# Multivariate Pattern Analysis of Genotype–Phenotype Relationships in Schizophrenia

Amanda B. Zheutlin[1,8], Adam M. Chekroud[1,2,3,8], Renato Polimanti[3], Joel Gelernter[3], Fred W. Sabb[4], Robert M. Bilder[5], Nelson Freimer[6], Edythe D. London[6], Christina M. Hultman[7], and Tyrone D. Cannon[*,1,3]

[1]Department of Psychology, Yale University, New Haven, CT; [2] Spring Health, New York, NY; [3]Department of Psychiatry, Yale University School of Medicine, New Haven, CT; [4]Lewis Center for Neuroimaging, University of Oregon, Eugene, OR; [5]Department of Psychology, University of California - Los Angeles, Los Angeles, CA; [6]Department of Psychiatry and Biobehavioral Sciences, University of California - Los Angeles, Los Angeles, CA; [7]Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[8]These authors contributed equally to the article.

*To whom correspondence should be addressed; Department of Psychology, Yale University, PO Box 208205, New Haven, CT 06520; tel: 203-436-1545, e-mail: tyrone.cannon@yale.edu

Genetic risk variants for schizophrenia have been linked to many related clinical and biological phenotypes with the hopes of delineating how individual variation across thousands of variants corresponds to the clinical and etiologic heterogeneity within schizophrenia. This has primarily been done using risk score profiling, which aggregates effects across all variants into a single predictor. While effective, this method lacks flexibility in certain domains: risk scores cannot capture nonlinear effects and do not employ any variable selection. We used random forest, an algorithm with this flexibility designed to maximize predictive power, to predict 6 cognitive endophenotypes in a combined sample of psychiatric patients and controls ($N = 739$) using 77 genetic variants strongly associated with schizophrenia. Tenfold cross-validation was applied to the discovery sample and models were externally validated in an independent sample of similar ancestry ($N = 336$). Linear approaches, including linear regression and task-specific polygenic risk scores, were employed for comparison. Random forest models for processing speed ($P = .019$) and visual memory ($P = .036$) and risk scores developed for verbal ($P = .042$) and working memory ($P = .037$) successfully generalized to an independent sample with similar predictive strength and error. As such, we suggest that both methods may be useful for mapping a limited set of predetermined, disease-associated SNPs to related phenotypes. Incorporating random forest and other more flexible algorithms into genotype–phenotype mapping inquiries could contribute to parsing heterogeneity within schizophrenia; such algorithms can perform as well as standard methods and can capture a more comprehensive set of potential relationships.

*Keywords:* cognition/machine learning/endophenotype/ genetics/computational psychiatry/polygenic risk scores

## Introduction

Substantial progress has been made toward identifying the genetic loci conferring risk for schizophrenia.[1,2] Known risk variants have been leveraged extensively to predict related phenotypes such as cognitive impairment,[3–5] symptom severity,[5–7] structural brain abnormalities,[8,9] as well as other psychiatric illnesses,[10,11] yielding insight into the shared (and discrepant) etiologies of these clinical and biological phenotypes. This suggests that individual variation in genetic risk for schizophrenia may be informative in parsing the phenotypic heterogeneity within schizophrenia—both in disease development and clinical presentation.[12,13] Within psychiatry, genotype–phenotyping mapping usually employs polygenic risk scores—aggregate measures of disease-associated single nucleotide polymorphisms (SNPs). While this approach has many benefits, it is less flexible than other statistical methods used commonly in prediction.

Several recent studies have promoted the use of random forest and other machine learning algorithms to improve genomic mapping because they can detect complex, nonlinear, high-dimensional patterns of effects that may inform predictions.[14–17] Interactions of this kind are widely believed to play a role in the etiology of schizophrenia[18–20] and contribute variance above and beyond additive genetic effects.[21] Additionally, random forest selects only those variants among a list that boost predictive power, providing a data-driven approach to identifying subsets of SNPs from among those known to be disease-associated that are important for individual related phenotypes. These 2 features of random forest— nonlinearity and feature selection—allow for modeling

genotype–phenotype interactions more flexibly. Such flexibility has resulted recently in discovery of novel disease-related loci[16] and increased predictive strength of genetic variants on case–control status in psychiatry[14] and medicine more broadly.[22] Here, we are interested in the application of this algorithm to mapping disease-associated SNPs to intermediate phenotypes associated with schizophrenia. For reference, we compare it to linear, additive models, including polygenic risk scores given their popularity in the field.

In the current study, we used random forest, general linear regression, and polygenic risk scores to test the effects of 77 loci previously associated with schizophrenia[23] on a range of cognitive domains known to be impaired in schizophrenia. Statistical models were trained to predict measures including verbal intelligence, working memory, and processing speed within a sample of European-ancestry psychiatric patients and controls ($N = 739$). However, the utility of any model depends critically on its generalizability to data beyond the original discovery context,[24,25] which cannot be fully examined in one sample alone.[24] With this in mind, models were then tested in an ancestry-matched independent cohort ($N = 336$), and the predictive ability of random forest, linear regression, and polygenic risk scores were assessed. We aimed to compare how well each of these algorithms could map the same schizophrenia risk variants to 6 similar, but distinct, neuropsychological measures, as improving the precision of genotype–phenotype mapping may aid in understanding heterogeneity in etiology and clinical presentation in schizophrenia.

## Methods and Materials

### Participants

*American Sample.* Neuropsychiatric subjects and healthy controls were recruited as part of the Consortium for Neuropsychiatric Phenomics at the Semel Institute at the University of California, Los Angeles, a study examining underlying genetic and neural factors and their link to 3 target neuropsychiatric illnesses. Details of recruitment and study protocol are listed in supplementary information S1. In total, 1254 subjects were evaluated for this study; however, to minimize issues associated with population stratification, only individuals with self-reported European ancestry were included in the analyses ($N = 739$). Of these, 645 were healthy individuals, 24 were schizophrenia patients, 33 were bipolar patients, and 37 were attention-deficit, hyperactive disorder (ADHD) patients.

*Swedish Sample.* Same-sex twin pairs with at least one member diagnosed with schizophrenia or bipolar disorder when discharged from a hospital, and who were born in Sweden between 1940 and 1985, were identified on a population basis via the Swedish Twin Registry and Swedish National Patient Registry. Details of recruitment and study protocol are listed in supplementary information S1. The final sample included 55 schizophrenia patients, 58 unaffected schizophrenia co-twins, 62 bipolar patients, 45 unaffected bipolar co-twins, and 116 control twins ($N = 336$). Demographic information for all subjects is listed in table 1.

### Cognitive Performance

Participants in American and Swedish samples completed neuropsychological batteries that were partially overlapping. To examine the generalizability of all statistical models, we only analyzed target measures that were available in both samples, though in some cases the included measures were from different neurocognitive batteries or different versions of the same battery in the 2 samples. Six cognitive tasks in the American sample met this criterion: the California Verbal Learning Test (CVLT)[26]; the Vocabulary subtest of the Wechsler Adult Intelligence

**Table 1.** Demographic and Cognitive Performance Metrics

| | American Sample | Swedish Sample | Statistic | $P$ |
|---|---|---|---|---|
| N | 739 | 336 | — | — |
| Age (years) | 31.6 (8.6) | 49.5 (10.5) | 25.56 | <.001 |
| Sex (%F) | 50.6% | 52.7% | 0.32 | .573 |
| % Patients | 12.7% | 34.8% | 70.13 | <.001 |
| CVLT | 55.4 (9.6) | 49.8 (11.5) | 0.93 | .353 |
| Vocabulary | 43.5 (9.4) | 30.8 (6.1) | -15.07 | <.001 |
| Trails 1/A | 32.5 (11.9) | 34.6 (13.9) | -3.78 | <.001 |
| Digit Span | 30.7 (5.6) | 23.0 (5.6) | -10.63 | <.001 |
| VR I | 37.6 (5.1) | 31.8 (6.2) | -3.41 | <.001 |
| VR II | 31.2 (8.9) | 25.4 (9.0) | 0.99 | .324 |

*Note:* Sample means and standard deviations were reported for demographic and cognitive performance variables. Cognitive performance was adjusted for sample differences in scale ranges when necessary, but otherwise reflect raw scores. Chi-square tests were performed for sex and % patients. For all other variables, mixed effect regressions were run with family as a random effect. Age and diagnosis were included as covariates for cognitive models as samples differed by these metrics significantly. $t$ statistics for the fixed effect of sample and corresponding $P$-values were listed. CVLT, California Verbal Learning Test; VR I and VR II, Visual Reproduction I and II.

Scale—IV (WAIS-IV)[27]; the Visual Reproduction—Immediate Recall (VR I), Visual Reproduction—Delayed Recall (VR II), and Digit Span subtests of the Wechsler Memory Scale—IV (WMS-IV)[28]; and Color Trails 1.[29] In the Swedish sample, the Vocabulary subtest of the Wechsler Abbreviated Scale of Intelligence,[30] the VR I and II and Digit Span subtests of the WMS-R,[31] and Trail Making Test A[32] were used instead and adjusted according to scale differences relative to the versions used in the American sample. Respectively, these tasks assessed verbal learning and memory, visual memory, working memory, and processing speed. Cognitive variables were z-scored according to the American sample means and standard deviations, and due to extreme outliers in the Swedish sample ($|z| > 5$), they were also winsorized (trim = 0.01). Table 1 lists descriptive statistics for each measure in each sample. See supplementary information S2 for a complete list of tasks administered in each study.

Since predictive accuracy could reflect performance-related variance either specific to that task or attributable to general cognitive ability more broadly, we tested each model in the validation sample for its prediction of both (ie, task performance and general ability). We generated 6 general ability scores, each excluding one cognitive task (so that general ability in each case is independent from performance on the task of interest), by extracting the first principal component from the scaled scores of the included 5 tasks.

To account for age, sex, and ancestry (see Genotyping and Quality Control for details), linear regression models were run using these variables to predict cognitive performance. Residual values were extracted and used in all modeling and validation analyses in both cohorts.

### Genotyping and Quality Control

Blood samples were collected from individuals recruited for the American sample and genotyped using the Illumina OmniExpress (769 152 SNPs) SNP array. DNA extracted from blood samples from the Swedish sample was genotyped in 3 batches, the first two on the Illumina Omni 1M SNP array and the third on the Illumina Omni 2M SNP array. For all genotype data, markers were excluded if they had <95% genotyping rate (3325 variants excluded), a minor allele frequency (MAF) <1% (53 143 excluded), or deviated significantly from Hardy–Weinberg equilibrium expectations ($P < 10^{-4}$; 18 684 excluded). Individuals were excluded for missing genotypic data, missing phenotypic data, and unexpected relatedness as revealed by inflated pairwise identity-by-descent ($\hat{\pi} > .2$) in PLINK.[33] Data retained after these steps were imputed using SHAPEIT[34] for the prephasing, IMPUTE2[35] for imputation, and the 1000 Genomes Project Phase 1[36] as the reference panel; high quality imputed markers were retained (INFO score > .8, MAF > .01; 8 430 008 SNPs). To generate hard-call genotypes from the imputed

dosage data, we used maximum posterior probability ≥.8, MAF ≥.05, and SNP missingness rate <.01, yielding a final data set of 3 986 016 SNPs.

A multidimensional scaling matrix[33] was generated for each cohort using hard-call genotypes, the first 10 dimensions of which were used to control for ancestry. Genotypes available at the best SNP from each of the 108 loci identified in the Psychiatric Genomics Consortium (PGC) schizophrenia genome-wide association study (GWAS)[23] were extracted for all subjects. Minor allele dosage at 59 of these SNPs present in the imputed data sets and 18 additional proxy SNPs ($r^2 > .8$), identified by SNAP[37] using the CEU population panel from the 1000 Genomes Pilot 1 reference data set, were included in the analyses (77 SNPs total). No high-quality markers or surrogate markers were available for the remaining 31 schizophrenia-associated SNPs.

### Statistical Modeling

*Training and Validation Procedures.* Random forest and linear regression models were constructed and examined with repeated 10-fold cross-validation (3 repeats), which partitioned the American sample into 10 distinct subsets, used 9 of those subsets in the training process, and then made predictions on the remaining subset. To avoid opportune data splits, model performance metrics were averaged across the test folds and repeats. During training (within the US sample, $N = 739$), model performance was assessed by calculating a Pearson correlation between predicted and residual observed values.

For external validation, the model built in the American cohort was applied without modification to predict the same phenotypic measure in the Swedish cohort. To explore the specificity of these predictions, the same model was also used to predict general ability. To maintain power, we accounted for relatedness statistically rather than including only one individual from each twin pair. Specifically, model performance in the Swedish sample was assessed using linear mixed effects models, which included the fixed effect of model prediction on residual observed performance, and a random intercept for family to account for relatedness. One-tailed P-values were used for significance testing in this sample as correlations should be positive in all cases.

*Random Forest.* Machine learning methods identify patterns of information in data that are useful in predicting targets at the single-subject level.[24,38] Here, we implemented an algorithm based on a class of models called decision trees, which are a nonparametric approach for mapping observations about an individual (in this case, minor allele dosage at each SNP) to a target of interest. They naturally perform an implicit form of variable selection because irrelevant predictor variables will not be used to partition observations. Recently, ensemble decision tree

approaches have been developed that improve generalizability: here, we implemented the random forest.[38,39] To create a random forest, many bootstrap samples are drawn from the data and tree models are fit to each sample (a process known as bagging). Crucially, each tree model can only use a small random subset of the available predictor variables. Final predicted scores for each subject are determined by taking the average prediction of all tree models.

To select the most appropriate model parameters, we performed a grid search over a predetermined range of reasonable parameters, using an $R^2$-optimization process. We evaluated $\in \{3,4,10\}$ as the number of randomly selected predictors of the total 77 SNPs, as the global optimum is thought to be reached when the range overlaps the square root of the number of predictors.[40] We selected the least complex parameter combination that was within one standard error of the best performing combination to minimize risk of overfitting.[41] Parameter optimization was only conducted within the American sample: once the optimal model was determined during cross-validation, it was applied to the Swedish cohort without modification.

*Linear Regression Models.* A linear modeling approach including all 77 SNPs as predictors was used for each of the 6 cognitive tasks. Cross-validation and external validation procedures were conducted exactly as in random forest.

*Polygenic Risk Scores.* Polygenic risk scores typically sum across SNPs weighted by odds ratios from a case–control GWAS. One advantage of random forest and linear regression relative to polygenic risk scores is that these algorithms weight SNPs in training with respect to the desired outcome measure (cognitive performance), rather than a different phenotype (case–control status). To test this approach when tailored to the chosen outcome, we calculated task-specific risk scores in PLINK[33] for each cognitive measure. We summed the number of minor alleles for each of the available 77 SNPs weighted by the beta from the association of that SNP with residual cognitive performance in the American sample. Mixed effect models of each cognitive polygenic risk score on corresponding residual task performance were run in the Swedish sample, including family as a random variable.

Unless otherwise specified, all analyses were implemented in the freely available R statistical environment (Version 3.2.2; http://cran.r-project.org/). Models were built using the caret package[42] as a wrapper to the R implementation of the Random Forest algorithm.[43,44] All R code developed for statistical modeling is available upon request (Chekroud, Zheutlin).

## Results

### Model Development and External Validation

Cross-validated performance metrics (Pearson correlations) were modest and comparable for all random forest and linear regression models ($R^2$s 1.0%–1.6%; supplementary information S4) within the American sample. The random forest models were relatively complex, with 216–270 terminal nodes. After training, all models—random forest, linear regression, and polygenic risk scores—were applied without modification to predict corresponding performance measures in the Swedish sample (table 2). Two random forest models significantly predicted the corresponding cognitive measure, Trails 1/A and Visual Reproduction II (table 2; supplementary information S3) and permutation analyses revealed that these models were unlikely to have generalized by chance ($P$s ≤ .05; supplementary information S7). However, neither model significantly predicted general ability (all $P$s > .11), suggesting they generated relatively task-specific predictions. Two task-specific polygenic risk scores also predicted the corresponding measure, CVLT and Digit Span (table 2) and did not predict general ability significantly (all $P$s > .07). Linear regression models did not significantly predict any cognitive measures in the Swedish sample.

We also trained random forest and linear regression models in exactly the same way as described in the primary analyses using only control subjects ($N$ = 645), tested these for external validation, and found similar effect sizes and errors (supplementary information S5) in both samples, suggesting these genotype–phenotype relationships do not reflect unrelated case–control differences. One final concern was regarding multiple comparisons as we built models for 6 independent phenotypes. While we believe that external validation and permutation testing should address concerns of false positives, we encourage readers to consider that these results do not survive correction for multiple comparisons and so they should be seen as preliminary.

### Variable Importance

For the 2 random forest models that generalized to the Swedish sample, we inspected which SNPs were selected. Random forest models perform a form of implicit feature selection because a variable is only used to partition a sample if an informative split is found. For both the Trails 1/A and VR II models, nearly all (76 of the 77) SNPs were used. We also estimated relative variable importance using a permutation based test. This method rests on the logic that if a variable is not important (the null hypothesis), then rearranging the values of that variable will not degrade prediction accuracy. For each tree, the mean squared error is computed on the out-of-bag data, and then the same is computed after permuting a variable. The differences are averaged and normalized by the standard error. If the standard error is equal to 0 for a variable, the division is not done. The relative importance of all variables for each model that successfully generalized was illustrated in figure 1 along with the rankings of

**Table 2.** Predictive Ability of Random Forest, Linear Regression, and Polygenic Risk Scores for Cognitive Performance

| | Random Forest | | | General Linear Model | | | Polygenic Risk Scores | | |
|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMSE | $P$ | $R^2$ | RMSE | $P$ | $R^2$ | RMSE | $P$ |
| CVLT | 0 | 1.91 | .475 | .004 | 1.90 | .135 | **.010** | **1.75** | **.042** |
| Vocabulary | .002 | 1.17 | .222 | 0 | 1.17 | .488 | .002 | 1.19 | .200 |
| Trails A | **.013** | **1.70** | **.019** | .003 | 1.71 | .179 | .005 | 1.72 | .108 |
| Digit Span | 0 | 1.10 | .375 | .005 | 1.11 | .103 | **.009** | **1.17** | **.037** |
| VR I | .001 | 1.20 | .258 | .004 | 1.20 | .135 | 0 | 1.22 | .405 |
| VR II | **.010** | **1.40** | **.036** | .001 | 1.41 | .291 | 0 | 1.46 | .431 |

*Note:* All models were developed in the American sample ($N = 739$) and applied without modification to an independent Swedish sample ($N = 336$). Residual cognitive performance excluding variance attributable to age, sex, and ancestry was used in all cases. We reported mean model performance—the square of the correlation between model predictions and true performance ($R^2$)—and RMSE from mixed effect linear regressions of model predictions on observed performance with a random effect of family in the independent sample. Significant effects were bolded. RMSE, root mean square error; CVLT, California Verbal Learning Test; VR I and II, Visual Reproduction I and II.
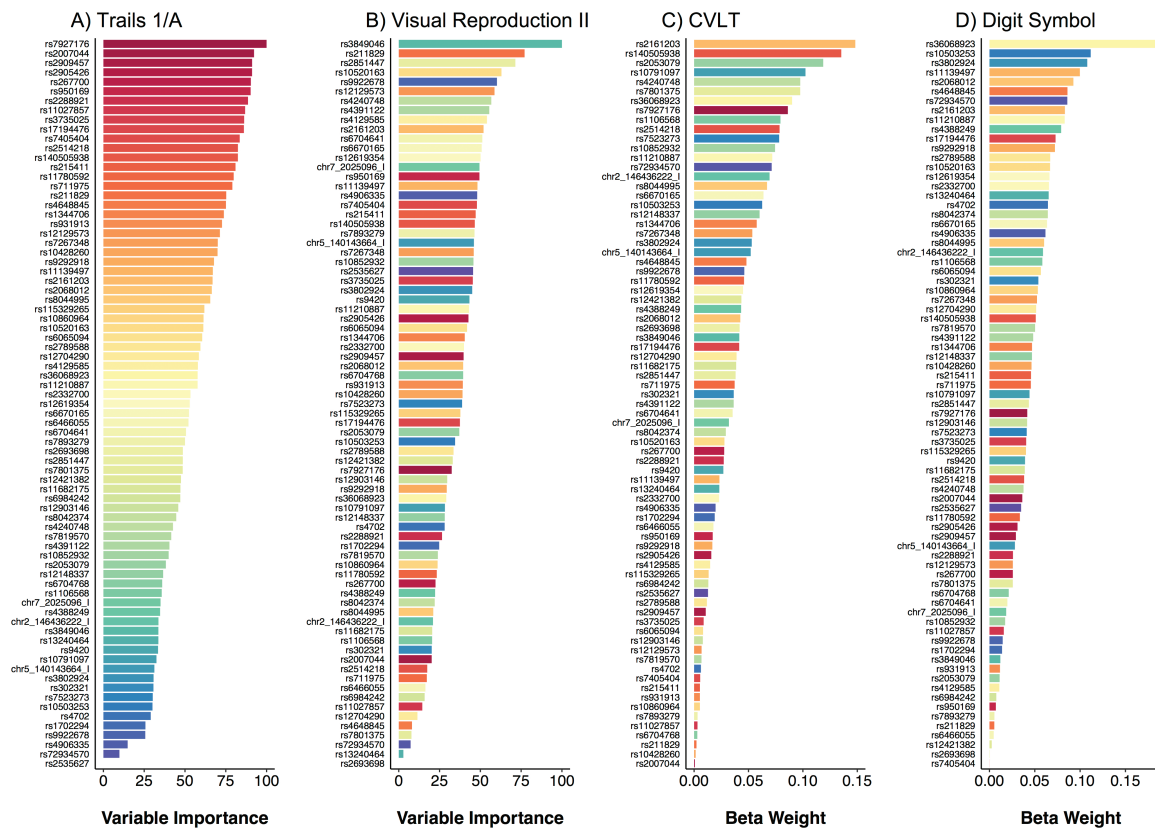


**Fig. 1.** The relative importance or beta-weight magnitude (absolute value of beta weight) of each SNP to the random forest models (A, B) and polygenic risk scores (C, D) that generalized successfully in an independent sample. Colors were assigned to SNPs arbitrarily based on their ranking for the Trails A model to demonstrate that the ranking of SNPs by importance/beta weight varied considerably across cognitive models.

SNPs by magnitude of beta-weight for the 2 polygenic risk scores that generalized.

## Discussion

We tested how well linear and nonlinear models could map schizophrenia risk loci to 6 similar, but distinct, neuropsychological measures. Random forest and polygenic risk scores each generated 2 externally valid models (for 4 different cognitive tasks) with similar predictive strength and error. All of these models showed some task-specificity with respect to their predictions as well, signaling that these models picked up on variance relating to a particular cognitive task above and beyond that shared across cognition broadly. Furthermore, while all models were given the exact same genotypic information, the

1049

importance of each predictor within a given phenotype varied considerably (figure 1), suggesting individual variation in genetic risk for schizophrenia may correspond to nuances in cognitive impairment that can be captured by random forest and polygenic risk scores.

One important component of this study was including independent validation. While all cross-validated models performed well in training, most did not generalize to an external sample as even rigorous *k*-fold internal cross-validation routines can be susceptible to over-fitting, especially when using powerful algorithms in small samples.[24,45-49] Only interpreting the results of externally validated models minimizes the risk of overanalyzing sample-specific associations. This is especially true when distinguishing signal from noise while measuring small effects. Random forest fit comparable models for permuted data in our training sample. However, critically, these models did not produce meaningful predictions when tested out of sample (supplementary information S6). Although such concerns also apply equally to non-machine learning approaches, these demonstrations are sure to become a functional requirement for machine learning studies as they gain popularity.[24,25,46,50,51]

Improving the precision of phenotypic prediction using genotypic risk factors may aid in understanding heterogeneity in etiology and clinical presentation in schizophrenia. Towards this end, both random forest and task- or outcome-specific polygenic risk scores may be useful and our results do not indicate one method is superior. It is possible that as more risk SNPs are identified, more flexible algorithms will perform favorably—this would be likely in the case of nonlinear allele effects or selectivity of SNPs for certain outcomes—but this remains a hypothesis to be tested in future work.

Many limitations should be considered when interpreting these results. First, though we restricted analyses to SNPs and endophenotypes associated with schizophrenia, and our results confirm previous genotype–phenotype findings,[3,52] the inclusion of other patient groups in the study may suggest that these relationships are not specific to schizophrenia and may apply broadly among these illnesses. Further, in selecting only a subset of SNPs significantly associated to schizophrenia, we cannot consider patterns including all causal risk loci nor explain the large majority of variance in any endophenotype. We also tested only schizophrenia-related SNPs, whereas SNPs associated with bipolar disorder[53] and general cognition[54,55] may also have been appropriate, though derived from smaller samples and less well-characterized. However, the current samples were not adequately powered to discover novel loci via a genome-wide search, nor test all applicable SNP lists, so we limited selection to those markers already known to associate with schizophrenia. Nonetheless, for models that generalized externally, most SNPs were retained as predictors, suggesting that this a priori selection was relatively successful, as random forest models drop uninformative predictors. Although this approach precludes the opportunity for gene discovery, it may offer a more sensitive analysis method for moderate sample sizes.

Second, even for the models that successfully generalized, predictive performance (variance explained) was modest. This is unsurprising given that models were built using only 77 SNPs—a tiny fraction of the total SNPs hypothesized to account for complex phenotypes such as schizophrenia or cognition—even though these were the SNPs best associated with schizophrenia in the PGC GWAS. Nonetheless, performance in the current study, using less than 100 SNPs, was comparable to the performance typically seen in traditional efforts examining complex traits using thousands of loci.

Finally, while 4 models did generalize, the rest did not. Typically, it is assumed that models do not generalize because of "overfitting"; that is, when an algorithm is capturing sample-specific (noise) relationships, a problem that is exaggerated when the number of predictors is far larger than the number of observations. In any case, a priori selection in the current study meant that we had around 10 times as many observations as variables and some models did generalize. As such, a more likely explanation is that the external validation sample was not sufficiently representative of the primary training sample. Indeed, the patient composition of the samples was not matched with respect to diagnosis type nor proportion of cases relative to controls. Allelic frequencies differed between samples, as well (supplementary information S7). In addition to sample differences, there were also differences in the task constructs across the 2 samples: 5 tasks required numerical adjustments for differences in scoring ranges, tasks were administered by different practitioners in different languages on different continents, and outliers were observed in the Swedish sample. Given that these discrepancies between samples could only diminish the generalizability of these models (ie, they would inflate type II error rate), those that did generalize are likely robust.

Here, random forest and task-specific polygenic risk scores were able to generate externally valid models with similar predictive strength and error. As such, we suggest that both methods may be useful for mapping a limited set of predetermined, disease-associated SNPs to related phenotypes. Ideally, with the incorporation of more SNPs and phenotypes, both linear and nonlinear methods together will eventually identify how general risk variants for a disorder pool together to indicate likelihood of certain clinical features within an individual. Incorporating random forest and other more flexible algorithms into these inquiries could contribute to parsing heterogeneity within schizophrenia; such algorithms can perform as well as standard methods and can capture a more comprehensive set of potential relationships.

## Supplementary Material

Supplementary data are available at *Schizophrenia Bulletin* online.

## Funding

## References

1. Cannon TD, Kaprio J, Lönnqvist J, Huttunen M, Koskenvuo M. The genetic epidemiology of schizophrenia in a Finnish twin cohort. A population-based modeling study. *Arch Gen Psychiatry*. 1998;55:67–74.

2. Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry*. 2003;60:1187–1192.

3. Lencz T, Knowles E, Davies G, et al. Molecular genetic evidence for overlap between general cognitive ability and risk for schizophrenia: a report from the Cognitive Genomics consorTium (COGENT). *Mol Psychiatry*. 2014;19:168–174.

4. Hatzimanolis A, Bhatnagar P, Moes A, et al. Common genetic variation and schizophrenia polygenic risk influence neurocognitive performance in young adulthood. *Am J Med Genet B Neuropsychiatr Genet*. 2015;168B:392–401.

5. Martin AK, Robinson G, Reutens D, Mowry B. Common genetic risk variants are associated with positive symptoms and decision-making ability in patients with schizophrenia. *Psychiatry Res*. 2015;229:606–608.

6. Derks EM, Vorstman JA, Ripke S, Kahn RS, Ophoff RA; Schizophrenia Psychiatric Genomic Consortium. Investigation of the genetic association between quantitative measures of psychosis and schizophrenia: a polygenic risk score analysis. *PLoS One*. 2012;7:e37852.

7. Fanous AH, Zhou B, Aggen SH, et al.; Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study of clinical dimensions of schizophrenia: polygenic effect on disorganized symptoms. *Am J Psychiatry*. 2012;169:1309–1317.

8. Whalley HC, Sprooten E, Hackett S, et al. Polygenic risk and white matter integrity in individuals at high risk of mood disorder. *Biol Psychiatry*. 2013;74:280–286.

9. Terwissscha van Scheltinga AF, Bakker SC, van Haren NE, et al.; Psychiatric Genome-wide Association Study Consortium. Genetic schizophrenia risk variants jointly modulate total brain and white matter volume. *Biol Psychiatry*. 2013;73:525–531.

10. Ruderfer DM, Fanous AH, Ripke S, et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; Bipolar Disorder Working Group of the Psychiatric Genomics Consortium; Cross-Disorder Working Group of the Psychiatric Genomics Consortium. Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol Psychiatry*. 2014;19:1017–1024.

11. Maier R, Moser G, Chen GB, et al.; Cross-Disorder Working Group of the Psychiatric Genomics Consortium. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am J Hum Genet*. 2015;96:283–294.

12. Gandal MJ, Leppa V, Won H, Parikshak NN, Geschwind DH. The road to precision psychiatry: translating genetics into disease mechanisms. *Nat Neurosci*. 2016;19:1397–1407.

13. Domenici E. Schizophrenia genetics comes to translation. *NPJ Schizophr*. 2017;3:10.

14. Chuang LC, Kuo PH. Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm. *Sci Rep*. 2017;7:39943.

15. Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet*. 2014;10:e1004754.

16. Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genet*. 2010;11:49.

17. Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics*. 2012;99:323–329.

18. Andreasen NC, Wilcox MA, Ho BC, et al. Statistical epistasis and progressive brain change in schizophrenia: an approach for examining the relationships between multiple genes. *Mol Psychiatry*. 2012;17:1093–1102.

19. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*. 2012;109:1193–1198.

20. Guan L, Wang Q, Wang L, Wu B, Chen Y, Liu F, et al. Common variants on 17q25 and gene–gene interactions conferring risk of schizophrenia in Han Chinese population and regulating gene expressions in human brain. *Mol Psychiatry* 2016;21:1244–1250.

21. Nicodemus KK, Hargreaves A, Morris D, et al.; Schizophrenia Psychiatric Genome-wide Association Study (GWAS) Consortium; Wellcome Trust Case Control Consortium 2. Variability in working memory performance explained by epistasis vs polygenic scores in the ZNF804A pathway. *JAMA Psychiatry*. 2014;71:778–785.

22. Botta V, Louppe G, Geurts P, Wehenkel L. Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS One*. 2014;9:e93379.

23. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014;511:421–427.

24. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 2016;3:243–250.

25. Kessler RC, van Loo HM, Wardenaar KJ, et al. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Mol Psychiatry*. 2016;21:1366–1371.

26. Delis DC, Kramer JH, Kaplan E, Ober BA. California Verbal Learning Test – second edition. Adult version. Manual. San Antonio, TX: Psychological Corporation; 2000.

27. Wechsler D. *Wechsler Adult Intelligence Scale—IV*. San Antonio, TX: The Psychological Corporation; 2008.

28. Wechsler D. *Wechsler Memory Scale—IV*. San Antonio, TX: The Psychological Corporation; 2009.

29. D'Elia LF, Satz P, Uchiyama CL, White T. *Color Trails Test. Professional Manual*. Odessa, FL: Psychological Assessment Resources; 1996.

30. Wechsler D. (1999). Wechsler Abbreviated Scale of Intelligence (WASI). San Antonio, TX: Psychological Corporation; 1999.

31. Elwood RW. The Wechsler Memory Scale-Revised: psychometric characteristics and clinical application. *Neuropsychol Rev*. 1991;2:179–201.

32. Reitan RM, Wolfson D. *The Halstead-Reitan Neuropsychological Test Battery : Theory and Clinical Interpretation*. Tucson, AZ: Neuropsychology Press; 1985.

33. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–575.

34. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2011;9:179–181.

35. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet*. 2010;11:499–511.

36. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56–65.

37. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*. 2008;24:2938–2939.

38. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. *Elements*. 2009;1:337–387.

39. Friedman JH. Recent advances in predictive (machine) learning. *J Classif*. 2006;23:175–197.

40. Bernard S, Heutte L, Adam S. Influence of hyperparameters on random forest accuracy. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; 2009:171–180.

41. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Wadsworth, Belmont, CA. 1984.

42. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28:1–26.

43. Liaw A, Wiener M. Classifcation and regression by random-Forest. *R News* 2002;2:18–22.

44. Breiman L. Random forest. *Mach Learn*. 1999;45:1–35.

45. Ng, A.Y. Preventing "overfitting" of cross-validation data. In: Proceedings of the 14th International Conference on Machine Learning (Nashville, TN, 1997). Morgan Kaufmann, San Mateo, CA, 245–253.

46. Uher R, Perlis RH, Henigsberg N, et al. Depression symptom dimensions as predictors of antidepressant treatment outcome: replicable evidence for interest-activity symptoms. *Psychol Med*. 2012;42:967–980.

47. Klockars AJ, Sax G. *Multiple Comparisons*. Thousand Oaks, CA: SAGE Publications, Inc; 1986.

48. Tukey JW. Comparing individual means in the analysis of variance. *Biometrics*. 1949;5:99–114.

49. Tukey JW. The philosophy of multiple comparisons. *Stat Sci*. 1991;6:100–116.

50. Paulus MP, Huys QJ, Maia TV. A roadmap for the development of applied computational psychiatry. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2016;1:386–392.

51. Chekroud A, Gueorguieva R, Krumholz H, et al. Reevaluating the efficacy and predictability of antidepressant treatments. *JAMA Psychiatry*. 2017;74:370–378. doi: 10.1001/jamapsychiatry.2017.0025.

52. Mark W, Toulopoulou T. Cognitive intermediate phenotype and genetic risk for psychosis. *Curr Opin Neurobiol*. 2016;36:23–30.

53. Hou L, Bergen SE, Akula N, et al. Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. 2016. doi:10.1101/044412.

54. Davies G, Armstrong N, Bis JC, et al. Genetic contributions to variation in general cognitive function : a meta-analysis of genome-wide association studies in the CHARGE consortium ($N$ = 53 949). *Mol Psychiatry*. 2015;20:183–192.

55. Trampush JW, Yang MLZ, Yu J, et al. GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: a report from the COGENT consortium. *Mol Psychiatry*. 2017;22:336–345.