean=media,sigma=sigma,corr=corr)
        #generate the error following a standard normal distribution
        e = rnorm(1000,0,0.5)
        #The first variable
        v2<-x[,1]

        #the outcome is generated as the linear model
        y=0.05*v2+e

#dataframe with the outcome and the predictors
data <-data.frame(cbind(y,x))
#save each database in each iteration
write.table(data, paste("data80_40.",i, sep=""), row.names=FALSE, quote=FALSE)
}
```

Code A.2. Code of the weakly-associated single data generation, r=0.80 and N=40 as an example.

```
#Vector of the 100 means, zero means
media<-c(rep(0,100))
#matrix of the variance, identity matrix with 100x100 dimension
sigma<-diag(length(media))

#Number of variables correlated
N=40
#Strength of correlation
correlation=0.80

#Correlation matrix
#First a matrix for the correlated variables
corr1<-matrix(0,nrow=N,ncol=N)
for(i in 1:N){
        for(j in 1:N){
                if(i==j){
                        corr1[i,j]=1 # 1 for the correlation of each variable with itself
                }
                else if(i!=j){
                        if(i==3 | j==3){
                                # negative correlation of the 3rd variable with others
                                corr1[i,j]=-correlation
#the value of the correlation between the variable i and j is the correlation between j and i
                                corr1[j,i]=corr1[i,j]
                        }
                        else if(i!=3 & j!=3){
                                corr1[i,j]=correlation #all other variables are positive correlated
                                corr1[j,i]=corr1[i,j] #the same between j and i, and i and j
                        }
                }
        }
}
#the rest of the correlation matrix has 0 values between different variables
ceros1<-matrix(0,nrow=(100-N),ncol=N)
corr1<-rbind(corr1,ceros1)
ceros2<-matrix(0,nrow=N,ncol=(100-N))
corr2<-diag(100-N)
corr2<-rbind(ceros2,corr2)
corr<-cbind(corr1,corr2)

#number of databases
n=500

#loop to create the databases
for(i in 1:n){
        #generate the 100 standard normal variables with 1000 observations
        #with the mean, sigma and corr fix above
        x<-rmvnormc(1000,mean=media,sigma=sigma,corr=corr)
        #generate the error following a standard normal distribution
        e = rnorm(1000,0,0.5)
        #The first variable (correlated)
        v2<-x[,1]
        #The 89th variable (uncorrelated)
        v90<-x[,89]
        #the outcome is generated as the linear model
        y=0.033*v2+0.033*v90+0.09*v2*v90+e
#dataframe with the outcome and the predictors
data <-data.frame(cbind(y,x))
#save each database in each iteration
write.table(data, paste("DATA80_40intweakfull033_09.",i, sep=""), row.names=FALSE, quote=FALSE)
}
```

Code A.3. Code of the weakly-associated interaction data generation, r=0.80 and N=40 as an example.

```
#Vector of the 100 means, zero means
media<-c(rep(0,100))
#matrix of the variance, identity matrix with 100x100 dimension
sigma<-diag(length(media))

#Number of variables correlated
N=40
#Strength of correlation
correlation=0.80

#Correlation matrix
#First a matrix for the correlated variables
corr1<-matrix(0,nrow=N,ncol=N)
for(i in 1:N){
        for(j in 1:N){
                if(i==j){
                        corr1[i,j]=1 # 1 for the correlation of each variable with itself
                }
                else if(i!=j){
                        if(i==3 | j==3){
                                # negative correlation of the 3rd variable with others
                                corr1[i,j]=-correlation
#the value of the correlation between the variable i and j is the correlation between j and i
                                corr1[j,i]=corr1[i,j]
                        }
                        else if(i!=3 & j!=3){
                                corr1[i,j]=correlation #all other variables are positive correlated
                                corr1[j,i]=corr1[i,j] #the same between j and i, and i and j
                        }
                }
        }
}


#the rest of the correlation matrix has 0 values between different variables
ceros1<-matrix(0,nrow=(100-N),ncol=N)
corr1<-rbind(corr1,ceros1)
ceros2<-matrix(0,nrow=N,ncol=(100-N))
corr2<-diag(100-N)
corr2<-rbind(ceros2,corr2)
corr<-cbind(corr1,corr2)

#number of databases
n=500

#loop to create the databases
for(i in 1:n){
        #generate the 100 standard normal variables with 1000 observations
        #with the mean, sigma and corr fix above
        x<-rmvnormc(1000,mean=media,sigma=sigma,corr=corr)
        #generate the error following a standard normal distribution
        e = rnorm(1000,0,0.5)

        #the outcome is generated as the linear model
        y=e
#dataframe with the outcome and the predictors
data <-data.frame(cbind(y,x))
#save each database in each iteration
write.table(data, paste("DATA80_40nullerror05.",i, sep=""), 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.

```
library(lmtest) #download the libray for the LRT tests
n=500 #number of databases
diff<-c() #Null vector for the bias
upper1<-c() #Null vector for the upper limit of the 95%CI
lower1<-c() #Null vector for the lower limit of the 95%CI
p<-c() #Null vector for the p-values
#loop to caculate them
for(i in 1:n){
        data<-read.table(paste("data80_40.",i,sep=""),header=T) #input data
        y=data[,1] #take the outcome
        x1=data[,2] #take V2 (variable associated)
        reg1<-glm(y~x1) #regression with V2
        reg0<-glm(y~1) #regression only the intercept
        a1<-summary(reg) #summary of the regression with V2
        diff<-append(diff,a1$coefficients[2,1]-0.05,after=length(diff)) #bias
        #upper limit of the 95%CI
        upper1<-append(upper1,a1$coef[2,1]+1.96*sqrt(a1$coef[2,2]^2),after=length(upper1))
        #lower limit of the 95%CI
        lower1<-append(lower1,a1$coef[2,1]-1.96*sqrt(a1$coef[2,2]^2),after=length(lower1))
        lrtc<-lrtest(reg1,reg0) #LRT test
        p<-append(p,lrtc$"Pr(>Chisq)"[2],after=length(p)) #LRT p-value
}
sig<-length(p[p<0.0001]) #Number of p-values less than Bonferroni threshold
insl1<-length(lower1[lower1>0.05]) #Number of expected values lower than the lower limit of the CI
#Number of expected values greater than the upper limit of the CI
insu1<-length(upper1[upper1<0.05])
cover1<-(insl1+insu1)*100/500 #sum of the points outside the CI
bias<-mean(diff) #mean of the bias
bias #bias
100-cover1 #percentage of the coverage
sig #number of p-values less than Bonferroni threshold
```

Code A.5. Code of the bias, coverage and p-values, r=0.80 and N=40 weakly-associated single study under H_A as an example.

```
#number of variables correlated
N=40
#number of datasets
n=500
#Null vector for the medians of the correlation
medic<-c(); medcind<-c()
#loop for each database
for(i in 1:n){
        data<-read.table(paste("data80_40strong.",i,sep=""),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
        medc<-numeric(N); meanc<-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,-j] #For each variable take the correlation with others
                medc[j]<-median(corri) #take the median of the correlation with others
        }
        #save the median in the vector for each database
        medic<-append(medic,median(medc),after=length(medic))
        for(h 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[h,-h] #For each variable take the correlation with others
                medcind[h]<-median(corri) #take the median of the correlation with others
        }
        #take the median of the absolute value of the correlation with others
        medcind<-append(medcind,median(medcind),after=length(medcind))

}
#median of the correlated
median(medic)
#median for the uncorrelated
median(medcind)
```

Code A.6. Code for the correlation, r=0.80 and N=40 strong single study as an example.

```
#number of databases
n=500

#Null vectors to take the 5% cut-off of each database
gininullw<-c();rawnullw<-c();breinullw<-c();liawnullw<-c();mengnullw<-c();condnullw<-c()
#take the median values of each VIM
tginiw<-rbind();traww<-rbind();tbreiw<-rbind();tliaww<-rbind();tmengw<-rbind();tcondw<-rbind()

#loop for each RF output
for(i in 1:n){

#Gini output
null1w<-read.table(paste(paste("nullout80_40nullerror05.",i,sep=""),".importance",sep=""),
header=TRUE)
#RF PVIM output
null2w<-read.table(paste(paste("nullout80_40nullerror05.",i,sep=""),".importance2",sep=""),
header=TRUE)
RF conditional PVIM output
null3w<-read.table(paste(paste("outcond80_40nullerror05k3.",i,sep=""),".importance2",sep=""),
header=TRUE)

    #take the 100 null values

giniw<-null1w[,4];raww<-null2w[,5];breiw<-null2w[,6];liaww<-null2w[,7];mengw<-null2w[,8]
condw<-null3w[,4]


        #order the importance values under the NULL

        oginiw<-giniw[order(giniw,decreasing=T)];oraww<-raww[order(raww,decreasing=T)]

        obreiw<-breiw[order(breiw,decreasing=T)];oliaww<-liaww[order(liaww,decreasing=T)]

        omengw<-mengw[order(mengw,decreasing=T)]

        ocondw<-condw[order(condw,decreasing=T)]


        #take the 5% cut-off

        cginiw<-oginiw[5];craww<-oraww[5];cbreiw<-obreiw[5];cliaww<-oliaww[5];

        cmengw<-omengw[5];   ccondw<-ocondw[5]


        #safe the cut-off of each database

        gininullw<-c(gininullw,cginiw);rawnullw<-c(rawnullw,craww);breinullw<-c(breinullw,cbreiw)

        liawnullw<-c(liawnullw,cliaww); mengnullw<-c(mengnullw,cmengw);

        condnullw<-c(condnullw,ccondw)
        #Null vector for the VIM scores

        giniwv<-c();rawwv<-c();breiwv<-c();liawwv<-c();mengwv<-c();condwv<-c()


for(j in 2:101){ #take the score for each variable
    on1w=null1w[,3]==paste("V",j,sep=""); on2w=null2w[,3]==paste("V",j,sep="")

    on3w=null3w[,3]==paste("V",j,sep="")

    giniwv<-c(giniwv,null1w[on1w,4]);rawwv<-c(rawwv,null2w[on2w,5]);

    breiwv<-c(breiwv,null2w[on2w,6]);liawwv<-c(liawwv,null2w[on2w,7])

    mengwv<-c(mengwv,null2w[on2w,8]);condwv<-c(condwv,null3w[on3w,4])

        }

        #save the scores for all variables

        tginiw<-rbind(tginiw,giniwv);traww<-rbind(traww,rawwv);tbreiw<-rbind(tbreiw,breiwv)

        tliaww<-rbind(tliaww,liawwv);tmengw<-rbind(tmengw,mengwv);

        tcondw<-rbind(tcondw,condwv)

}


#VIM party and AUC, and the minimal depth for both mtry values under H0

vimw<-read.table("VIM80_40nullerror05",header=T)

vimaucw<-read.table("VIMAUC80_40nullerror05",header=T)

mindw<-read.table("mindepth80_40_27nullerror05",header=T)

mindw39<-read.table("mindepth80_40_39nullerror05",header=T)


#Null vectors to save the cut-off

vimnullw<-c();vimaucnullw<-c();mindnullw39<-c();mindnullw<-c()

for(i in 1:n){

        #order the values and take the 5%

        cvimw<-vimw[i,][order(vimw[i,],decreasing=T)][5];

        cvimaucw<-vimaucw[i,][order(vimaucw[i,],decreasing=T)][5]

        #5% of the lower values for minimal depth

        cmindw39<-mindw39[i,][order(mindw39[i,])][5]

        cmindw<-mindw[i,][order(mindw[i,])][5]
```

```
        #save the values for each database
        vimnullw<-c(vimnullw,cvimw)
        vimaucnullw<-c(vimaucnullw,cvimaucw)
        mindnullw39<-c(mindnullw39,cmindw39)
        mindnullw<-c(mindnullw,cmindw)
}
#number of variables correlated
N=40

#take the median values of the importance for the correlated variables
corr<-
c(median(apply(tginiw,2,median)[1:N]),median(apply(traww,2,median)[1:N]),median(apply(tbreiw,2,median)[1:N]),
median(apply(tliaww,2,median)[1:N]),median(apply(tmengw,2,median)[1:N]),median(apply(tcondw,2,median)[1:N]),
median(apply(vimw,2,median)[1:N]),median(apply(vimaucw,2,median)[1:N]),median(apply(mindw,2,median)[1:N]),
median(apply(mindw39,2,median)[1:N]))
#take the median values of the importance for the uncorrelated variables
uncorr<-
c(median(apply(tginiw,2,median)[(N+1):100]),median(apply(traww,2,median)[(N+1):100]),median(apply(tbreiw,2,me
dian)[(N+1):100]),median(apply(tliaww,2,median)[(N+1):100]),median(apply(tmengw,2,median)[(N+1):100]),median
(apply(tcondw,2,median)[(N+1):100]),median(apply(vimw,2,median)[(N+1):100]),median(apply(vimaucw,2,median)[(
N+1):100]),median(apply(mindw,2,median)[(N+1):100]),median(apply(mindw39,2,median)[(N+1):100]))
##########################################
##Check if the cut-off was well-defined
#Null matrix to save the average of number of importance values above the cut-off for each iteration
gininullt<-rbind();rawnullt<-rbind();breinullt<-rbind();liawnullt<-rbind();mengnullt<-rbind();condnullt<-rbind()

#loop to take the number of null importance values are greater than the 5%cut-off (reject H0)
for(i in 1:n){
null1<-read.table(paste(paste("nullout80_40nullerror05.",i,sep=""),".importance",sep=""), header=TRUE) #Gini
null2<-read.table(paste(paste("nullout80_40nullerror05.",i,sep=""),".importance2",sep=""), header=TRUE) #RF PVIM
#Conditional PVIM
null3<-read.table(paste(paste("outcond80_40nullerror05k3.",i,sep=""),".importance2",sep=""), header=TRUE)
        #NULL vecctor for the average in each iteration
        ginit<-c();rawt<-c();breit<-c();liawt<-c();mengt<-c();condt<-c()
#check for all variables the number of times are greater than all 500 cut-off and divided by that total for each VIM
        for(j in 1:100){
                ginit<-c(ginit,sum(null1[j,4]>=gininullw)/500)
```

```
                rawt<-c(rawt,sum(null2[j,5]>=rawnullw)/500)
                breit<-c(breit,sum(null2[j,6]>=breinullw)/500)
                liawt<-c(liawt,sum(null2[j,7]>=liawnullw)/500)
                mengt<-c(mengt,sum(null2[j,8]>=mengnullw)/500)
                condt<-c(condt,sum(null3[j,4]>=condnullw)/500)
        }
        #save the average of all variables for each iteration for each VIM
        gininullt<-rbind(gininullt,ginit)
        rawnullt<-rbind(rawnullt,rawt)
        breinullt<-rbind(breinullt,breit)
        liawnullt<-rbind(liawnullt,liawt)
        mengnullt<-rbind(mengnullt,mengt)
        condnullt<-rbind(condnullt,condt)
}
#matrix to save the average of importance values in each iteration compared to all 500 cut-offs
vimntf<-cbind();vimaucntf<-cbind();mindntf39<-cbind();mindntf<-cbind()
#loop for each database
for(i in 1:n){
        #save the average in this vector for all variables in each database
        vimnt<-c();vimaucnt<-c();mindnt39<-c();mindnt<-c()
        #take the average for each variable
        for(j in 1:100){
        vimnt<-c(vimnt,sum(vimw[i,j]>=vimnullw)/500)
        vimaucnt<-c(vimaucnt,sum(vimaucw[i,j]>=vimaucnullw)/500)
        mindnt39<-c(mindnt39,sum(mindw39[i,j]<=mindnullw39)/500)
        mindnt<-c(mindnt,sum(mindw[i,j]<=mindnullw)/500)
        }
        #save the mean of all variables in each database in the matrix
        vimntf<-rbind(vimntf,vimnt)
        vimaucntf<-rbind(vimntf,vimaucnt)
        mindntf39<-rbind(vimntf,mindnt39)
        mindntf<-rbind(mindntf,mindnt)
}
```

```
        #calculate the percentage of the average across all 500 databases and for all variables (all non-associated) for each VIM
        ginite<-mean(apply(gininullt,2,mean))*100;rawte<-mean(apply(rawnullt,2,mean))*100
        breite<-mean(apply(breinullt,2,mean))*100;liawte<-mean(apply(liawnullt,2,mean))*100
        mengte<-mean(apply(mengnullt,2,mean))*100;condte<-mean(apply(condnullt,2,mean))*100
        vimte<-mean(apply(vimntf,2,mean))*100;vimaucte<-mean(apply(vimaucntf,2,mean))*100
        mindte39<-mean(apply(mindntf39,2,mean))*100;mindte<-mean(apply(mindntf,2,mean))*100

        #Take the importance values for V2, V3, V6, V42 and V90 under H0
        giniwp<-cbind(tginiw[,1],tginiw[,2],tginiw[,5],tginiw[,41],tginiw[,89])
        rawwp<-cbind(traww[,1],traww[,2],traww[,5],traww[,41],traww[,89])
        breiwp<-cbind(tbreiw[,1],tbreiw[,2],tbreiw[,5],tbreiw[,41],tbreiw[,89])
        liawwp<-cbind(tliaww[,1],tliaww[,2],tliaww[,5],tliaww[,41],tliaww[,89])
        condwp<-cbind(tcondw[,1],tcondw[,2],tcondw[,5],tcondw[,41],tcondw[,89])
        vimwp<-cbind(vimw[,1],vimw[,2],vimw[,5],vimw[,41],vimw[,89])
        vimaucwp<-cbind(vimaucw[,1],vimaucw[,2],vimaucw[,5],vimaucw[,41],vimaucw[,89])
        #negative values of minimal depth to show more values more association
        mindw39p<-cbind(-mindw39[,1],-mindw39[,2],-mindw39[,5],-mindw39[,41],-mindw39[,89])
        mindwp<-cbind(-mindw[,1],-mindw[,2],-mindw[,5],-mindw[,41],-mindw[,89])

        #column names for the matrix of the five variables
        colnames(giniwp)<-c("V2","V3","V6","V42","V90")
```

```
colnames(rawwp)<-c("V2","V3","V6","V42","V90")
colnames(breiwp)<-c("V2","V3","V6","V42","V90")
colnames(liawwp)<-c("V2","V3","V6","V42","V90")
colnames(condwp)<-c("V2","V3","V6","V42","V90")
colnames(vimwp)<-c("V2","V3","V6","V42","V90")
colnames(vimaucwp)<-c("V2","V3","V6","V42","V90")
colnames(mindw39p)<-c("V2","V3","V6","V42","V90")
colnames(mindwp)<-c("V2","V3","V6","V42","V90")

#create a pdf for the different VIMs under H0 in each correlation condition
pdf("VIMs_mindepth_80_40_H0plots.pdf")
par(oma=c(0,0,2,0))
par(mfrow=c(3,3))
boxplot(giniwp,main="VIM Gini",xlab="Variable",ylab="Importance")
boxplot(rawwp,main="VIM rawperm-RF",xlab="Variable",ylab="Importance")
boxplot(breiwp,main="VIM Breiman",xlab="Variable",ylab="Importance")
boxplot(liawwp,main="VIM Liaw",xlab="Variable",ylab="Importance")
boxplot(condwp,main="VIM rawperm-CF",xlab="Variable",ylab="Importance")
boxplot(vimwp,main="VIM Party",xlab="Variable",ylab="Importance")
boxplot(vimaucwp,main="VIM AUC",xlab="Variable",ylab="Importance")
boxplot(mindw39p,main="-Minimal depth mtry=39",xlab="Variable",ylab="Importance")
boxplot(mindwp,main="-Minimal depth mtry=27",xlab="Variable",ylab="Importance")
title(main="Under H0, correlation 0.80 N=40",outer=T)
dev.off()
```

Code A.7. Code for VIMs and minimal depth under H_0, 5% Cut-offs, and plot under H_0 for all VIMs and minimal depth when r=0.80 and N=40 as an example.

```
#number of iterations
n=500
#create empty vector to fill then with each iteration to save each cut-off
gininullw<-c();rawnullw<-c();breinullw<-c();liawnullw<-c();mengnullw<-c();condnullw<-c()

for(i in 1:n){ #loop for each RF output under the null
#call RF Gini ouput
null1w<-read.table(paste(paste("nullout80_40nullerror05.",i,sep=""),".importance",sep=""),
header=TRUE)
#call RF uscaled and scaled PVIMs output
null2w<-read.table(paste(paste("nullout80_40nullerror05.",i,sep=""),".importance2",sep=""),
header=TRUE)
#call RF conditional PVIM output
null3w<-read.table(paste(paste("outcond80_40nullerror05k3.",i,sep=""),".importance2",sep=""),
header=TRUE)
#take the VIMs under the null
        giniw<-null1w[,4];raww<-null2w[,5];breiw<-null2w[,6];liaww<-null2w[,7];mengw<-null2w[,8]
        condw<-null3w[,4]
#order the VIMs
        oginiw<-giniw[order(giniw,decreasing=T)]
        oraww<-raww[order(raww,decreasing=T)]
        obreiw<-breiw[order(breiw,decreasing=T)]
        oliaww<-liaww[order(liaww,decreasing=T)]
        omengw<-mengw[order(mengw,decreasing=T)]
        ocondw<-condw[order(condw,decreasing=T)]
#take the 5% cut-off for each dataset (there are 100 values)
        cginiw<-oginiw[5];craww<-oraww[5];cbreiw<-obreiw[5];cliaww<-oliaww[5];
        cmengw<-omengw[5];ccondw<-ocondw[5]
#save each cut-off in the vector
        gininullw<-c(gininullw,cginiw);rawnullw<-c(rawnullw,craww);breinullw<-c(breinullw,cbreiw)
        liawnullw<-c(liawnullw,cliaww); mengnullw<-c(mengnullw,cmengw);
        condnullw<-c(condnullw,ccondw)
}
#call the VIMs from CIF, AUC and party
vimw<-read.table("VIM80_40nullerror05",header=T)
vimaucw<-read.table("VIMAUC80_40nullerror05",header=T)
#call the minimal depth with both mtry values
mindw<-read.table("mindepth80_40_27nullerror05",header=T)
mindw39<-read.table("mindept80_40_39nullerror05",header=T)
#create NUll vector to then fill it with the cut-off values
vimnullw<-c();vimaucnullw<-c();mindnullw39<-c();mindnullw<-c()

#loop for each RF iteration
for(i in 1:n){
        #order the VIM and minimal depth values
        cvimw<-vimw[i,][order(vimw[i,],decreasing=T)][5]
        cvimaucw<-vimaucw[i,][order(vimaucw[i,],decreasing=T)][5]
        #in minimal depth the order is the opposite, as  less variance more importance
        cmindw39<-mindw39[i,][order(mindw39[i,])][5]
        cmindw<-mindw[i,][order(mindw[i,])][5]
        #fill the vector the value of each iteration
vimnullw<-c(vimnullw,cvimw)
        vimaucnullw<-c(vimaucnullw,cvimaucw)
        mindnullw39<-c(mindnullw39,cmindw39)
        mindnullw<-c(mindnullw,cmindw)
}
#create NUll matrices for RF PVIM under HA in the weak and strong studies for the power
ginialtw<-rbind();rawaltw<-rbind();breialtw<-rbind();liawaltw<-rbind();mengaltw<-rbind()
condaltw<-rbind();ginialts<-rbind();rawalts<-rbind();breialts<-rbind();liawalts<-rbind()
mengalts<-rbind();condalts<-rbind()
#create NUll matrices for RF PVIM under HA in the weak and strong studies for the median VIM
ginialtwv<-rbind();rawaltwv<-rbind();breialtwv<-rbind();liawaltwv<-rbind();mengaltwv<-rbind()
condaltwv<-rbind();ginialtsv<-rbind();rawaltsv<-rbind();breialtsv<-rbind();liawaltsv<-rbind()
mengaltsv<-rbind();condaltsv<-rbind()

#loop for each RF ouput under HA
for(i in 1:n){
        #Gini VIM for each database in the weak study
        imp1wa<-read.table(paste(paste("outcond80_40weak.",i,sep=""),".importance",sep=""),
header=TRUE)
        #unscaled and scaled RF PVIMS for each database in the weak study
        imp2wa<-read.table(paste(paste("outcond80_40weak.",i,sep=""),".importance2",sep=""),
header=TRUE)
        #RF conditional PVIM for each database in the weak study
        imp3wa<-read.table(paste(paste("out80_40weakcond75.",i,sep=""),".importance2",sep=""),
header=TRUE)

        #Gini VIM for each database in the strong study
        imp1sa<-read.table(paste(paste("out80_40strong.",i,sep=""),".importance",sep=""),
header=TRUE)
        #unscaled and scaled RF PVIMS for each database in the strong study
        imp2sa<-read.table(paste(paste("out80_40strong.",i,sep=""),".importance2",sep=""),
header=TRUE)
        #RF conditional PVIM for each database in the strong study
        imp3sa<-read.table(paste(paste("outcond80_40strongK3.",i,sep=""),".importance2",sep=""),
header=TRUE)

        #create NULL vector to fill them with the power of detecting each variable
        #in each database for the weak and the strong studies
        giniwa<-c();rawwa<-c();breiwa<-c();liawwa<-c();mengwa<-c();condwa<-c()
        ginisa<-c();rawsa<-c();breisa<-c();liawsa<-c();mengsa<-c();condsa<-c()

        #create NULL vector to fill them with the median VIM  of each variable
        #in each database for the weak and tha strong studies
        giniwv<-c();rawwv<-c();breiwv<-c();liawwv<-c();mengwv<-c();condwv<-c()
        ginisv<-c();rawsv<-c();breisv<-c();liawsv<-c();mengsv<-c();condsv<-c()
        for(j in 2:101){ #loop for each variable
        #take the row where each variable belongs to in each input
    on1wa=imp1wa[,3]==paste("V",j,sep="");on2wa=imp2wa[,3]==paste("V",j,sep="")
    on3wa=imp3wa[,3]==paste("V",j,sep="");on1sa=imp1sa[,3]==paste("V",j,sep="")
    on2sa=imp2sa[,3]==paste("V",j,sep=""); on3sa=imp3sa[,3]==paste("V",j,sep="")
```

```
#calculate the number of times the VIM for the variable is greater than the cut-off in both association
studies
    giniwa<-c(giniwa,sum(imp1wa[on1wa,4]>=gininullw)/500)
    rawwa<-c(rawwa,sum(imp2wa[on2wa,5]>=rawnullw)/500)
    breiwa<-c(breiwa,sum(imp2wa[on2wa,6]>=breinullw)/500)
    liawwa<-c(liawwa,sum(imp2wa[on2wa,7]>=liawnullw)/500)
    mengwa<-c(mengwa,sum(imp2wa[on2wa,8]>=mengnullw)/500)
    condwa<-c(condwa,sum(imp3wa[on3wa,4]>=condnullw)/500)
    ginisa<-c(ginisa,sum(imp1sa[on1sa,4]>=gininullw)/500)
    rawsa<-c(rawsa,sum(imp2sa[on2sa,5]>=rawnullw)/500)
    breisa<-c(breisa,sum(imp2sa[on2sa,6]>=breinullw)/500)
    liawsa<-c(liawsa,sum(imp2sa[on2sa,7]>=liawnullw)/500)
    mengsa<-c(mengsa,sum(imp2sa[on2sa,8]>=mengnullw)/500)
    condsa<-c(condsa,sum(imp3sa[on3sa,4]>=condnullw)/500)

#take the importance of each variable each database
    giniwv<-c(giniwv,imp1wa[on1wa,4]);rawwv<-c(rawwv,imp2wa[on2wa,5])
    breiwv<-c(breiwv,imp2wa[on2wa,6]);liawwv<-c(liawwv,imp2wa[on2wa,7])
    mengwv<-c(mengwv,imp2wa[on2wa,8]);condwv<-c(condwv,imp3wa[on3wa,4])
    ginisv<-c(ginisv,imp1sa[on1sa,4]); rawsv<-c(rawsv,imp2sa[on2sa,5])
    breisv<-c(breisv,imp2sa[on2sa,6]); liawsv<-c(liawsv,imp2sa[on2sa,7])
    mengsv<-c(mengsv,imp2sa[on2sa,8]);condsv<-c(condsv,imp3sa[on3sa,4])
        }
        #save the values in the matrices
        ginialtw<-rbind(ginialtw,giniwa);rawaltw<-rbind(rawaltw,rawwa)
        breialtw<-rbind(breialtw,breiwa);liawaltw<-rbind(liawaltw,liawwa)
        mengaltw<-rbind(mengaltw,mengwa);condaltw<-rbind(condaltw,condwa)
        ginialts<-rbind(ginialts,ginisa);rawalts<-rbind(rawalts,rawsa)
        breialts<-rbind(breialts,breisa);liawalts<-rbind(liawalts,liawsa)
        mengalts<-rbind(mengalts,mengsa);condalts<-rbind(condalts,condsa)

        ginialtwv<-rbind(ginialtwv,giniwv);rawaltwv<-rbind(rawaltwv,rawwv)
        breialtwv<-rbind(breialtwv,breiwv);liawaltwv<-rbind(liawaltwv,liawwv)
        mengaltwv<-rbind(mengaltwv,mengwv);condaltwv<-rbind(condaltwv,condwv)
        ginialtsv<-rbind(ginialtsv,ginisv);rawaltsv<-rbind(rawaltsv,rawsv)
        breialtsv<-rbind(breialtsv,breisv);liawaltsv<-rbind(liawaltsv,liawsv)
        mengaltsv<-rbind(mengaltsv,mengsv);condaltsv<-rbind(condaltsv,condsv)
}
#input of minimal depth in the weak study
mindw<-read.table("mindepth80_40_27weakfin",header=T)
mindw39<-read.table("mindepth80_40_39weakfin",header=T)
#input of the AUC and party in the weak study
vimw<-read.table("VIM80_40weak",header=T)
vimaucw<-read.table("VIMAUC80_40weak",header=T)
#inputs from the strong study
minds<-read.table("mindepth80_40_27strong",header=T)
minds39<-read.table("mindepth80_40_39strong",header=T)
vims<-read.table("VIM80_40strong",header=T)
vimaucs<-read.table("VIMAUC80_40strong",header=T)
```

```
#create null matrices to save the power of detecting each variables in all databse
vimaltw<-rbind();vimaucaltw<-rbind();mindaltw39<-rbind();mindaltw<-rbind()
vimalts<-rbind();vimaucalts<-rbind();mindalts39<-rbind();mindalts<-rbind()

#loop for each database
for(i in 1:n){
	#null vectors for the power of each variable in each database
	vimwa<-c();vimaucwa<-c();mindwa39<-c();mindwa<-c()
	vimsa<-c();vimaucsa<-c();mindsa39<-c();mindsa<-c()

	#take the number of times the VIM is greater than or equal to all cutt-offs
	#divided by the 500 total number of cut-offs in the weak and strong studies
	for(j in 1:100){
	vimwa<-c(vimwa,sum(vimw[i,j]>=vimnullw)/500)
	vimaucwa<-c(vimaucwa,sum(vimaucw[i,j]>=vimaucnullw)/500)
	mindwa39<-c(mindwa39,sum(mindw39[i,j]<=mindnullw39)/500)
	mindwa<-c(mindwa,sum(mindw[i,j]<=mindnullw)/500)
	vimsa<-c(vimsa,sum(vims[i,j]>=vimnullw)/500)
	vimaucsa<-c(vimaucsa,sum(vimaucs[i,j]>=vimaucnullw)/500)
	mindsa39<-c(mindsa39,sum(minds39[i,j]<=mindnullw39)/500)
	mindsa<-c(mindsa,sum(minds[i,j]<=mindnullw)/500)

	}
	#Fill the matrices with the power of all variables in each database
	vimaltw<-rbind(vimaltw,vimwa)
	vimaucaltw<-rbind(vimaucaltw,vimaucwa)
	mindaltw39<-rbind(mindaltw39,mindwa39)
	mindaltw<-rbind(mindaltw,mindwa)
	vimalts<-rbind(vimalts,vimsa)
	vimaucalts<-rbind(vimaucalts,vimaucsa)
	mindalts39<-rbind(mindalts39,mindsa39)
	mindalts<-rbind(mindalts,mindsa)
}
N=40 #number of variables correlated
#mean of the power across all 500 databases for V2, weak study
v2<-
c(mean(ginialtw[,1])*100,mean(rawaltw[,1])*100,mean(breialtw[,1])*100,mean(liawaltw[,1])*100,m
ean(condaltw[,1])*100,mean(vimaltw[,1])*100,mean(vimaucaltw[,1])*100,mean(mindaltw39[,1])*1
00,mean(mindaltw[,1])*100)

#median of VIM values for V2
v2v<-
c(median(ginialtwv[,1]),median(rawaltwv[,1]),median(breialtwv[,1]),median(liawaltwv[,1]),median(c
ondaltwv[,1]),median(vimw[,1]),median(vimaucw[,1]),median(mindw39[,1]),median(mindw[,1]))

#median of the median VIM for correlated variables
corrv<-
c(median(apply(ginialtwv,2,median)[2:N]),median(apply(rawaltwv,2,median)[2:N]),median(apply(br
eialtwv,2,median)[2:N]),median(apply(liawaltwv,2,median)[2:N]),median(apply(condaltwv,2,median)
[2:N]),median(apply(vimw,2,median)[2:N]),median(apply(vimaucw,2,median)[2:N]),median(apply(mi
ndw39,2,median)[2:N]),median(apply(mindw,2,median)[2:N]))
```

```
#median of the median VIM for uncorrelated variables
uncorrv<-
c(median(apply(ginialtwv,2,median)[(N+1):100]),median(apply(rawaltwv,2,median)[(N+1):100]),med
ian(apply(breialtwv,2,median)[(N+1):100]),median(apply(liawaltwv,2,median)[(N+1):100]),median(a
pply(condaltwv,2,median)[(N+1):100]),median(apply(vimw,2,median)[(N+1):100]),median(apply(vim
aucw,2,median)[(N+1):100]),median(apply(mindw39,2,median)[(N+1):100]),median(apply(mindw,2,
median)[(N+1):100]))

#the same in the strong association study
v2s<-
c(mean(ginialts[,1])*100,mean(rawalts[,1])*100,mean(breialts[,1])*100,mean(liawalts[,1])*100,mea
n(condalts[,1])*100,mean(vimalts[,1])*100,mean(vimaucalts[,1])*100,mean(mindalts39[,1])*100,me
an(mindalts[,1])*100)
v2sv<-
c(median(ginialtsv[,1]),median(rawaltsv[,1]),median(breialtsv[,1]),median(liawaltsv[,1]),median(cond
altsv[,1]),median(vims[,1]),median(vimaucs[,1]),median(minds39[,1]),median(minds[,1]))
corrsv<-
c(median(apply(ginialtsv,2,median)[2:N]),median(apply(rawaltsv,2,median)[2:N]),median(apply(brei
altsv,2,median)[2:N]),median(apply(liawaltsv,2,median)[2:N]),median(apply(condaltsv,2,median)[2:
N]),median(apply(vims,2,median)[2:N]),median(apply(vimaucs,2,median)[2:N]),median(apply(minds
39,2,median)[2:N]),median(apply(minds,2,median)[2:N]))
uncorrsv<-
c(median(apply(ginialtsv,2,median)[(N+1):100]),median(apply(rawaltsv,2,median)[(N+1):100]),media
n(apply(breialtsv,2,median)[(N+1):100]),median(apply(liawaltsv,2,median)[(N+1):100]),median(apply
(condaltsv,2,median)[(N+1):100]),median(apply(vims,2,median)[(N+1):100]),median(apply(vimaucs,2
,median)[(N+1):100]),median(apply(minds39,2,median)[(N+1):100]),median(apply(minds,2,median)[
(N+1):100]))

#VIM and negative minimal depth for V2, V3, V6, V42 and V90 under HA in the weak study
giniwp<-cbind(ginialtwv[,1],ginialtwv[,2],ginialtwv[,5],ginialtwv[,41],ginialtwv[,89])
rawwp<-cbind(rawaltwv[,1],rawaltwv[,2],rawaltwv[,5],rawaltwv[,41],rawaltwv[,89])
breiwp<-cbind(breialtwv[,1],breialtwv[,2],breialtwv[,5],breialtwv[,41],breialtwv[,89])
liawwp<-cbind(liawaltwv[,1],liawaltwv[,2],liawaltwv[,5],liawaltwv[,41],liawaltwv[,89])
condwp<-cbind(condaltwv[,1],condaltwv[,2],condaltwv[,5],condaltwv[,41],condaltwv[,89])
vimwp<-cbind(vimw[,1],vimw[,2],vimw[,5],vimw[,41],vimw[,89])
vimaucwp<-cbind(vimaucw[,1],vimaucw[,2],vimaucw[,5],vimaucw[,41],vimaucw[,89])
mindwp39<-cbind(-mindw39[,1],-mindw39[,2],-mindw39[,5],-mindw39[,41],-mindw39[,89])
mindwp<-cbind(-mindw[,1],-mindw[,2],-mindw[,5],-mindw[,41],-mindw[,89])

#VIM and negative minimal depth for V2, V3, V6, V42 and V90 under HA in the strong study
ginisp<-cbind(ginialtsv[,1],ginialtsv[,2],ginialtsv[,5],ginialtsv[,41],ginialtsv[,89])
rawsp<-cbind(rawaltsv[,1],rawaltsv[,2],rawaltsv[,5],rawaltsv[,41],rawaltsv[,89])
breisp<-cbind(breialtsv[,1],breialtsv[,2],breialtsv[,5],breialtsv[,41],breialtsv[,89])
liawsp<-cbind(liawaltsv[,1],liawaltsv[,2],liawaltsv[,5],liawaltsv[,41],liawaltsv[,89])
condsp<-cbind(condaltsv[,1],condaltsv[,2],condaltsv[,5],condaltsv[,41],condaltsv[,89])
vimsp<-cbind(vims[,1],vims[,2],vims[,5],vims[,41],vims[,89])
vimaucsp<-cbind(vimaucs[,1],vimaucs[,2],vimaucs[,5],vimaucs[,41],vimaucs[,89])
mindsp39<-cbind(-minds39[,1],-minds39[,2],-mindw39[,5],-minds39[,41],-minds39[,89])
mindsp<-cbind(-minds[,1],-minds[,2],-minds[,5],-minds[,41],-minds[,89])
#Column names for the matrix of the VIMs for the 5 variables in the weak study
colnames(giniwp)<-c("V2","V3","V6","V42","V90")
colnames(rawwp)<-c("V2","V3","V6","V42","V90")
colnames(breiwp)<-c("V2","V3","V6","V42","V90")
colnames(liawwp)<-c("V2","V3","V6","V42","V90")
colnames(condwp)<-c("V2","V3","V6","V42","V90")
colnames(vimwp)<-c("V2","V3","V6","V42","V90")
colnames(vimaucwp)<-c("V2","V3","V6","V42","V90")
colnames(mindwp39)<-c("V2","V3","V6","V42","V90")
colnames(mindwp)<-c("V2","V3","V6","V42","V90")
#Column names for the matrix of the VIMs for the 5 variables in the strong study
colnames(ginisp)<-c("V2","V3","V6","V42","V90")
colnames(rawsp)<-c("V2","V3","V6","V42","V90")
colnames(breisp)<-c("V2","V3","V6","V42","V90")
colnames(liawsp)<-c("V2","V3","V6","V42","V90")
colnames(condsp)<-c("V2","V3","V6","V42","V90")
colnames(vimsp)<-c("V2","V3","V6","V42","V90")
colnames(vimaucsp)<-c("V2","V3","V6","V42","V90")
colnames(mindsp39)<-c("V2","V3","V6","V42","V90")
colnames(mindsp)<-c("V2","V3","V6","V42","V90")

#plots for the weak study under the HA
pdf("VIMs_mindepth_80_40_HAplots_weak.pdf")
par(oma=c(0,0,2,0))
par(mfrow=c(3,3))
boxplot(giniwp,main="VIM Gini",xlabel="Variable",ylabel="Importance")
boxplot(rawwp,main="VIM rawperm-RF",xlabel="Variable",ylabel="Importance")
boxplot(breiwp,main="VIM Breiman",xlabel="Variable",ylabel="Importance")
boxplot(liawwp,main="VIM Liaw",xlabel="Variable",ylabel="Importance")
boxplot(condwp,main="VIM rawperm-CF",xlabel="Variable",ylabel="Importance")
boxplot(vimwp,main="VIM Party",xlabel="Variable",ylabel="Importance")
boxplot(vimaucwp,main="VIM AUC",xlabel="Variable",ylabel="Importance")
boxplot(mindwp39,main="-Minimal depth mtry=39",xlabel="Variable",ylabel="Importance")
boxplot(mindwp,main="-Minimal depth mtry=27",xlabel="Variable",ylabel="Importance")
title(main="Under HA - Weak association, correlation 0.80 N=40",outer=T)
dev.off()
#plots for the weak study under the HA
pdf("VIMs_mindepth_80_40_HAplots_strong.pdf")
par(oma=c(0,0,2,0))
par(mfrow=c(3,3))
boxplot(ginisp,main="VIM Gini",xlabel="Variable",ylabel="Importance")
boxplot(rawsp,main="VIM rawperm-RF",xlabel="Variable",ylabel="Importance")
boxplot(breisp,main="VIM Breiman",xlabel="Variable",ylabel="Importance")
boxplot(liawsp,main="VIM Liaw",xlabel="Variable",ylabel="Importance")
boxplot(condsp,main="VIM rawperm-CF",xlabel="Variable",ylabel="Importance")
boxplot(vimsp,main="VIM Party",xlabel="Variable",ylabel="Importance")
boxplot(vimaucsp,main="VIM AUC",xlabel="Variable",ylabel="Importance")
boxplot(mindsp39,main="-Minimal depth mtry=39",xlabel="Variable",ylabel="Importance")
boxplot(mindsp,main="-Minimal depth mtry=27",xlabel="Variable",ylabel="Importance")
title(main="Under HA - Strong association, correlation 0.80 N=40",outer=T)
dev.off()
```

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.

Appendix B

Tables chapter 3

H0, different studies	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Xj~N(0,1) e~N(0,1)	62.9221	62.8311	62.7081	62.678	62.6089	63.2982	62.859	63.3346	62.8924	63.361
Different precision	52.4815	62.2177	64.2041	64.0603	64.2184	64.5318	64.6213	64.5983	64.3695	64.0536
Different variance	62.6818	62.617	62.4143	62.949	62.7934	62.7979	63.2267	62.9403	63.0363	62.6844
e~N(0,0.5)	15.8618	15.611	15.5937	15.6527	15.8198	15.8157	15.7157	15.6832	15.6893	15.7281

Table B.1. Median of VIM_{Gini} for all ten variables in the four different studies under H_0. Outcome continuous.

H0, different studies	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Xj~N(0,1) e~N(0,1)	31.2927	31.4244	31.3612	31.5361	31.4132	31.5126	31.4971	31.3432	31.5011	31.5486
Different precision	26.1307	31.2199	32.1809	32.3349	32.1453	32.1277	32.0929	32.1071	32.0568	32.1136
Different variance	31.3959	31.5198	31.5403	31.4252	31.4771	31.4414	31.3482	31.4009	31.4868	31.4201
e~N(0,0.5)	31.3662	31.5319	31.4082	31.4965	31.511	31.5271	31.5193	31.3521	31.3288	31.334

Table B.2. Median of VIM_{Gini} for all ten variables in the four different studies under H_0. Outcome binary.

HA, xj~N(0,1), e~N(0,1)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1 associated	119.023	63.1907	63.2225	63.237	63.411	63.731	63.3299	63.3741	63.5434	63.6112
X2 associated	62.5475	117.177	63.1652	62.7548	62.5191	63.1882	63.1568	63.3188	63.1333	63.0636
X3 associated	62.679	62.7183	118.499	63.1459	62.9821	63.2999	63.298	63.211	63.0888	63.0716
X4 associated	63.0417	62.527	62.5774	117.925	63.037	62.9427	62.7206	62.9106	63.3407	63.1398
X5 associated	63.0273	62.8677	63.5761	62.9983	117.015	63.6788	63.2698	63.1303	63.2412	62.8546
X6 associated	62.8513	62.7945	62.5644	63.3541	63.2972	117.479	63.0979	62.8141	63.0488	63.2025
X7 associated	62.8663	63.2159	63.1041	62.7904	63.3521	63.5549	116.711	62.8946	63.0605	62.9786
X8 associated	62.9558	62.3869	62.9407	62.7225	62.9905	63.088	62.5413	117.95	63.1133	62.8502
X9 associated	62.9768	62.6551	62.5283	62.9138	62.9857	63.157	63.1918	63.2856	116.743	62.9745
X10 associated	63.1169	62.5745	62.9598	62.7475	63.1775	63.1782	63.2238	63.4287	63.0448	117.561

Table B.3. Median of VIM_{Gini} for all ten variables when all variables and the error are standard normal distributed under H_A in the ten association studies. Outcome continuous.

HA, xj~N(0,1), e~N(0,1)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1 associated	46.034	29.8585	29.9278	29.8821	29.7525	29.7971	29.932	29.8084	29.8824	29.8144
X2 associated	29.8273	45.857	29.822	29.855	29.8362	29.8717	29.9198	29.7442	29.639	29.9039
X3 associated	29.9533	29.7351	45.9971	29.8904	29.8681	29.9071	29.8305	29.735	29.899	29.7792
X4 associated	29.8286	29.9293	29.8451	46.0523	29.7872	29.8421	29.7835	29.8908	29.9046	29.8158
X5 associated	29.7303	29.7013	29.9555	29.8653	45.9434	29.8614	29.9575	29.8405	29.7226	30.0451
X6 associated	29.882	29.9596	29.7682	29.8944	29.8874	46.2839	29.7625	29.9316	29.8672	29.8152
X7 associated	29.8547	29.7806	29.767	29.8967	29.7567	29.7129	46.2134	29.8966	29.7968	29.8351
X8 associated	29.869	29.7679	29.8648	29.8724	29.8507	29.75	29.7657	46.4996	29.6813	29.9282
X9 associated	29.8009	29.7786	29.8447	29.8601	29.8204	29.8616	29.7647	29.7746	46.2893	29.6852
X10 associated	29.8222	29.7605	29.7045	29.8975	29.848	29.7855	29.8227	29.849	29.8035	46.3124

Table B.4. Median of VIM_{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.

HA, Different precision	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1 associated	100.339	63.3089	65.3545	65.3852	65.2904	65.5001	65.2534	65.0637	65.605	65.0978
X2 associated	53.3511	116.49	64.3514	64.662	64.1794	64.5586	64.5335	64.5422	65.089	64.328
X3 associated	53.2001	62.6837	119.033	64.445	64.213	64.3208	64.4411	64.4951	64.8469	64.0126
X4 associated	53.3683	62.9945	63.8048	119.937	63.9796	64.7709	64.3922	64.0794	64.2445	64.2556
X5 associated	53.4403	62.7712	64.0174	64.1662	120.141	64.7326	64.5422	64.2086	64.427	64.0494
X6 associated	53.1937	62.5086	63.8525	64.3627	64.0856	119.362	63.9776	64.448	64.2555	64.2571
X7 associated	53.3783	62.5451	63.7194	64.1364	64.4073	64.7211	120.262	64.4709	64.3918	64.0277
X8 associated	52.7411	62.6389	63.9117	64.3752	64.3256	64.3619	64.1936	121.116	64.2059	64.3668
X9 associated	53.0516	62.5846	63.7718	64.4168	64.2923	64.6197	64.3321	64.366	119.014	64.1118
X10 associated	53.3321	62.4944	63.736	64.0733	64.5771	64.7573	64.8163	64.4405	64.3617	120.48

Table B.5. Median of VIM_{Gini} when all variables follow 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.

HA, Different precision	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1 associated	37.8389	30.1165	30.6864	30.7436	30.8254	30.6854	31.0194	30.7516	30.7562	30.8123
X2 associated	24.9635	45.7278	30.3591	30.6587	30.5433	30.6436	30.6012	30.4329	30.5116	30.4371
X3 associated	24.9625	29.749	47.053	30.3972	30.3246	30.419	30.4348	30.4258	30.6652	30.5705
X4 associated	24.9897	29.7508	30.3816	47.0267	30.5116	30.5338	30.4641	30.4713	30.4909	30.454
X5 associated	24.8532	29.6988	30.5625	30.6373	46.3625	30.4685	30.4158	30.5343	30.3497	30.5045
X6 associated	24.9074	29.7312	30.5166	30.5588	30.3915	46.8233	30.404	30.5319	30.4978	30.4396
X7 associated	24.8039	29.6161	30.4149	30.5654	30.5703	30.5467	46.7827	30.3272	30.6434	30.3897
X8 associated	24.9908	29.6175	30.3913	30.4642	30.5519	30.48	30.5899	46.7833	30.3474	30.593
X9 associated	24.8194	29.7246	30.3867	30.6837	30.4862	30.4382	30.4648	30.4223	46.9143	30.4602
X10 associated	24.9239	29.737	30.4726	30.5497	30.541	30.4284	30.6795	30.4513	30.4251	46.5528

Table B.6. Median of VIM_{Gini} when all variables follow 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 binary.

HA, Different variance	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1 associated	2478.86	107.185	107.198	107.51	107.854	106.114	107.855	107.749	107.819	106.827
X2 associated	104.939	2228.02	104.579	104.744	104.589	105.195	104.383	105.144	104.703	105.62
X3 associated	99.3306	97.6085	1999.94	98.7569	98.9127	99.0391	99.6391	98.125	99.3194	99.1639
X4 associated	94.1938	94.0224	93.4793	1760.41	94.5221	94.0474	93.4961	94.147	93.9245	93.341
X5 associated	89.746	88.3718	89.314	89.0518	1535.15	88.5215	89.2763	88.9912	89.8794	88.5399
X6 associated	84.8603	84.8254	85.0341	84.428	85.2924	1281.3	85.607	85.2379	85.6085	85.295
X7 associated	79.706	79.2596	79.2412	80.1114	80.1693	79.5691	1035.64	79.7442	80.4661	79.3398
X8 associated	75.4034	75.4049	75.4757	75.4771	75.2774	75.228	75.4661	797.889	76.1737	76.2123
X9 associated	70.1521	70.2061	69.9071	70.0853	70.4992	70.5206	70.7234	70.2437	559.598	70.276
X10 associated	62.8589	62.6819	63.0586	62.837	62.4103	63.1241	62.9535	63.0095	62.8967	118.463

Table B.7. Median of VIM_{Gini} for all ten variables when all variables follow a standard normal distribution but each one with different variance ($\sum$=diag(50,45,40,35,30,25,20,15,10,1)), under H_A in the ten association studies. Outcome continuous.

HA, Different variance	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1 associated	192.07	13.5788	13.6632	13.6225	13.6031	13.7239	13.7395	13.6045	13.6804	13.663
X2 associated	14.0464	188.117	14.1736	14.0745	14.1429	14.085	14.1144	14.1025	14.1933	14.0332
X3 associated	14.7323	14.7162	183.07	14.7164	14.705	14.567	14.6141	14.6707	14.6867	14.6823
X4 associated	15.4285	15.4204	15.4359	176.281	15.5682	15.3853	15.4746	15.5394	15.5274	15.4204
X5 associated	16.2572	16.1905	16.3007	16.3225	168.627	16.3194	16.2745	16.3634	16.2491	16.2796
X6 associated	17.2422	17.2055	17.2622	17.3333	17.2865	159.901	17.22	17.2538	17.2085	17.311
X7 associated	18.4044	18.5416	18.6428	18.4906	18.401	18.3707	149.184	18.3864	18.367	18.547
X8 associated	20.3339	20.2242	20.2796	20.1379	20.2333	20.2931	20.1831	133.448	20.0142	20.2415
X9 associated	22.2464	22.2372	22.4644	22.3367	22.3366	22.2368	22.491	22.3391	114.405	22.3227
X10 associated	29.9166	29.9386	29.7888	29.731	29.8383	29.7921	29.8748	29.7025	29.7869	46.4456

Table B.8. Median of VIM_{Gini} for all variables that follow a normal distribution but each one with different variance (∑=diag(50,45,40,35,30,25,20,15,10,1)), under H_A in the ten association studies. Outcome binary.

HA, xj~N(0,1), e~N(0,0.5)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1 associated	67.6897	16.2471	16.1967	16.3196	16.3499	16.2883	16.349	16.4003	16.3229	16.2318
X2 associated	16.5012	66.2296	16.3807	16.4226	16.3982	16.4897	16.3424	16.35	16.4249	16.3547
X3 associated	16.3894	16.2738	67.0189	16.3466	16.3161	16.4492	16.3228	16.4059	16.3974	16.3131
X4 associated	16.348	16.2987	16.3251	66.6842	16.3489	16.3223	16.34	16.2988	16.4039	16.3617
X5 associated	16.3777	16.2364	16.3239	16.186	67.3314	16.3514	16.3151	16.1869	16.3958	16.1386
X6 associated	16.2244	16.2798	16.214	16.2669	16.2591	66.6315	16.3148	16.2981	16.3193	16.2756
X7 associated	16.2456	16.1754	16.247	16.2412	16.3172	16.2577	66.7554	16.3264	16.3712	16.3603
X8 associated	16.4698	16.2417	16.3274	16.2994	16.3316	16.4281	16.2094	66.9516	16.26	16.2681
X9 associated	16.1486	16.2261	16.1957	16.1846	16.3886	16.2541	16.2626	16.3153	67.5406	16.2153
X10 associated	16.3004	16.1829	16.3318	16.3498	16.2686	16.4334	16.434	16.3245	16.3033	66.8746

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.

HA, xj~N(0,1), e~N(0,0.5)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1 associated	76.8576	26.4452	26.5369	26.5626	26.5479	26.5847	26.5033	26.3697	0.60985	26.4347
X2 associated	26.5406	76.9216	26.3962	26.4851	26.4421	26.5497	26.6054	26.4134	26.4964	26.4993
X3 associated	26.5043	26.4783	76.7638	26.2937	26.3489	26.6336	26.3601	26.3837	26.5017	26.3765
X4 associated	26.6472	26.4945	26.4834	76.6708	26.6288	26.4448	26.4585	26.345	26.4076	26.372
X5 associated	26.5274	26.5433	26.5267	26.4882	76.5348	26.5618	26.5647	26.4181	26.3993	26.4744
X6 associated	26.4525	26.3641	26.4819	26.4283	26.5529	76.653	26.505	26.4112	26.3795	26.4071
X7 associated	26.6347	26.5451	26.8141	26.5018	26.5375	26.521	75.8759	26.4165	26.4849	26.5811
X8 associated	26.4354	26.416	26.4245	26.45	26.709	26.4755	26.4235	76.547	26.3731	26.4899
X9 associated	26.4601	26.4027	26.6163	26.3814	26.4427	26.5171	26.5543	26.5254	76.7394	26.4801
X10 associated	26.509	26.4902	26.5218	26.55	26.4126	26.4232	26.4516	26.3276	26.4291	76.7302

Table B.10. Median of VIM_{Gini} for all ten variables when all variables follow a standard normal distribution but error has less variance N(0,0.5), under H_A in the ten association studies. Outcome binary.

Codes chapter 3

```
#Number of databases
n=500
#Import the library to generate multivariate normal distribution (independent)
library(mvtnorm)
#Import the library for the binary data, to dichotomize with threshold zero.
library(bindata)
#All ten variables are standard normal
media=rep(0,10)
sigma<-diag(length(media))

#Create each database
for(i in 1:n){
        #10 independent standard normal with 1000 observations
        x<-rmvnorm(1000,mean=media,sigma=sigma)
        e = rnorm(1000,0,1) #the error follow a standard normal
        yo=e #Under the null, the outcome
        y<-ra2ba(yo)     #Create a binary outcome with with threshold 0
        data <- data.frame(cbind(y,x)) #dataframe with the outcome and the predictors
#Save each database
write.table(data, paste("DATAginibinnullstdnormal.",i, sep=""), row.names=FALSE, quote=FALSE)
}
#######
#Now with an error with less variance
#create the databases with error ~ N(0,0.5), standard deviation equals to 0.5
for(i in 1:n){
        #10 independent standard normal with 1000 observations
        x<-rmvnorm(1000,mean=media,sigma=sigma)
        #generate normal distributed error with mean 0 and standard deviation 0.5
        e = rnorm(1000,0,0.5)
        yo=e #Under the null, the outcome
        y<-ra2ba(yo) #Create a binary outcome with with threshold 0
data <- data.frame(cbind(y,x)) #dataframe with the outcome and the predictors
#Save each database
write.table(data, paste("DATAginibinnullstdnormalerror05.",i, sep=""), row.names=FALSE,
quote=FALSE)
}
##########################
### predictor with different variance

#the mean 0 and each variable different variance
media=c(rep(0,10))
sigma=diag(c(50,45,40,35,30,25,20,15,10,1))

#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=media,sigma=sigma)
        e = rnorm(1000,0,1)
        yo=e # the outcome under the null
        y<-ra2ba(yo) #Create a binary outcome with with threshold 0
        #dataframe with the outcome and the predictors
        data <- data.frame(cbind(y,x))
#Save each database
write.table(data, paste("DATAginibinnulldiffvariance.",i,sep=""), row.names=FALSE, quote=FALSE)
}
}
```

```
############################
### Different number of decimal places
#Set the Mean to 0 and variance to 1
media=c(rep(0,10))
sigma=diag(length(media))

for(i in 1:n){
        #10 predictors follow a standard normal distribution
        xo<-rmvnorm(1000,mean=media,sigma=sigma)
        x<-cbind() #create a new matrix xo variables but different number of decimal places
        for(j in 1:10){
                #the variable xoj has j decimal places
                x1<-c(round(xo[,j],j))
                x<-cbind(x,x1)
        }
        #column names for each predictor
        colnames(x)<-c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10")

        e = rnorm(1000,0,1) #standard normal error
        yo=e #NUll association
        y<-ra2ba(yo) #Binary outcome
        data <- data.frame(cbind(y,x)) #daframe with predictors and outcome
#Save each database
write.table(data, paste("DATAginibinnullstdnormalcutpoint.",i, sep=""), row.names=FALSE,
quote=FALSE)
}
```

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 ~ N(0,1) and the error has less variance.

```
#500 databases
n=500

#Import the library to generate multivariate normal distribution (independent)
library(mvtnorm)
#Import the library for the binary data, to dichotomize with threshold zero.
library(bindata)

#Mean 0 and variance 1
media=rep(0,10)
sigma<-diag(length(media))

for(j in 1:10){ #Loop for each association study
        for(i in 1:n){ #500 databases under each association study
        #generate 10 predictors and error standard normal distributed
        x<-rmvnorm(1000,mean=media,sigma=sigma)
        e = rnorm(1000,0,1)
        #association study, beta is 0.3 and xj is the variable associated
        yo=0.3*x[,j]+e
        #dichotomize the outcome
        y<-ra2ba(yo)
        #data frame with the outcome and predictors
         data <- data.frame(cbind(y,x))

#save each data frame
write.table(data, paste(paste(paste("DATAginibinstdnormal03_",j, sep=""),".",sep=""),i,sep=""),
row.names=FALSE, quote=FALSE)
}
}
##############################
### Error less variance
for(j in 1:10){ #For each association stud
        #for each database
        for(i in 1:n){
         x<-rmvnorm(1000,mean=media,sigma=sigma) #generate standard normal predictors
        #generate normal error with mean 0 and standard deviation 0.5
        e = rnorm(1000,0,0.5)
        #the variable j is associated with coefficient 0.3
        yo=0.3*x[,j]+e
        y<-ra2ba(yo) #binary outcome
        data <- data.frame(cbind(y,x)) #data frame the outcome and predictors
#save each database
write.table(data,paste(paste(paste("DATAginibinstdnormal03error05_",j, sep=""),".",sep=""),i,sep=""),
row.names=FALSE, quote=FALSE)
}
##############################
### predictors with different variance

media=c(rep(0,10)) #save the mean 0
sigma=diag(c(50,45,40,35,30,25,20,15,10,1)) # different sigma for each variable
for(j in 1:10){ #Loop for each association study
        for(i in 1:n){ #loop to create 500 databases in each association study
        #multivariate standard normal, 1000 observations and 10 predictors
        x<-rmvnorm(1000,mean=media,sigma=sigma)
        x1<-x[,j] #take the predictor that is going to be associated with outcome
        e = rnorm(1000,0,1) #standard normal error
        yo=0.3*x1+e #the outcome is model by the following linear model
        y<-ra2ba(yo) #transform to a binary variable
        data <- data.frame(cbind(y,x)) #data frame with the binary outcome and the predictor
#Save each database
write.table(data, paste(paste(paste("DATAginibindiffvariance03_",j, sep=""),".",sep=""),i,sep=""), row.names=FALSE,
quote=FALSE)
}
}
```

```
################################
### Predictors with different precision


for(j in 1:10){ #loop for each association study
        for(i in 1:n){ #500 databases for each study
        #generate 10 independent standard normal variables
        xo<-rmvnorm(1000,mean=media,sigma=sigma)
        x<-cbind()#NUll matrix to fill it with xo predictors, each one with a different precision
        for(k in 1:10){ #loop for each predictor
                x1<-c(round(xo[,k],k)) #round predictor k with k number of decimal places
                x<-cbind(x,x1)
        }
        #names of the predictors
        colnames(x)<-c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10")
        #standard normal error
        e = rnorm(1000,0,1)
        #yo genarated by a linear model, Xj is the associated variable
        yo=0.3*x[,j]+e
        #transform the continuous outcome into a binary outcome
        y<-ra2ba(yo)
        #dataframe with the 10 predictors and the outcome
        data <- data.frame(cbind(y,x))
#Save each database in each associaton study
write.table(data, paste(paste(paste("DATAginibinstdnormalcutpoint03_",j, sep=""),".",sep=""),i,sep=""),
row.names=FALSE, quote=FALSE)
}
}
```

Code B.2. Generation of the databases with binary outcome under H_A, 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.

```
#import library for multivariate normal distribution
library(mvtnorm)
#500 databases
n=500

media=c(rep(0,10)) #set the mean to 0
sigma=diag(length(media)) # set the variance to 1
for(i in 1:n){ #for each database
        x<-rmvnorm(1000,mean=media,sigma=sigma) #generate 10 standard normal predictors
        e = rnorm(1000,0,1) #standard normal error
        y=e #outcome for the null, no associated predictor
        data <- data.frame(cbind(y,x)) #data frame with the continuous outcome and 10 predictors
#Save each database
write.table(data, paste("DATAgininullstdnormal.",i, sep=""), row.names=FALSE, quote=FALSE)
}
##############
### predictors with different variance
media=c(rep(0,10)) #set the mean to 0
sigma=diag(c(50,45,40,35,30,25,20,15,10,1)) #Now, different variance for each predictor
for(i in 1:n){ #Loop for each database
        #10 independent normal predictos with mean 0 and variance sigma (different variance)
        x<-rmvnorm(1000,mean=media,sigma=sigma)
        e = rnorm(1000,0,1) #standard normal error
        y=e #any predictor associated with the outcome
        data <- data.frame(cbind(y,x) #data frame with outcone and the 10 predictors
#save each database
write.table(data,paste("DATAgininulldiffvariance.",i, sep=""), row.names=FALSE, quote=FALSE)
}
############################
#### less error variance

media=c(rep(0,10)) #set the mean to 0
sigma=diag(length(media)) # the variance to 1
for(i in 1:n){ #Loop for each database
        #10 independent standard normal variables with 1000 observations
        x<-rmvnorm(1000,mean=media,sigma=sigma)
        e = rnorm(1000,0,0.5) #generate a normal error with mean 0 and standard deviation 0.5
        y=e #No association
        #data frame with the predictors and the outcome
        data <- data.frame(cbind(y,x))
#save each database
write.table(data, paste("DATAgininullstdnormalerror05.",i, sep=""), row.names=FALSE, quote=FALSE)
}
#################################
### predictors with different precision (cutpoints)

media=c(rep(0,10)) #set the mean to 0
sigma=diag(length(media)) #the variance to 1
for(i in 1:n){ #Create each databases
        xo<-rmvnorm(1000,mean=media,sigma=sigma) #10 independent standard normal predictors
        x<-cbind()#Save the variables with different precision in the matrix
        for(j in 1:10){ #For each variable
                x1<-c(round(xo[,j],j)) #the predictor xj is the xo(j) with "j" (number) decimal places
                x<-cbind(x,x1)
        }
        colnames(x)<-colnames(xo) #The name of the predictors is the same
        e = rnorm(1000,0,1) #standard normal error
        y=e #Null model, no association
        data <- data.frame(cbind(y,x)) #data frame with the outcome and the 10 predictors
#save each database
write.table(data, paste("DATAgininullstdnormalcutpoint.",i, sep=""), row.names=FALSE, quote=FALSE)
}
```

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.

```
library(mvtnorm) #import the library to generate the multivariate normal distribution
n=500 #Number of databases
media=c(rep(0,10)) #Set the mean of each variable to 0
sigma=diag(length(media)) #Matrix with the variance set to 1
for(j in 1:10){ #Loop for each association study
        for(i in 1:n){ #500 databases under each study
        #error and predictors follow a standard normal distribution
        x<-rmvnorm(1000,mean=media,sigma=sigma)
        e = rnorm(1000,0,1)
        x1<-x[,j] #Take the variable j that is going to be associated with the outcome
        y=0.3*x1+e #linear model with coefficient 0.3 for the variable j
        data <- data.frame(cbind(y,x)) #data with the continuous outcome and the 10 predictors
        write.table(data, paste(paste(paste("DATAginistdnormal03_",j, sep=""),"-",sep=""),i,sep=""),
row.names=FALSE, quote=FALSE) #save each database in each association study
}
}
##############################
#### predictors with different variance
media=c(rep(0,10)) #all predictors with mean 0
sigma=diag(c(50,45,40,35,30,25,20,15,10,1)) # but different variance
for(j in 1:10){ #For each associated study
        for(i in 1:n){ #create all 500 databases for each association study
        #10 normal distributed variables with mean 0, but different variance
        x<-rmvnorm(1000,mean=media,sigma=sigma)
        e = rnorm(1000,0,1) #standard normal error
        x1<-x[,j]
        y=0.3*x1+e #outcome modeled by the linear model with coefficient 0.3 for the variable j
        data <- data.frame(cbind(y,x)) #data frame with the outcome and predictors
        #save each database for each study
        write.table(data, paste(paste(paste("DATAginidiffvariance03_",j, sep=""),"-",sep=""),i,sep=""),
row.names=FALSE, quote=FALSE)
        }
}
#######################################
##### less error variance
media=c(rep(0,10)) ; sigma=diag(length(media)) #predictos with mean 0 and variance 1
for(j in 1:10){ # for each association study
        for(i in 1:n){ # for each database
        x<-rmvnorm(1000,mean=media,sigma=sigma) #10 independent standard normal predictors
        e = rnorm(1000,0,0.5) # generate normal error with mean 0 and standard deviation 0.5
        x1<-x[,j] #variable j, associated with the outcome
        y=0.3*x1+e #generating model
        data <- data.frame(cbind(y,x)) #data frame with the outcome and with the predictors
        write.table(data, paste(paste(paste("DATAginistdnormal03error05_",j, sep=""),"-",sep=""),i,sep=""),
row.names=FALSE, quote=FALSE) #save each database
        }
}
#####################################
#### Predictors with different precision
#Again the mean is 0 and variance 1 (set up before)
for(j in 1:10){ #For each assocation study f
        for(i in 1:n){ #for each database of 500
        xo<-rmvnorm(1000,mean=media,sigma=sigma) #standard normal distributed predictors
        x<-cbind()#Null matrix to fill with xo predictors, but each one with different precision
        for(k in 1:10){ #For each predictor
                x1<-c(round(xo[,k],k)) #the predictor k of xo has k decimal points
                x<-cbind(x,x1)
        }
        colnames(x)<-colnames(xo) #the name of the predictors are the same
        e = rnorm(1000,0,1) #normal distributed error
        y=0.3*x[,j]+e #linear generating model
        data <- data.frame(cbind(y,x)) #data with outcome and predictors
        write.table(data, paste(paste(paste("DATAginistdnormalcutpoint03_",j, sep=""),"-",sep=""),i,sep=""),
row.names=FALSE, quote=FALSE) #save each dataset for each study
        }
}
```

Code B.4. Generation of the databases with continuous outcome under H_A, 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.

```
library(lmtest) #import the library
n=500 #number of databases
results<-cbind()#NULL matrix where to save the results

for(j in 2:11){ #For each of the 10 values we extract the bias, the coverage and the p-values
        #null vectors for the bias, the upper and lower CI limits, and the p-values
        diff<-c();upper1<-c();lower1<-c();p<-c()
        for(i in 1:n){ #loop to calculate it across all 500 databases
        data<-read.table(paste("DATAginibinnullstdnormal.",i,sep=""),header=T) #input data
        y=data[,1] #the outcome is the first variable
        x1=data[,j] #variable under study
        reg1<-glm(y~x1,family=binomial(probit)) #logistic regression with probit link cinsudering the variable
        reg0<-glm(y~1,family=binomial(probit)) #regression with with probit link with only the intercept
        a1<-summary(reg1) #summary of the regression with the variable
        diff<-append(diff,a1$coefficients[2,1]-0,after=length(diff)) #extract the bias
        #extract the upper  and lower limit for the 95% CI
        upper1<-append(upper1,a1$coef[2,1]+1.96*sqrt(a1$coef[2,2]^2),after=length(upper1))
        lower1<-append(lower1,a1$coef[2,1]-1.96*sqrt(a1$coef[2,2]^2),after=length(lower1))
        lrtc<-lrtest(reg1,reg0) #perform LRT test for the models
        p<-append(p,lrtc$"Pr(>Chisq)"[2],after=length(p)) #extract the LRT p-values
        }
        sig<-length(p[p<0.05]) #save the number of times the variables were significant before Bonferroni
        sigb<-length(p[p<0.0001]) #save the number of times the variables were significant after Bonferroni
        insl1<-length(lower1[lower1>0]) #check the times the expected value is lower than the lower CI limit
        insu1<-length(upper1[upper1<0]) #check the times the expected value is higher than the upper CI limit
        cover1<-(insl1+insu1)*100/500 # average of the times the expected value is outside the CI in percentage
        bias<-mean(diff) #average of the bias for each database
        res<-c(bias,100-cover1,sig,sigb)  #vector of each output for each database
        results<-cbind(results,res)  #save for each value the outputs across all databases

}
results # print the results
```

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.

```
library(lmtest) # import library for the LRT tests
n=500 #Number of databases
results<-cbind()  #Null matrix to fill it with the bias, coverage and with p < Bonferroni threshol
for(j in 2:11){ #For associated study (the first variable of each database is the outcome)
        #Null vectors to save the bias, the upper and lower 95% limits of the CI
        diff<-c();upper1<-c();lower1<-c();p<-c()
        for(i in 1:n){ #For each database in each association study
        data<-read.table(paste(paste(paste("DATAginistdnormal03_",j, sep=""),".",sep=""),i,sep=""),header=T)
        y=data[,1] #outcome
        x1=data[,j] #variable associated
        reg1<-glm(y~x1)#regression model with the variable j
        reg0<-glm(y~1) #regression with only the intercept
        a1<-summary(reg1) #summary of the regression with the variable
        diff<-append(diff,a1$coefficients[2,1]-0.3,after=length(diff)) # Bias
        #upper limit and lower limit of the 95% CI
        upper1<-append(upper1,a1$coef[2,1]+1.96*sqrt(a1$coef[2,2]^2),after=length(upper1))
        lower1<-append(lower1,a1$coef[2,1]-1.96*sqrt(a1$coef[2,2]^2),after=length(lower1))
        lrtc<-lrtest(reg1,reg0) #LRT test between both regressions
        p<-append(p,lrtc$"Pr(>Chisq)"[2],after=length(p)) #LRT p-values
        }
        sig<-length(p[p<0.05]) #Number of p-values less than 0.05
        sigb<-length(p[p<0.0001]) #Number of p-values less than Bonferroni threshold
        insl1<-length(lower1[lower1>0.3]) #Number of lower values greater then the expected values
        insu1<-length(upper1[upper1<0.3]) #Number of upper values lower then the expected values
        cover1<-(insl1+insu1)*100/500 #sum both values dived by the total and take the percentage
        bias<-mean(diff)# mean of the bias
        #Save the bias, the coverage (inside of CI, not outside), the number of p-values less than both thresholds
        res<-c(bias,100-cover1,sig,sigb) #include them in the matrix of the results
        results<-cbind(results,res)
}
Results #Desplay the matrix of the results
```

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.

```
n=500 #Number Gini outputs

#matrix to save the Gini output for each variable, one when all variables are standard normal, another table for
#when variables have different precision (cutpoints)

gini1<-matrix(0,nrow=n,ncol=10)

gini2<-matrix(0,nrow=n,ncol=10)

for(i in 1:n){

#Gini output when all preditors are standard normal distributed under the null

t1<-read.table(paste(paste("outginibinnullstdnormal.",i,sep=""),".importance",sep=""), header=TRUE)

#Gini output when all predictor have different precision under the null

t2<-read.table(paste(paste("outginibinnullstdnormalcutpoint.",i,sep=""),".importance",sep=""), header=TRUE)

        a<-c("V2","V3","V4","V5","V6","V7","V8","V9","V10","V11") #names of the variables in t1

        b<-c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10") #names of the variables in t2

         g1<-numeric(10);g2<-numeric(10) #vectors with ten zeros to get the VIM for each variable in the dataset

        for(j in 1:10){

        o1f=t1$varname==a[j] #take the row of the variable in t1

        o2f=t2$varname==b[j] #take the row of the variable in t2

        g1[j]<-t1[o1f,4] #Take the VIMgini for that variable from t1

        g2[j]<-t2[o2f,4] #Take the VIMgini for that variable from t2

         }

        gini1[i,]<-g1 #save the VIMgini for all predictor of each VIMgini output from t1

        gini2[i,]<-g2 #save the VIMgini for all predictor of each VIMgini output from t1

}

#set the column of names to be the same

colnames(gini1)<-c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10")

colnames(gini2)<-c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10")

pdf("ginibinnullcomparecutpointH0.pdf") #Create a pdf with 2 plots to compare both cases

par(mfrow=c(2,1)) #2 plots in the figure, 2 in different rows

#boxplots with the VIMgini when the 10 variables follow a standard normal

boxplot(gini1,main="Under H0, y Binary, all predictors N(0,1), error~N(0,1)")

#boxplots with VIMgini when the predictors are normal distributed with different precision

boxplot(gini2,main="Under H0, y Binary, Different cutpoints, error~N(0,1)")

dev.off() #finish the pdf

medgini1<-apply(gini1,2,median) #median of VIMgini for all predictors from t1

medgini2<-apply(gini2,2,median) #median of VIMgini for all predictors from t2
```

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 H_0. 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).

```
n=500 #500 Gini outputs
gini<-list()#list to save the VIMgini for each association study
for(j in 1:10){ #loop for each association study
        gini1<-matrix(0,nrow=n,ncol=10) #matrix with 0 to fill with the VIM for all predictors from each output
        for(i in 1:n){ #Loop for each RF output
        #RF based on VIMgini output from each database when the variable j was associated
        t1<-
read.table(paste(paste(paste(paste("outginibinstdnormal03_",j,sep=""),"_",sep=""),i,sep=""),".importance",sep=""),
header=TRUE)
        a<-c("V2","V3","V4","V5","V6","V7","V8","V9","V10","V11") #names of the variables in the output
        g1<-numeric(10) #zero vector for save the VIM in each (i) output for all variables
        for(h in 1:10){ #For each variable
                o1f=t1$varname==a[h] #take the output for the variable h
                g1[h]<-t1[o1f,4] #save the VIMgini of the variable h in the vector
                }
        gini1[i,]<-g1 #save VIMgini for all variables from each output
        }
        gini[[j]]<-gini1 #save VIMgini for all variables and all 500 outputs
}
colnames(gini1)<-c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10") #names for the matrices on the list
pdf("ginibinstdnormalHAlim_1.pdf") #create the pdf for the plots
par(oma=c(0,0,2,0)) ; par(mfrow=c(3,2)) #outer margin for the title and first 6 plots in the pdf, 2 per row
for(j in 1:6){
        #boxplot with the VIMgini for each association study, limit the y axis with the same values, title and labels
boxplot(gini[[j]],main=paste("Variable associated X",j,sep=""),ylim=c(23,63),xlabel="Variable",ylim="Importance")
}
title(main="Under HA, y binary, all predictors follow N(0,1), error~N(0,1)",outer=T) #title
dev.off()#end the pdf
pdf("ginibinstdnormalHAlim_2.pdf") # pdf for the remain 4 plots
par(mfrow=c(3,2)) #6 plots in the pdf, only four places are going to be fill
for(j in 7:10){ # exactly the same as before
boxplot(gini[[j]],main=paste("Variable associated X",j,sep=""),ylim=c(23,63),xlabel="Variable",ylim="Importance")
}
dev.off()#end the pdf
medgini1<-apply(gini[[j]],2,median) #median of VIM gini for each predictor j. It was displayed for all predictors.
```

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 that 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, t1 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).