SciMed Central®

JSM Schizophrenia

Research Article

The use of Multiple SNP Data to Predict Schizophrenia Risk

Jules Hernández-Sánchez[1,2]*, Cameron Hurst[1,3], Dimitrios Vagenas[1], Christopher D Swagell[1], Ian P Hughes[1], Bruce R Lawford[1,4], Ross McD Young[1], Charles P Morris[1], and Joanne Voisey[1]

[1]Institute of Health and Biomedical Innovation, Queensland University of Technology, Australia
[2]Papworth Hospital (NHS UK), UK
[3]Faculty of Medicine, KhonKaen University, Thailand
[4]Division of Mental Health, Royal Brisbane and Women's Hospital, Australia

***Corresponding author**

Jules Hernández-Sánchez, Papworth Hospital, Papworth Everard, Cambridgeshire, UK, Tel/Fax: 004407843970264; Email: jules.hernandez.sanchez@gmail.com

Submitted: 16 February 2017

Accepted: 17 March 2017

Published: 20 March 2017

Keywords

- Schizophrenia
- SNP
- Machine learning

Abstract

Schizophrenia affects one percent of the population and has life-long debilitating consequences for those affected. It is a complex genetic disorder that results from the interaction of multiple gene variants and environmental factors. To date, numerous polymorphisms have been identified that are associated with schizophrenia but it is clear that no single polymorphism can accurately predict schizophrenia status.

Using a multiple candidate gene approach, 30 candidate genes were selected to test association with schizophrenia. Single nucleotide polymorphisms (SNPs) in these genes were selected because they were either Hap Map tag-SNPs or because of their location in a functional gene domain. Initially, a total of 273 SNPs were genotyped in 160 DSM-IV diagnosed schizophrenia patients and 250 control samples. After quality control, 151 SNPs in 29 different genes were used in this study. In order to evaluate the best method to combine multiple SNP data, five different statistical classifiers were used to predict schizophrenia risk. The five classifiers evaluated were; binary logistic regression (BLR), support vector machines (SVM), decision trees (DT), adaptive boosting (AB), and partial-least-squares with linear-discriminant-analysis (PLS-LDA).

The best classifier was BLR but it was more informative to use several classifiers. The synonymous SNP, rs7301328, in the glutamate receptor 2B gene (*GRIN2B*) was consistently selected among several classifiers (BLR, DT and AB). All classifiers utilised main effects of SNPs only but given that all genes were functional candidates for schizophrenia, we hypothesised they may interact. As BLR was the best classifier, we used it to estimate direct and interaction effects between all pair-wise combinations of SNPs. Additive-additive, additive-dominant and dominant-dominant interactions that averaged z-scores > 3.5 are reported. The greatest number of additive-additive interactions involved the catechol methyl Transferase (*COMT*) gene but both *COMT* and dopa decarboxylase (*DDC*) showed a large number of dominant-dominant interactions and *DDC* was over-represented in terms of additive-dominant interactions. While this panel of SNPs does not have the required sensitivity or specificity to be used as a diagnostic test, it is anticipated that the approach described in this study will lead to a test for improved early diagnosis of schizophrenia. Such a test will enable early intervention strategies with the ultimate objective of preventing schizophrenia onset and progression.

ABBREVIATIONS

BLR: Binary Logistic Regression; SVM: Support Vector Machines; DT: Decision Trees; AB: Adaptive Bootstrapping; PLS-LDA: Partial Least Squares and Linear Discriminant Analysis; Sens: Sensitivity; Spec: Specificity; F1 = 2 Sens PPV / (Sens + PPV); PPV: The Positive Predictive Value

INTRODUCTION

Schizophrenia is a devastating psychiatric disorder affecting about 1% of people worldwide and often results in self-harm or harming others and long periods of institutional care [1,2]. Symptoms of schizophrenia include hallucinations, delusions, thought disorder and general negative symptoms such as social withdrawal and self-neglect [2].

Schizophrenia is a complex trait involving the interaction between many genes and environmental factors, such as traumatic experiences during puberty or perinatal hypoxia [2]. Twin studies reveal the heritability of schizophrenia is up to 80% [3,4].

The high genetic risk for schizophrenia has led to considerable research efforts aimed at the identification of susceptibility genes. These studies have involved linkage analysis, candidate gene association studies and genome wide association studies (GWAS). As a result, numerous candidate genes have been identified that show consistent association with schizophrenia including; catechol methyl transferees (*COMT*), dystrobrevin binding protein 1 (dysbindin, *DTNBP1*), disrupted in schizophrenia 1 (*DISC1*), proline dehydrogenase 1 (*PRODH*), gamma-amino butyric acid A receptor 1 (*GABRA1*), regulator of G-protein signalling 4 (*RGS4*), neuregulin 1 (*NRG1*) and D-amino acid oxidase activator (*DAOA*).

This study aims to test a genetic model of schizophrenia prediction incorporating data from multiple candidate genes and SNPs. We compared the performance of five different classifiers for predicting schizophrenia: binary logistic regression (BLR), decision trees (DT), adaptive boosting (AB), partial least squares - linear discriminant analysis (PLS-LDA), and support vector machines (SVM). Their performance was measured in terms of sensitivity, specificity, positive predictive value, negative predictive value, F1-score, and overall misclassification error.

MATERIALS AND METHODS

Subjects

All subjects were recruited in the Brisbane region (a city of approximately 2 million inhabitants on the east coast of Australia). Subjects were all of mixed European descent, consisting of 157 unrelated cases (134 males and 23 females) and 237 unrelated controls (139 males and 98 females). The age range was between 18 and 65 years with a mean and sd of 36 ± 12 among cases and 37 ± 13 among controls.

Diagnoses were separately confirmed by two psychiatrists using the Diagnostic and Statistical Manual of Mental Disorders version IV (DSM-IV). Cases had never been diagnosed with other psychiatric disorders and all were maintained on a constant dose of antipsychotic medication for a minimum of three weeks excluding antidepressants, anxiolytic agents or mood-stabilizing psychotropic medications. The Positive and Negative Symptom Scale (PANSS) was administered to all cases to assess psychotic severity [5]. Cases were relatively severe and many had a family history of psychosis. They had been diagnosed on average for 13 years and were all still experiencing positive and negative symptoms despite antipsychotic medication.

Controls consisted of volunteers from the general public, hospital nursing and medical staff, and university staff and students. While formal screening for schizophrenia was not undertaken, this control sample represents a group of individuals who are unlikely to have cognitive deficits found in those with schizophrenia or their first-degree relatives [6].

Ethics approval was obtained from all institutions involved.

Genotypes

Thirty candidate genes were selected because either they had been previously associated with schizophrenia or they are part of the dopaminergic or glutamatergic pathways which have been hypothesized to be important in the pathogenesis of schizophrenia [7]. The genes were: *ANKK1, BDNF, HTR2A, GABBR1, G72, GRM3, NRG1, RGS4, RELN, DRD1, AR, DRD2, KPNA3, DRD3, AKT1, DRD4, PRODH, MAOB, CAPON, DAT, CNR1, GRIN1, DDC, COMT, GABRA1, DISC1, DTNBP1, GRIN2B, GRIN2C* and *GRIN2D*. A total of 162 SNPs were genotyped across all genes. Whenever possible, they were chosen from functional domains, promoters, 3´-untranslated regions or when they ‘tagged’ haplotype blocks as defined in HapMap phase II [8]. Genotyping was performed essentially as previously described, using a homogeneous Mass EXTEND (hME) Sequenom assay performed by the Australian Genome Research Facility [9]. Six SNPs with a minor allele frequency (MAF) of <1% were removed. Five SNPs that were not in Hardy-Weinberg equilibrium (HWE) were also removed as they may indicate genotyping errors [10]. Thus, a final total of 151 SNPs in 29 different genes were subjected to further analysis.

Missing data

The mean number of missing genotypes per SNP was 19 and the distribution was right skewed (skewness = 1.4; supplementary material). The number of missing genotypes per individual ranged from 1 to 13%. Only 112 individuals (28%) had complete genotype s. In order to prevent losing 72% of all data, missing data were randomly imputed 10 times given observed allele frequencies and assuming HWE and linkage equilibrium (LE). Those assumptions were adequate given that SNPs not in HWE had been discarded and that, apart from a few adjacent SNPs, pair wise linkage disequilibrium was practically nil (Figures in supplementary material).

Genetic models

Two nested genetic models were assumed in terms of SNP effects, additive and genotypic. In the additive model, a straight regression line was fitted through the three genotypes at each SNP. In the genotypic model, the means of all genotypes were estimated at each SNP. In this model, additive and dominant effects are combined. A third genetic model, in which additive and dominant effects were disentangled, was also implemented to build genetic networks.

Genetic networks

None of the classifiers used SNP interactions as predictive features in the main analysis of this study. However, in a secondary analysis, BLR was used to test additive- additive, additive-dominant and dominant-dominant interactions across all pairs of SNPs. Results were averaged over the 10 augmented data sets. Those interactions (and main effects) with a standardised absolute effect ≥ 3.5 were retained for visual inspection. This value was chosen in order to produce visually appealing plots, there was no biological or statistical basis for it.

Machine learning classifiers

In machine learning, a classifier is a statistical model with parameters estimated on training data and predictions made on cross-validating data. The training data consisted of 2/3 of all individuals chosen at random and the cross-validating data consisted of the remaining 1/3. Learning means choosing the parameters that minimise prediction error among the training set. Performance was reported on the validating set. All classifiers used the same SNPs, training and cross-validating data sets.

The task was to use SNPs to classify cases and controls. Five different classifiers were compared: PLS-LDA, BLR, SVM, DT, and AB. A summary of the main features of each classifier is given in the supplementary materials. Performance was measured as total misclassification error (error), sensitivity (sens), specificity (spec) and F1 score. The F1 score is the harmonic mean of sens and positive predicted value (ppv), i.e. F1 = 2 sensppv / (sens + ppv), and therefore combines two positive performance features into a single value simplifying comparisons across classifiers.

RESULTS AND DISCUSSION

Genotyping

Using a systematic approach from the literature and the

HapMap project, 273 SNPs were chosen from 30 schizophrenia candidate genes. In total, 160 DSM-IV diagnosed schizophrenia patients and 250 control samples were genotyped for all 273 SNPs. After quality controls (see Methods), 151 SNPs in 29 different genes were used for further analysis.

Performance of classifiers

In order to evaluate the best method to combine multiple SNP data, five different machine learning classifiers were trialled. The five classifiers evaluated were; binary logistic regression (BLR), support vector machines (SVM), decision trees (DT), adaptive boosting (AB) and partial-least-squares with linear-Discriminant-analysis (PLS-LDA). Cross-validated performance of classifiers is shown in Table (1). The best classifier across all performance measures was BLR, e.g. under the genotypic gene action model, the cross-validation misclassification error of BLR was 23%, whereas the cross-validation misclassification error varied from 30 to 40% across all the other classifiers. The F1 score of BLR was 68% under the genotypic gene action model, whereas it was between 42 and 56% across all the other classifiers.

Under the additive gene action model all classifiers performed worse than under the genotypic model. Nevertheless, BLR was still the best classifier with a cross-validation error of 29% compared to 36 to 40% across all the other classifiers and an F1 score of 60% compared to between 36 and 46% across the other classifiers. The difference in performance for the other measures (SVM, PLS-LDA, DT and AB) were relatively minor (Table 1).

Noteworthy genes and SNPs

Table (2), shows the common SNPs selected by the BLR, DT and AB classifiers across all 10 augmented data sets. Only one SNP, rs7301328 from the glutamate receptor 2B gene (*GRIN2B*), was consistently identified by each classifier (Table 2). This SNP appears to increase the risk of schizophrenia mainly through a dominant gene action (Table 3). The rs7301328 SNP showed a strong association with schizophrenia at the genotype level (χ^2= 15.3, df = 2, p-value = 0.0005), but the additive linear trend was not significant (χ^2= 2.2, df = 1, p-value = 0.135). This implies a strong dominant effect. Table (3) shows that there are proportionally more heterozygotes among cases than among controls, and that potentially being homozygous GG confers some protection against developing schizophrenia.

Common performance features among classifiers

There were two common performance features across classifiers. First, specificity was higher than sensitivity across all classifiers, averaging 81% and 43% respectively, regardless of which gene action model was used. This implies that controls were predicted with less error than cases. A plausible explanation is that cases may be heterogeneous, with several subtypes of schizophrenia grouped together as a single disease. Second, apart from DT, all classifiers performed better under the genotypic rather than the additive gene action model. For example, the mean F1 score across all classifiers under the additive gene action model was 46% compared to 55% under the genotypic gene action model. Moreover, the average error rates across all classifiers were 36% and 31% for the additive and genotypic models, respectively. This was probably due to the fact that the genotypic model estimated up to 3 parameters per SNP (3 means) whereas the additive model estimated only 2 (intercept and slope), and the additional parameter was necessary to capture the non-negligible dominant genetic variation present in schizophrenia.

Differential performance among tree-based classifiers (DT and AB)

Both AB and DT are classifiers based on generating random trees. However, only AB was sensitive to the assumed gene action model with, for example, F1 being ~10% higher when assuming a genotypic model over an additive one. In contrast, DT showed no change in F1 or any other measure of performance with regards to the gene action model (Table 1). Another difference is that AB required 100 random trees and DT required 10,000 trees to reach similar F1 scores.

Table 1: Performance of 5 classifiers in discriminating between schizophrenic patients and controls. The best classifier within each performing measure was shown in bold. (a) Additive genetic model (SNPs as covariates). (b) Genotypic genetic model (SNPs as factors). the R package *plsgenomics* (PLS-LDA) does not accept SNPs as factors.

	Additive model					**Genotypic model**			
	BLR	**SVM**	**DT**	**AB**	**PLS-LDA**	**BLR**	**SVM**	**DT**	**AB**
Error	29±2.8	38±2.2	40±1.8	36±4.2	39±1.7	23±2.2	30±2.2	40±1.8	32±3.2
Sens	53±3.7	27±3.1	37±3.3	38±4.5	36±4.1	64±3.8	49±2.8	37±3.8	45±4.6
Spec	84±2.6	85±2.3	75±2.9	81±5.2	78±5.0	85±3.1	84±4.2	75±2.5	83±4.3
F1	60±4.0	36±3.8	43±2.8	46±5.2	43±2.1	68±2.9	56±2.3	42±3.2	53±4.4

Abbreviations: Percentages Rounded to the Nearest Integer; ± Standard Deviations Across Augmented Data Sets BLR: Binary Logistic Regression; SVM: Support Vector Machines; DT: Decision Trees; AB: Adaptive Bootstrapping; PLS-LDA: Partial Least Squares And Linear Discriminant Analysis; Error: Misclassification Error; Sens: Sensitivity; Spec: Specificity; F1 = 2 Sens PPV / (Sens + PPV) Where PPV Is The Positive Predictive Value. There Were 10,000 Random Trees Generated In DT, and 100 in AB

Table 2: SNP selection across GLM, DT and AB classifiers.

Classifier	Selected SNPs
Genotypic GLM	rs165774, rs1923730, rs410557, rs7301328
Additive GLM	rs165774, rs3924999, rs4975646
DT$_{rca}$	rs7301328
DT$_{gini}$	rs9370822, rs40184, rs7301328, rs936460, rs6313
AB	rs7301328

Abbreviations: DT$_{gini}$: Feature selection for DT using the Gini index 2π(1-π); where π is the proportion of individuals of class 1 at a particular node. DT$_{rca}$: Feature selection for DT using out-of-bag permutation to calculate average classification error rate

Table 3: Genotypic distribution at rs7301328 before imputing missing data.

Genotypes	CC	CG	GG
Controls	41 (18.2%)	88 (39.1%)	96 (42.7%)
Cases	21 (13.5%)	92 (59.4%)	42 (27.1%)

Abbreviations: Pearson's χ^2 = 15.3, df = 2, p-value = 0.0005

Support vector machines

SVMs were used to extract additional information about this classification problem. A learning curve of training and cross-validating errors against sample size suggested that increasing sample size was more likely to reduce classification error than increasing model complexity (supplementary material). Moreover, the number of support vectors was high. Support vectors denote individuals used to define the classification boundary (or margin) between cases and controls. A large number of support vectors are indicative of a highly complex classification boundary. SVMs use kernels in the hope of finding simpler boundaries resulting in a model less likely to overfit the sample in hand. However, averaging across augmented data sets, out of 394 total individuals, there were 260 support vectors, 122 of which defined the actual margin and 137 of which lay within the margin. The latter set of support vectors were non-separable and contained all the misclassifications. Approximately 70% of all cases and 64% of all controls became support vectors.

Gene networks

All classifiers utilised only the main effects of SNPs. However, given that all genes were functional candidates for schizophrenia, we hypothesised that genetic interactions existed. Given that BLR was the best classifier and its model parameters have a clear genetic interpretation as additive or dominant gene effects, BLR was used to estimate direct SNP effects and interaction effects between all pairs of SNPs. The greatest number of additive-additive interactions involved the catechol methyltransferasegene (*COMT*) but both *COMT* and dopa decarboxylase (*DDC*) showed a large number of dominant-dominant interactions. Finally, *DDC* was over-represented in terms of additive-dominant interactions. Figure (1), shows a circular plot of additive-additive interactions, and plots for additive-dominant and dominant-dominant interactions are given in the supplementary material.

CONCLUSION

The main outcomes

In this study we compared five different methods to classify individuals either as cases or controls for schizophrenia based on 151 SNPs in 29 candidate genes. BLR was the best classifier among those tested, and all the other classifiers were roughly equivalent in performance. The most consistently featured SNP was located in the glutamate receptor B gene (*GRIN2B*), which is part of the glutamatergic pathway and is consistently reported to be involved in schizophrenia pathogenesis. Estimating both additive and dominant genetic effects rather than just additive genetic effects rendered generally better classifications, with the exception of DT. Specificity was greater than sensitivity across all classifiers, meaning that it was easier to predict controls than cases. Genetic networks were built with BLR supporting the hypothesis of existing pair-wise genetic interactions between genes.

Despite all the above outcomes, none of the classifiers could predict schizophrenia reliably enough to be used diagnostically, mainly because the average sensitivity was too low at 43%. However, the performance of this approach is likely to be improved by future studies that include a larger number of cases and controls, the analysis of more SNPs in other schizophrenia candidate genes and the selection of more clinically uniform cases that represent specific molecular subtypes of schizophrenia. However, it will not be clear how generalisable these results are until this approach has been applied to additional independent patient cohorts from diverse ethnogeographic origins.

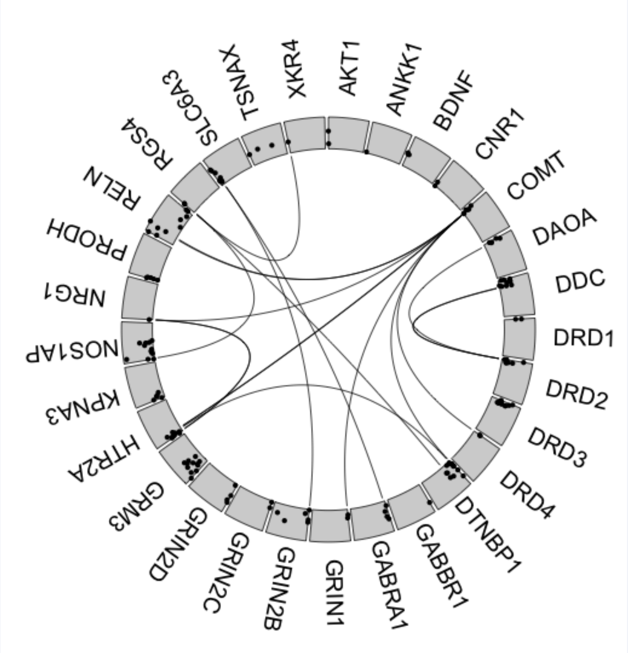

Figure 1 Additive-by-additive interactions (lines connecting two genes) averaging z>3.5 (standardised effect) across all 10 augmented data sets. Each gene is represented by equally wide segments. The relative location of each SNP, to the first SNP in a gene, is represented by dots within segments.

Performance of classifiers

The observation that BLR classified better than SVM, PLS-LDA, DT and AB may be partly due to the more elaborated feature selection scheme implemented in BLR compared to the other classifiers (see Supplementary Material). Additionally, the other classifiers have some disadvantages compared to BLR, e.g. penalized parameter estimates in SVM suffer from bias and are less efficient (have greater standard errors) than BLR estimates, PLS-LDA may be discarding important information by using only the first principal component, and DT and AB may be too naïve to solve complex classification problems.

Although interactions between SNPs were ignored when classifying, BLR offered the possibility of modelling additive and dominant interactions. The model parameters of BLR have a direct biological interpretation absent in all other classifiers. This is perhaps the strongest advantage of BLR. Although the genetic networks have not been validated, the possibility of incorporating them into a classifier is a potential avenue for future research.

There were two noteworthy results obtained with SVM, i.e. learning curves and performance of SVM without kernel. Learning curves for SVM suggested that the model was sufficiently complex and that more data could be collected to reduce classification error. Furthermore, SVM without a kernel rendered simpler boundaries. They required less support vectors compared to a SVM with a Gaussian kernel, at the cost of greater misclassification error (with kernel: 37% misclassification, 49% sensitivity and 72% specificity; without kernel: 23%, 64% and 85%). While we evaluated five classifiers in this study there are many others that could have been tested though they all have their own strengths and weaknesses.

As the only differences between data sets were missing data imputations, the standard deviation of classifier performance denoted the sensitivity of each classifier to recovering missing genotypes. All classifiers seem to respond similarly to missing data imputations. Imputations were important in this study to prevent loosing approximately 70% of all data.

Causal relations between GRIND2B and schizophrenia

The most consistent signal detected in this study was from a *GRIN2B* SNP. The glutamatergic system has long been regarded as having a central role in the pathophysiology of schizophrenia [11]. N-methyl-D-aspartate (NMDA) binds to specific glutamate receptors and agonists of NMDA such as glycine, D-cycloserine and D-serine have been successfully used as adjunctive medication to improve both negative and cognitive symptoms of schizophrenia [12]. Moreover, gamma-aminobutyric acid (GABA) is produced from glutamate by glutamic acid decarboxylase in presynaptic neurons before being released in the synaptic cleft. Significantly, mRNA expression profiling of prefrontal cortex from matched pairs of schizophrenia and control subjects revealed that transcripts encoding proteins involved in the regulation of those presynaptic functions were decreased in all cases but not in controls [13].

NMDA glutamatergic receptors reduce the phosphorylation and activity of protein phosphatase 1 regulatory subunit 1B (PPP1R1B). The active form of PPP1R1B inhibits phosphatase 1, a ubiquitous regulator of receptors and ion channels in neurons. Reduced PPP1R1B expression among schizophrenic patients reduces dopaminergic function, which opposes the action of glutamate receptors on PPP1R1B. The inhibitory effects of the NMDA receptors on PPP1R1B may be related to important alterations in the dorsolateral prefrontal cortex such as reduction of thalamocortical afferents, dendritic spines, inhibitory interneurons, neuronal size and mesocortical afferents [14].

The difficulty of predicting schizophrenia from SNP data

Misclassification rates ranged from 23 to 39% in this study, being lower when assuming a genotypic rather than additive gene action. Potential causes of misclassification were the limited number of participants and limited candidate genes included in the study. Other factors that may have contributed to misclassification were that DSM-IV and demographic data were only collected among cases, hence only SNPs could be used as predictors, and the probable existence of multiple genetic subtypes within schizophrenia. It may be possible to improve the model by including other predictor variables such as family history of mental disorders and demographic data. Moreover, multiple genetic pathways are likely to lead to schizophrenia, in addition to the existence of complex gene-environment interactions. However, one of the strengths of these data was the use of cases with a relatively uniform severe presentation that was rigorously diagnosed.

Candidate genes for schizophrenia that were not included in this study have been reported. Allen et al., carried out 118 meta-analyses using over 1,000 genetic studies, and reported nominally significant effects in 24 variants mapping to 16 different genes with an average odds ratio of 1.23 [15]. While many of these genes were included in the present study, notable exceptions included *MTHFR* and *TPH1*. A GWAS consortium reported 7 SNPs significantly associated with schizophrenia, the strongest of which was a known regulator of neural development micro RNA (MIR137). Four other genome-wide significant loci contained targets of MIR137 (The International Schizophrenia Consortium, 2008)[15].

Predicting genetic susceptibility to schizophrenia is difficult because it is a genetically complex trait. The International Schizophrenia Consortium reported that rare chromosomal deletions and duplications increased risk of schizophrenia and Purcell et al. reported a polygenic burden of rare and disruptive polymorphisms that increased the risk of schizophrenia [16,17]. Another difficulty is that 100 years after the first reported diagnosis of schizophrenia, there is still no consensus about whether it is a single disease or a disease with subtypes. This is exemplified by the fact that the number and nature of the subtypes varies across different psychological instruments. Indeed, the DSM-V treats schizophrenia as a binary trait (present/absence) but DSM-IV identified 5 subtypes (paranoid, disorganized, catatonic, undifferentiated and residual) and the International Statistical Classification of Diseases and Related Health Problems (ICD-10) from the World Health Organization describes two additional subtypes (post-schizophrenic depression and simple schizophrenia).

Although current data sets contain millions of SNPs and thousands of cases and controls, it is yet unclear whether final predictive models must contain all that, mostly irrelevant, genetic information or just a subset with informative mutations. From a practical point of view, e.g. implementation in primary care, the latter is easier to implement, and we recommend BLR. As technology advances and the use of all genomic information (where n<p) becomes routine then machine learning models could provide better alternatives to BLR.

Moreover, interactions are mostly neglected in research because of the complexity of exploring exponentially large number of effects. Nevertheless, interactions must hold some of the elusive genetic effects yet to explain part of the genetic variation in complex traits. BLR can provide estimates of interaction effects more easily than machine learning methods.

Finally, the presence of dominance seems pervasive. Therefore, genotypic models (sensitive to both dominant and additive gene effects) rather than additive models (which ignore dominance) should be applied.

CONCLUSION

This study reports a panel of SNPs targeting candidate genes that could help in predicting risk of schizophrenia before its onset, thereby raising the possibility of preventative interventions based on individual patient pharmacogenetic profiles. We found that BLR was the best performing of the five classifiers compared in this study. The misclassification error was 23% and the F1-score 68%. It may be possible to improve the model by including other predictor variables such as family history of mental disorders and demographic data. Additional genetic data could also be used such as data emerges from SNP association, mRNA expression and methylation studies.

The novel findings in this work were: 1) in a n>p situation, e.g. candidate genes analysis and moderate sample size, the classical BLR is better than machine learning procedures in terms of specificity, sensitivity and classification error, 2) that BLR can easily be used to test interactions, and 3) nevertheless a battery of models can help identifying the strongest candidate mutations.

LIMITATIONS

Specifically, a recent publication [18] has shown that schizophrenia may have many (more than 30) risk-contributing common variants of small effect size. Polygenic risk scores have been proposed for many disorders, including schizophrenia [19,20]. Based on the results of all of these papers, it is unsurprising that a definite classifier could not be constructed with our data.

Rather than attempting to build a classifier from these data, it may be more fruitful to examine the models most appropriate for determining the genetic risk conferred by these SNPs. In most current association studies, an additive model is assumed. However, we have uncovered evidence for the existence of dominant effects as well as of interactions. Comparing predictive performance of complex models (including dominant effects and interactions) against simpler ones (additive effects) is required.

ACKNOWLEDGEMENTS

This work was financially supported by the Queensland State Government, the Queensland University of Technology and the Nicol Foundation. This study has been reviewed by the appropriate ethics committee and all persons gave their informed consent prior to their inclusion in the study.

REFERENCES

1. Karayiorgou M, Gogos JA. A turning point in schizophrenia genetics. Neuron. 1997; 19 : 967-979.

2. Picchioni MM, Murray RM. Schizophrenia. BMJ. 2007; 14: 91-95.

3. Herson M, John Wiley Sons, Hoboken. Adult psychopathology and diagnosis. 2011.

4. Cardno AG, Marshall EJ, Coid B, Macdonald AM, Ribchester TR, Davies NJ, et al. Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. Arch Gen Psychiatry. 1999; 56: 162-168.

5. Peralta V, Cuesta MJ. Psychometric properties of the positive and negative syndrome scale (PANSS) in schizophrenia. Psychiatry Res. 1994; 53: 31-40.

6. Niendam TA, Bearden CE, Rosso IM, Sánchez LE, Hadley T, Nuechterlein K.H, Canon TD. A Prospective Study of Childhood Neurocognitive Functioning in Schizophrenic Patients and Their Siblings. American J Psychiatry. 2003; 160: 2060-2062.

7. Brisch R, Saniotis A, Wolf R, Bielau H, Bernstein HG, Steiner J, et al. The role of dopamine in schizophrenia from a neurobiological and evolutionary perspective: old fashioned, but still in vogue. Front Psychiatry. 2014; 1: 5-47.

8. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449: 851-861.

9. Voisey J, Swagell CD, Hughes IP, Lawford BR, Young RM, Morris CP. Analysis of HapMap tag-SNPs in dysbindin (DTNBP1) reveals evidence of consistent association with schizophrenia. Eur Psychiatry. 2010; 25: 314-319.

10. Xu J, Turner A, Little J, Bleecker ER, Meyers DA. Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? Human Genetics. 2002; 111: 573-574.

11. Krystal JH, Karper LP, Seibyl JP, Freeman KG, Delaney R, Bremner JD,et al. Subanesthetic effects of the noncompetitive NMDA antagonist, Ketamine, in humans: phycotomimetic, perceptual, cognitive and neuroendocrine responses. Arch Gen Psychiatry. 1994; 51: 199-214.

12. Potvin S, Stip E, Roy JY. Trends in Schizophrenia Research. Nova Science Publishers Inc New York. 2005; 119-149.

13. Mirnics K, Middleton FA, Marquez A, Lewis DA, Levitt P. Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. Neuron. 2000; 28: 53-67.

14. Masciotra L, Landreau F, Conesa HA, de Erausquin GA. Trends in Schizophrenia Research. Nova Science Publishers Inc New York. 2005; 27-44.

15. Allen NC, Bagade S, McQueen MB, Ioannidis JP, Kavvoura FK, Khoury MJ, Tanzi RE, Bertram L. Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. Nature Genetics. 2008; 40: 827-834.

16. The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature Letters. 2008; 455: 237-241.

17. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014; 506: 185-190.

18. Ripke S. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511: 421-427.

19. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460: 748-752.

20. Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genetics. 2013; 21: 2013.

Cite this article

Hernández-Sánchez J, Hurst C, Vagenas D, Swagell CD, Hughes IP, et al. (2017) The use of Multiple SNP Data to Predict Schizophrenia Risk. JSM Schizophr 2(1): 1007.