### Title

A genetic risk score using human chromosomal-scale length variation can predict schizophrenia




### Journal

Scientific Reports, 11(1)

### ISSN

2045-2322



### Authors

Toh, Christopher
Brody, James P



### Publication Date

2021

### DOI

10.1038/s41598-021-97983-0

Peer reviewed

scientific reports

Check for updates

OPEN

# A genetic risk score using human chromosomal-scale length variation can predict schizophrenia

Christopher Toh & James P. Brody✉

**Studies indicate that schizophrenia has a genetic component, however it cannot be isolated to a single gene. We aimed to determine how well one could predict that a person will develop schizophrenia based on their germ line genetics. We compared 1129 people from the UK Biobank dataset who had a diagnosis of schizophrenia to an equal number of age matched people drawn from the general UK Biobank population. For each person, we constructed a profile consisting of numbers. Each number characterized the length of segments of chromosomes. We tested several machine learning algorithms to determine which was most effective in predicting schizophrenia and if any improvement in prediction occurs by breaking the chromosomes into smaller chunks. We found that the stacked ensemble, performed best with an area under the receiver operating characteristic curve (AUC) of 0.545 (95% CI 0.539–0.550). We noted an increase in the AUC by breaking the chromosomes into smaller chunks for analysis. Using SHAP values, we identified the X chromosome as the most important contributor to the predictive model. We conclude that germ line chromosomal scale length variation data could provide an effective genetic risk score for schizophrenia which performs better than chance.**

**Abbreviations**

| | |
|---|---|
| AUC | Area under the curve |
| CNV | Copy number variation |
| CSLV | Chromosomal-scale length variation |
| GBM | Gradient boosted machines |
| GLM | General linear model |
| ROC | Receiver operator curve |
| SHAP | SHapley Additive exPlanations |
| SNP | Single nucleotide polymorphism |

Schizophrenia is a highly heritable, complex psychiatric disorder[1,2]. Genome wide association studies have identified over one hundred genetic loci that contribute to its heritability[2–7]. However, these loci still account for less than half of the genetic risk for schizophrenia[3]. Environmental exposure to chemicals appears to play almost no role in the development of schizophrenia, but different forms of trauma experienced during development does appear to be a risk factor[8]. Twin studies have consistently shown a significant genetic contribution to schizophrenia, and many twin studies find that the environmental contribution to schizophrenia exists but that genetic effects provide significant liability to schizophrenia[9].

Genetic risk scores[10–12] have been developed for many different forms of disease, including breast cancer[13], coronary artery disease[14], and stroke[15]. Polygenic risk scores based on SNPs clearly can predict schizophrenia. One study measured an odds ratio of about 8 (95% CI 4–14) for the highest decile compared to the lowest decile[16]. A second study found that polygenic risk scores for schizophrenia (and bipolar disorder) are also associated with creativity[17]. A review of polygenic risk scores for schizophrenia highlighted the difficulty these studies had finding a consistent diagnosis of schizophrenia[18]. One limitation of polygenic risk scores is that they only consider linear combinations of SNPs.

Copy number variations (CNVs) in germ line DNA have also been associated with schizophrenia[4,5,19–24]. Evidence suggests that these CNVs associated with schizophrenia are represented also by SNPs[24]; the predictive power of CNVs does not add to the predictive power of SNPs when using linear prediction algorithms. The dimensionality of the data (many more SNPs than patients with schizophrenia) precludes the use of non-linear machine learning techniques.

Department of Biomedical Engineering, University of California, Irvine, USA. ✉email: jpbrody@uci.edu

Chromosome-scale length variation (CSLV) reduces the dimensionality of the data, while maintaining sufficient information for predictive algorithms. Combining CSLV with modern machine learning classification algorithms provides a powerful tool to predict phenotypes from a person's genome[25]. The CSLV values are averages, across all or most of a chromosome, of copy number variation (CNV) measured at each SNP location. This method is particularly appealing for genetic risk scores because it includes epistatic effects that might be missed with conventional genome wide association studies, which use logistic regression—a linear combination of SNP scores. By attempting to still utilize every CNV value, this model aims to demonstrate that there are likely global CNV interactions which may be missed by conventional genetic risk scores.

The purpose of this paper is to evaluate how well a genetic risk score based on chromosome-scale length variation and machine learning classification algorithms can predict schizophrenia in individuals. We evaluated this approach on a dataset of 1129 patients who had schizophrenia in the UK Biobank dataset. These patients were previously genotyped as part of the UK Biobank project.

## Methods

Data was obtained from the UK Biobank under Application Number 47850. The UK Biobank project collected extensive data from about 500,000 people who were between the ages of 40 and 69 during the 2006–2010 recruitment years. This data included genotyping data and medical records. In addition, most of the participants' medical records are linked, through the National Health Service, to the UK Biobank records. This linkage provides for ongoing follow-up of health conditions[26,27].

First, we downloaded the "l2r" files from the UK Biobank. Each chromosome has a separate "l2r" file. Each "l2r" file contained 488,377 columns and a variable number of rows. Each column represented a unique patient in the dataset, who can be identified with an encoded ID number. Each row represented a different location in the genome. The values in the file represent the log base 2 ratio of intensity relative to the expected two copies measured at the SNP location.

After downloading the "l2r" data from the UK Biobank, we computed the mean l2r value for different portions of each chromosome for each patient in the dataset. We created three different datasets, which we refer to as "splits". We split each chromosome into either 1, 4, or 8 nominally equal parts. Then, we compute the length for each person's chromosome split using the l2r files by taking the average of all l2r values measured within that portion of the chromosome split. A value of 0 represents the nominal average length of that portion of the particular chromosome. We call this dataset the chromosome-scale length variation (CSLV) dataset.

The CLSV numbers represent the copy number of the genomic DNA recognized by the probe. We computed a measure of the length of chromosomes, or chromosome fragments, by averaging these l2r measurements from different probes along the chromosome. For each person, we have 1 split, 4 split, and 8 split datasets. The 1 split data consists of 23 numbers, one for each of the autosomes and one for the X chromosome. The 4 split data consists of 92 numbers and the 8 split data has 184 numbers for each person.

This CSLV dataset was matched with the UK Biobank Health records dataset. UK Biobank matched the person in the Public Health England data with UK Biobanks internal records to produce the person's encoded participant ID. The dataset we have, provided by UK Biobank, contains the participant ID and date the patient was diagnosed by a doctor as having schizophrenia.

Using the CSLV-Schizophrenia dataset, we selected all people who had a diagnosis of schizophrenia and labelled them in the dataset. We constructed an age-matched control group of the same size that had an identical age profile as those in the schizophrenia group. The age-matched control group was selected from all those in the UK Biobank dataset having no indication of schizophrenia. Since only a small fraction of the people in the UK Biobank had a schizophrenia diagnosis, we could rerun the analysis with a different age-matched control group many times to build up statistics.

We used the $H_2O$ machine learning package in R[28,29]. We created 100 machine learning models that were trained to classify a person in the dataset, consisting of those who had schizophrenia and age-matched controls, based solely on their chromosome scale length variation data. Each model was trained with fivefold cross-validation. Each model had a distinct set of controls. These models were trained to perform a binary classification, distinguishing between those who had been diagnosed with schizophrenia and those who did not have schizophrenia. The models were evaluated by measuring the area under the curve of the receiver operating characteristic curve, known as the AUC.

The $H_2O$ package implements several common machine learning algorithms. Distributed Random Forest (drf) is based on an algorithm originally called "Extremely randomized trees"[30]. The Gradient Boosting Machine algorithm (gbm) builds regression trees in parallel[31,32]. The generalized linear model (glm) is implemented using an augmented linear model[33–35]. XGBoost is a refinement to the general Gradient Boosting Machine algorithm[36]. Ensembles are a combination of these other machine learning algorithms. This combination often provides superior results to any particular algorithm[37,38]. The $H_2O$ package implements stacked ensembles as super learner algorithms[39]. The $H_2O$ package also uses SHAP values to interpret the models[40]. SHAP values are measures of how important different features are to the prediction.

Our computer analysis system is a Linux server running Ubuntu 18.04. The system is a 64-bit system running two Intel Xeon E5-2690 2.90 GHz CPUs. It also has a GeForce GT 710 NVIDIA GPU. 32 GBs of RAM were also available with a 10 TB HDD.

**Ethics approval and consent to participate.** Ethics approval and participant consent was collected by UK Biobank at the time participants enrolled. All subjects in the database have given informed consent, and if under 18, consent from a parent and/or legal guardian. Additionally, all subjects have the ability to withdraw at any time from the UK Biobank. This paper is an analysis of anonymized data provided by UK Biobank. Accord-

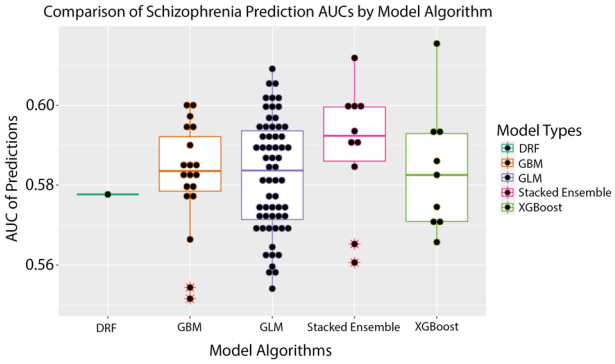


**Figure 1.** This boxplot figure presents the results of the machine learning predictions. We created 100 different datasets. For each dataset, we used the same set of schizophrenia patients with a distinct set of age matched people from the general UK Biobank population as controls. Then H2O was used to perform a grid-search of possible best algorithms. The best performing algorithm was then reported with an AUC. The differences between algorithms is reported here. The machine learning algorithms tested were distributed random forests (drf), gradient boosting machine (gbm), general linear model (GLM), stacked ensemble (a combination of the other four algorithms) and XGBoost (XGBoost).

ing to UC Irvine's IRB, analysis of anonymized data does not constitute Human Subjects Research. All methods and experimental research protocols were approved by the UK Biobank.

## Results

Figure 1 presents results showing the performance of different machine learning algorithms. We found that the stacked ensemble models consistently performed best. As Fig. 1 shows, we found a slight difference between algorithms and their performance. But all algorithms could predict schizophrenia significantly better than chance (AUC = 0.50). This finding indicates that germ line genetics of the patient, as represented by the set of chromosome-scale length variation numbers, demonstrates predictability of schizophrenia.

The AUC (area under the curve of the receiver operating characteristic curve) for the machine learning classification models was 0.583 (standard deviation 0.014, 95% confidence interval of 0.581–0.586). A classification model with an AUC of 0.50 is equivalent to random guessing. The measured AUC differs from 0.50 with $p < 0.00001$.

We also tested how well each model could predict schizophrenia on a holdout set of validation data. The holdout set was 30% of the original test data and was not included in the training of the models. The AUC of the holdout set was 0.5734 with a 95% confidence interval of 0.569–0.578.

We then tested whether increasing the number of splits improves model performance. We constructed three overlapping datasets with 1 split, 4 splits, and 8 splits. The phrase "1 split" represents the average l2r value measured across an entire chromosome for all 23 chromosomes giving a total of 23 numbers, "4 splits" represents the average of each quarter of the 23 chromosomes l2r values for a total of 92 numbers, and "8 splits" represent the average of each eighth of the 23 chromosomes' l2r values for a total of 184 numbers.

Figure 2 shows how models compare on the 3 different split datasets. Overall, a stacked ensemble had the best performance, however a general linear model (glm) was most often the best candidate model.

In all models, increasing splits improves model performance for the same runtime. Figure 3 demonstrates the difference of all models for 1 split, 4 splits, and 8 splits datasets. We tested whether finer splits of the dataset provided significantly improved AUCs. As shown in Table 1, the p-value of the 4 splits model compared to the 1 split model is $p = 1 \times 10^{-24}$. Comparing the mean AUC for the 8 splits model to the 1 split model gave a p-value of $p = 3 \times 10^{-30}$ indicating that finer splits significantly improved the predictive ability of the models. The 4 splits and 8 splits models performed better than the 1 split models by a significant amount.

We then calculated the odds ratio (OR) of our predictions drawn from the cross-validated model. Table 2 shows that a patient in the upper quintile is approximately twice as likely to have schizophrenia when compared to the lower quintile.

In order to understand how our models came to their conclusions, we created several plots to explain them from H2O's "explainability" framework. The first is a variable importance heatmap across the generated models which is shown in Fig. 4. Our analysis here indicated that chromosome X was one of the highest contributing variables in predicting Schizophrenia, especially in tree models such as GBM and XGBoost. We then confirmed this with a Shapley Additive exPlanation or SHAP plot in Fig. 5. This plot also indicates that chromosome X was the leading factor in our leading model for predicting schizophrenia.

Utilizing our findings above, we then proceeded to train new models from scratch using only CSLV values from chromosome X but with 64 CSLV splits. This model did not contain any information from the 22 autosomes but instead relied solely on CNVs in the X chromosome and our aim was to see if the model would be comparable to our previous 4-split and 8-split models. We found that on average these models had a comparable performance of about 0.58 with the highest being around 0.627 as shown in Fig. 6.

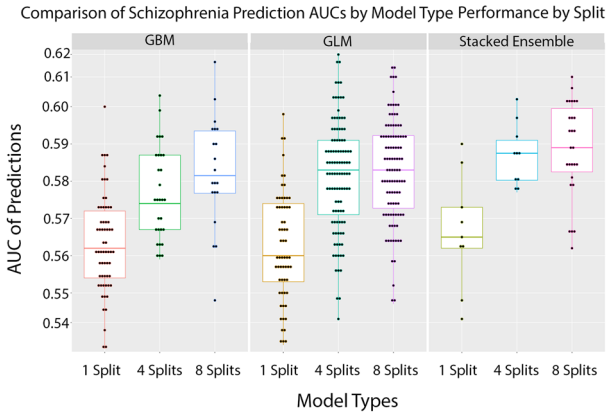


**Figure 2.** We tested whether finer splits of each chromosome lead to better predictability. We split each chromosome into either one, four, or eight subsections. We computed the chromosome scale length variation for each of these subsections for each person. This set of numbers was used to predict whether patients had schizophrenia. The quality of this prediction was characterized by the AUC. This plot demonstrates how the quality of these predictions increase with finer information on chromosome length variation. The Stacked Ensemble algorithm performs the best across all split variations.

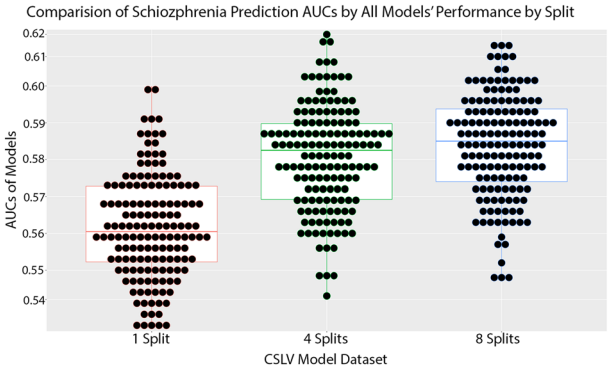


**Figure 3.** This plot represents the average performance of 150 models for each split type for a total of 450 models.

| Dataset | Mean AUC | Standard Deviation | P-value vs 1 split |
|---|---|---|---|
| 1 split | 0.5614 | 0.0148 | |
| 4 splits | 0.5807 | 0.0146 | $1 \times 10^{-24}$ |
| 8 splits | 0.5838 | 0.0141 | $3 \times 10^{-30}$ |

**Table 1.** The mean and standard deviation of the cross validated AUCs of 1 split, 4 splits, and 8 splits datasets of 150 models each.

| Quintile | Normal | Schizophrenia | Odds ratio | Count | 95% CI |
|---|---|---|---|---|---|
| 1 | 185 | 123 | 0.67 | 308 | 0.51–0.85 |
| 2 | 156 | 152 | 0.97 | 308 | 0.76–1.24 |
| 3 | 153 | 155 | 1.0 | 308 | 0.79–1.3 |
| 4 | 142 | 165 | 1.2 | 307 | 0.91–1.5 |
| 5 | 133 | 174 | 1.3 | 307 | 1.0–1.7 |

**Table 2.** This table represents the odds ratio between the quintiles of predicted results from our cross-validated results. The result indicates that the top quintile is twice as likely to have an accurate prediction for Schizophrenia as the bottom quintile.

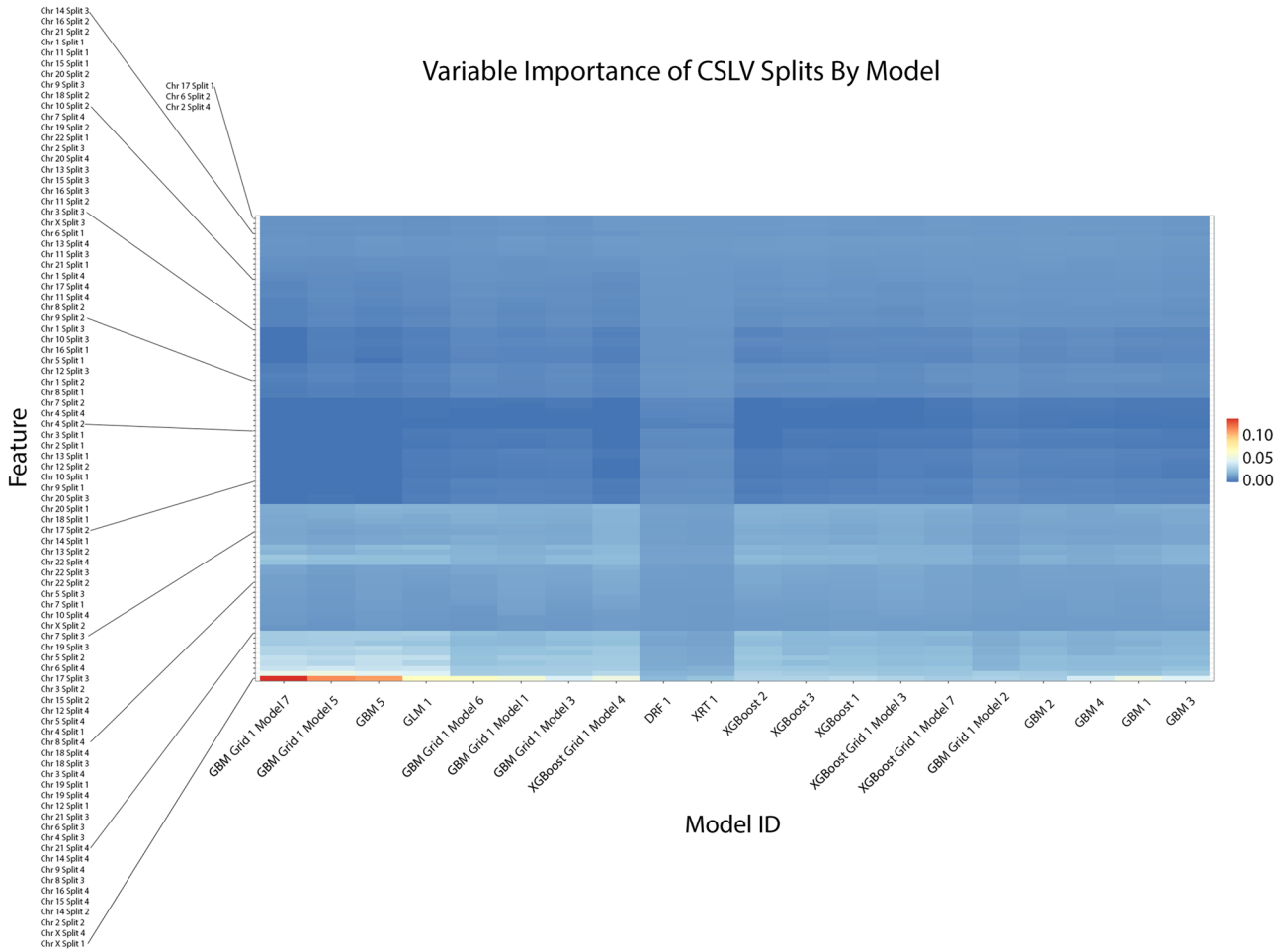


**Figure 4.** This variable importance heatmap shows the variables which most affected the performance and outcome of decisions made by the specified model. A value closer to 1.0 indicates higher importance of that variable. In most tree-based models the CSLV values for chromosome X have the highest importance.

We then again performed a variable importance heatmap analysis to get greater granularity of our understanding of the contributing CSLVs in chromosome X. We found that this was again consistent with the previous findings from the 4-split model. Figure 7 indicates that the top features of variable importance are again being found in the first and last regions of chromosome X. As such it appears that the majority of the predictive power of any model trained with CSLV and when predicting schizophrenia in an individual is a result of CNVs on chromosome X. We also report corresponding estimates of hg38 coordinates in Table 3.

We wanted to ensure these results were not due to inherent sex differences. We trained 50 models using the 64 split chromosome X dataset which were not only age-matched with the controls but also sex-matched. 25 of the AutoML models were trained with the actual data with correctly labeled disease states. The other 25 AutoML models were trained with the schizophrenia diagnosis randomly shuffled. The results are shown in Table 4. Here we can see that a portion of the previous performance is most likely due to CSLV differences inherent between males and females (Supplemental D). However, a portion of the prediction is statistically still better than random guessing.

## Discussion

These results indicate that germline genetic variation contributes at least to some degree to the onset of schizophrenia in individuals. Our results indicate that genetic structural variation across the global chromosomal scope is sufficient to predict, better than guessing, whether or not an individual will have schizophrenia. The patients were an equal number of patients by gender between the control and disease group and the ages of patients in the control group also were matched to the ages of patients in the disease group. Further analysis revealed that length variation in a handful of regions of the X chromosome was sufficient to reproduce the predictive model. Recently, there has been revived discussion of copy number variations as a large contributing factor to several neurological ailments including schizophrenia[41]. Additionally, hypotheses about sex chromosome links to schizophrenia inheritance have been discussed for several decades and our findings lend support to this idea[42].

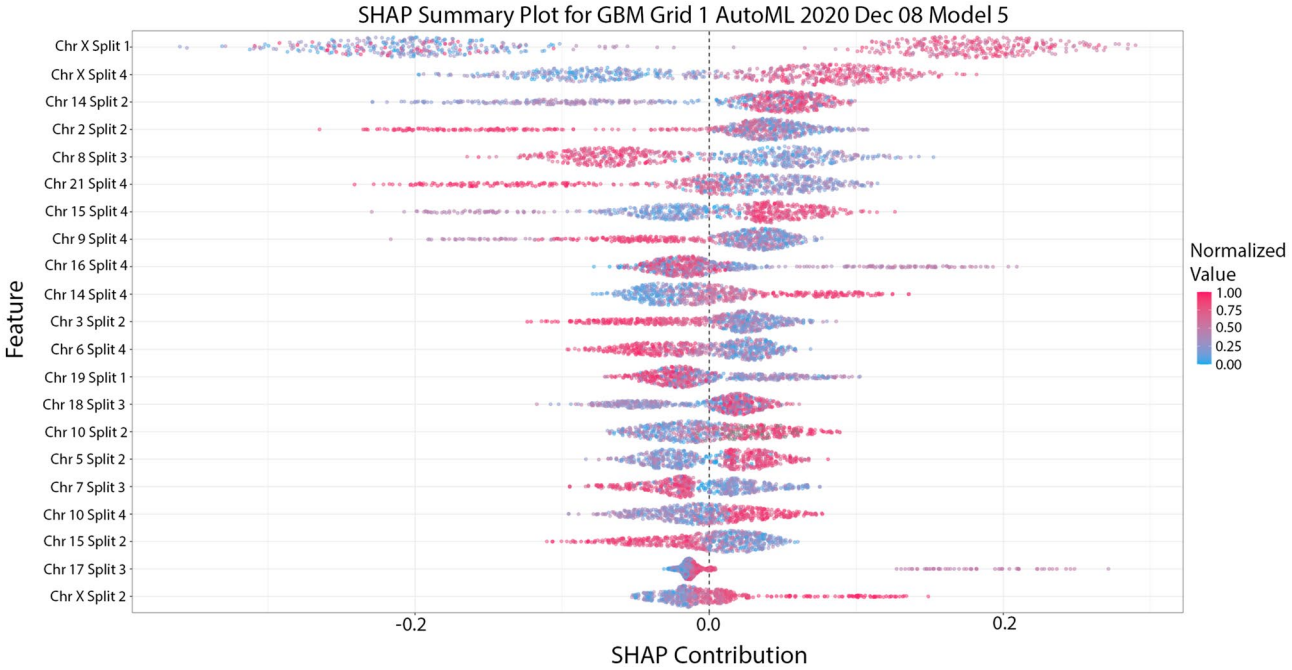


**Figure 5.** This SHAP plot indicates that the leading model for our 4-splits model relied heavily on the first quarter and last quarter value of chromosome X with some contribution from other regions and the second quarter of chromosome X.

On average, a stacked ensemble is the best approach to creating a predictive model for the prediction of schizophrenia. However, all models that were tested still created models with predictive power better than chance (Supplementary Information A, B, & C). Since $H_2O$'s AutoML performs a grid-search of all the possible datasets and each trial we ran included the same disease group but with a different control groups, we can see in Fig. 1 that a general linear model (GLM) oftentimes was still the best option. Gradient Boosted Machines (GBM) and XGBoost also typically performed the same as GLM.

Utilizing a more granularized dataset by splitting the autosomes into quarters and eighths performs significantly better than using a CSLV averaged across an entire chromosome. This observation suggests we can increase performance by increasing splits. In the future, we plan on exploring the trade off in run time and computational resources required by increasing splits. Other methods of dimensionality reduction may also yield better results without sacrificing runtime performance.

The CSLV values are averages of copy number variation (CNV) measured at each SNP location. Simply using every single CNV value introduces a dimensionality problem as our dataset only has roughly 488,000 individuals while the total number of CNV values is 764,257 across the 22 autosomes and an additional 18,857 CNV values for the X Chromosome. This means there is likely diminishing returns for using more splits unless it can be offset with increased data.

This approach has several limitations. First, CSLV is an averaged measure of copy-number variations across a large section of the entire chromosome. We used SHAP values to highlight the regions that seem to be more important, but this does not provide a mechanistic explanation. Second, the dataset lacks diversity. The UK Biobank population is primarily Caucasian individuals in the United Kingdom (although not exclusively). Third, the diagnosis of schizophrenia in an individual is difficult to quantify and the disease might consist of a heterogeneous group of underlying biological processes. Finally, this analysis is based on a single dataset and the conclusions would be stronger if the analysis could be replicated in an independent dataset. However, similar datasets are not currently available.

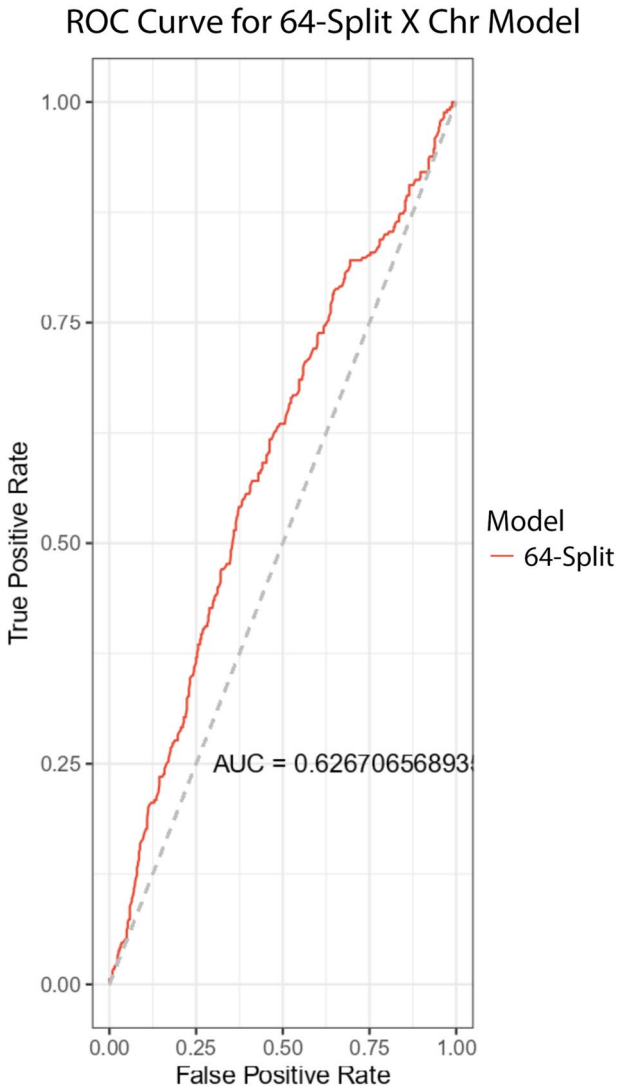


**Figure 6.** This ROC Curve for a schizophrenia prediction model utilizing 64-splits or 64 CSLVs of chromosome X only. The reported AUC is 0.627.

## Conclusion

We were able to create machine learning models for prediction of schizophrenia in patients. These models perform better than chance with an average AUC of 0.545. Prediction was performed with only chromosomal scale length variation measurements as the input variables. Further analysis of the SHAP values suggests that the length variation of several regions of the X chromosome are sufficient to reproduce this predictive value.

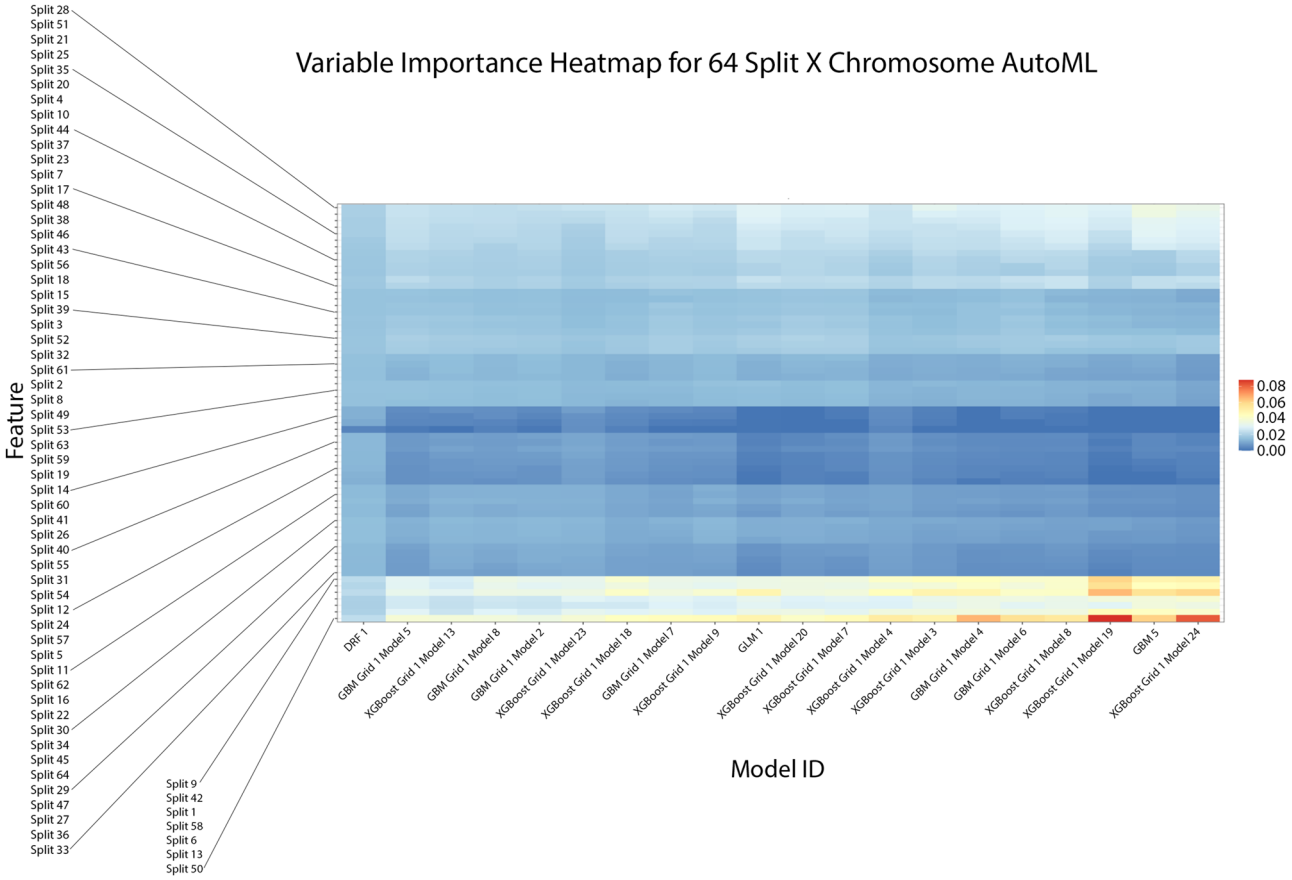


**Figure 7.** This variable importance heatmap shows the variables which most affected the performance and outcome of decisions made by the specified model. A value closer to 1.0 indicates higher importance of that variable. In most of the models we find that the CSLV values were mostly centered around split 50, 1, 9, 42, 13, 58, and 6. This is consistent with Fig. 4.

| CSLV Split | Estimated hg38 Coordinates |
|---|---|
| 1 | chrX:60425–634774 |
| 6 | chrX:5651118–7792613 |
| 9 | chrX:11426091–13234434 |
| 13 | chrX:20912585–22990332 |
| 42 | chrX:107331058–110669244 |
| 50 | chrX:128031497–130523635 |
| 58 | chrX:145709120–147908169 |

**Table 3.** This table shows the estimated hg38 coordinates for the corresponding CSLV splits with high variable importance as shown in Fig. 7.

| Dataset | Mean AUC | Standard Deviation |
|---|---|---|
| 64 Split × normal | 0.545 | 0.01373103 |
| 64 Split × random | 0.525 | 0.01363745 |
| Welch two sample t-test between normal and random | T = − 5.0111<br>df = 47.998 | p-value = 7.763e-06 |

**Table 4.** This table shows a comparison of the age and sex matched models using 64 Split chromosome X data. The reported mean AUCs demonstrates that a portion of the previous performance is attributed to differences between male and females in X Chromosome CSLV levels as shown in Supplementary Information D. However, it still performs better than randomly guessing.

## Data availability

The datasets analyzed during the current study are available from UK Biobank at https://www.ukbiobank.ac.uk/.

Received: 25 May 2021; Accepted: 1 September 2021
Published online: 22 September 2021

## References

1. Flint, J. & Munafò, M. Genesis of a complex disease. *Nature* **511**, 412–413. https://doi.org/10.1038/nature13645 (2014).
2. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* https://doi.org/10.1038/nature13595 (2014).
3. Lee, S. H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* https://doi.org/10.1038/ng.1108 (2012).
4. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* https://doi.org/10.1038/ng.2742 (2013).
5. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* https://doi.org/10.1038/ng.940 (2011).
6. Farrell, M. S. *et al.* Evaluating historical candidate genes for schizophrenia. *Mol. Psychiatry.* **20**, 555–562. https://doi.org/10.1038/mp.2015.16 (2015).
7. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* https://doi.org/10.1038/nature08185 (2009).
8. Van Os, J., Kenis, G. & Rutten, B. P. F. The environment and schizophrenia. *Nature* https://doi.org/10.1038/nature09563 (2010).
9. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait: Evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* https://doi.org/10.1001/archpsyc.60.12.1187 (2003).
10. Sugrue, L. P. & Desikan, R. S. What are polygenic scores and why are they important?. *JAMA* **321**, 1820. https://doi.org/10.1001/jama.2019.3893 (2019).
11. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-018-0018-x (2018).
12. Lello, L., Raben, T. G., Yong, S. Y., Tellier, L. C. A. M. & Hsu, S. D. H. Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. *Sci. Rep.* https://doi.org/10.1038/s41598-019-51258-x (2019).
13. Mavaddat, N. *et al.* Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.* https://doi.org/10.1016/j.ajhg.2018.11.002 (2019).
14. Fahed, A. C. *et al.* Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* https://doi.org/10.1038/s41467-020-17374-3 (2020).
15. Abraham, G. *et al.* Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat. Commun.* https://doi.org/10.1038/s41467-019-13848-1 (2019).
16. Agerbo, E. *et al.* Polygenic risk score, parental socioeconomic status, family history of psychiatric disorders, and the risk for schizophrenia: A Danish population-based study and meta-analysis. *JAMA Psychiat.* https://doi.org/10.1001/jamapsychiatry.2015.0346 (2015).
17. Power, R. A. *et al.* Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat. Neurosci.* https://doi.org/10.1038/nn.4040 (2015).
18. Mistry, S., Harrison, J. R., Smith, D. J., Escott-Price, V. & Zammit, S. The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: Systematic review. *Schizophr. Res.* https://doi.org/10.1016/j.schres.2017.10.037 (2018).
19. Ruderfer, D. M., Chambert, K., Moran, J., Talkowski, M., Chen, E. S., Gigek, C. *et al.* Mosaic copy number variation in schizophrenia. *Nature*. **508** (2014).
20. Szatkiewicz, J. P. *et al.* Copy number variation in schizophrenia in Sweden. *Mol. Psychiatry.* https://doi.org/10.1038/mp.2014.40 (2014).
21. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* https://doi.org/10.1038/nature12975 (2014).
22. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260. https://doi.org/10.1038/ng.237 (2008).
23. Derks, E. M. *et al.* Investigation of the genetic association between quantitative measures of psychosis and schizophrenia: A polygenic risk score analysis. *PLoS One* https://doi.org/10.1371/journal.pone.0037852 (2012).
24. Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* https://doi.org/10.1038/ng.3725 (2017).
25. Toh, C. & Brody, J. P. Analysis of copy number variation from germline DNA can predict individual cancer risk. *bioRxiv.* https://doi.org/10.1101/303339 (2018).
26. Sudlow, C. *et al.* UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* https://doi.org/10.1371/journal.pmed.1001779 (2015).
27. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209. https://doi.org/10.1038/s41586-018-0579-z (2018).
28. Click, C., Malohlava, M., Candel, A., Roark, H. & Parmar, V. Gradient Boosting Machine with $H_2O$. 30. Accessed 7 April 2021. https://www.H2OAi/Resources/ (2017).
29. Aiello, S., Eckstrand, E., Fu, A., Landry, M. & Abouyon, P. Machine learning with R and $H_2O$. $H_2O$ booklet, 550. It's available at this URL: https://h2o-release.s3.amazonaws.com/h2o/master/3283/docs-website/h2odocs/booklets/R_Vignette.pdf (2016).
30. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* https://doi.org/10.1007/s10994-006-6226-1 (2006).
31. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232. https://doi.org/10.1214/aos/1013203451 (2001).
32. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2 (2002).
33. Lee, Y. & Nelder, J. A. Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* https://doi.org/10.1093/biomet/88.4.987 (2001).
34. Lee, Y. & Nelder, J. A. Hierarchical generalized linear models. *J. R. Stat. Soc. Ser. B (Methodol.)* https://doi.org/10.1111/j.2517-6161.1996.tb02105.x (1996).
35. Nelder, J. A., Lee, Y., & Pawitan, Y. Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood, Second Edition (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315119953 (2017).
36. Chen, T. & Guestrin, C. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'16*. 785–794 (ACM Press, 2016). https://doi.org/10.1145/2939672.2939785.
37. Wolpert, D. H. Stacked generalization. *Neural Netw.* https://doi.org/10.1016/S0893-6080(05)80023-1 (1992).
38. Breiman, L. Stacked regressions. *Mach. Learn.* https://doi.org/10.1007/bf00117832 (1996).

39. Van Der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super learner. *Stat. Appl. Genet. Mol. Biol.* https://doi.org/10.2202/1544-6115.1309 (2007).
40. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777) https://proceedings.neurips.cc//paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf (2017).
41. Zarrei, M. *et al.* A large data resource of genomic copy number variation across neurodevelopmental disorders. *NPJ Genom. Med.* https://doi.org/10.1038/s41525-019-0098-3 (2019).
42. Bache, W. K. & DeLisi, L. E. The sex chromosome hypothesis of schizophrenia: Alive, dead, or forgotten? A commentary and review. *Mol. Neuropsychiatry.* **4**, 83–89. https://doi.org/10.1159/000491489 (2018).



## Acknowledgements
The data used in this study was obtained from the UK Biobank under Application Number 47850.


## Author contributions
C.T. conceived of the study. C.T. and J.B. analyzed the UK Biobank data. C.T. and J.B. contributed to the manuscript. All authors read and approved the final manuscript.


## Funding
No external funding supported this research.


## Competing interests
The authors declare no competing interests.

## Additional information
**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-97983-0.

**Correspondence** and requests for materials should be addressed to J.P.B.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.