Exploring the Intersection of Schizophrenia, Machine Learning, and Genomics: A Scoping Review

Alexandre Hudon, Mélissa Beaudoin, Kingsada Phraxayavong, Stéphane Potvin, Alexandre Dumais

Submitted to: JMIR Bioinformatics and Biotechnology
on: May 30, 2024

Table of Contents

Exploring the Intersection of Schizophrenia, Machine Learning, and Genomics: A Scoping Review

Alexandre Hudon[1, 2, 3] BEng, MD; Mélissa Beaudoin[4, 5] MSc; Kingsada Phraxayavong[6]; Stéphane Potvin[4, 1] PhD; Alexandre Dumais[6, 1, 4, 7] MD, PhD

[1]Centre de recherche de l'Institut universitaire en santé mentale de Montréal Montréal CA
[2]Department of psychiatry and addictology Faculty of Medicine Université de Montréal Montréal CA
[3]Institut universitaire en santé mentale de Montréal Montréal CA
[4]Department of psychiatry and addictology Université de Montréal Montréal CA
[5]Faculty of Medicine McGill University Montréal CA
[6]Services et Recherches Psychiatriques AD Montréal CA
[7]Institut nationale de psychiatrie légale Philippe-Pinel Montréal CA

Corresponding Author:
Alexandre Dumais MD, PhD
Department of psychiatry and addictology
Université de Montréal
2900 Edouard Montpetit Blvd
Montréal
CA

Abstract

Background: An increasing body of literature highlights the integration of machine learning with genomic data in psychiatry, particularly for complex mental health disorders such as schizophrenia. These advanced techniques offer promising potential for uncovering various facets of these disorders. A comprehensive review of the current applications of machine learning in conjunction with genomic data within this context can significantly enhance our understanding of the current state of research and its future directions.

Objective: The objective of this study is to conduct a systematic scoping review of the use of machine learning algorithms with genomic data in the field of schizophrenia.

Methods: A systematic, scoping review, search was performed in the electronic databases of Medline, Web of Science, PsycNet (PsycINFO), and Google Scholar from 2013 to 2024. Studies at the intersection of schizophrenia, genomic data, and machine learning were evaluated.

Results: The literature search identified 2437 eligible articles after removing duplicates. Following abstract screening, 143 full-text articles were assessed, and 121 were subsequently excluded. Therefore, 21 studies were thoroughly assessed. Various machine learning algorithms were used in the identified studies, with support vector machines being the most common. The studies notably used genomic data to predict schizophrenia, identifying schizophrenia features, discovering drugs, classifying schizophrenia amongst other mental health disorders, and predicting the quality-of-life of patients.

Conclusions: Several high-quality studies were identified. Yet, the application of machine learning with genomic data in the context of schizophrenia remains limited. Future research is essential to further evaluate the portability of these models and to explore their potential clinical applications.

(JMIR Preprints 30/05/2024:62752)
DOI: https://doi.org/10.2196/preprints.62752

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users. Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Review

Exploring the Intersection of Schizophrenia, Machine Learning, and Genomics: A Scoping Review

Abstract

Background: An increasing body of literature highlights the integration of machine learning with genomic data in psychiatry, particularly for complex mental health disorders such as schizophrenia. These advanced techniques offer promising potential for uncovering various facets of these disorders. A comprehensive review of the current applications of machine learning in conjunction with genomic data within this context can significantly enhance our understanding of the current state of research and its future directions.
Objective: The objective of this study is to conduct a systematic scoping review of the use of machine learning algorithms with genomic data in the field of schizophrenia.
Methods: A systematic, scoping review, search was performed in the electronic databases of Medline, Web of Science, PsycNet (PsycINFO), and Google Scholar from 2013 to 2024. Studies at the intersection of schizophrenia, genomic data, and machine learning were evaluated.
Results: The literature search identified 2437 eligible articles after removing duplicates. Following abstract screening, 143 full-text articles were assessed, and 121 were subsequently excluded. Therefore, 21 studies were thoroughly assessed. Various machine learning algorithms were used in the identified studies, with support vector machines being the most common. The studies notably used genomic data to predict schizophrenia, identifying schizophrenia features, discovering drugs, classifying schizophrenia amongst other mental health disorders, and predicting the quality-of-life of patients.
Conclusions: Several high-quality studies were identified. Yet, the application of machine learning with genomic data in the context of schizophrenia remains limited. Future research is essential to further evaluate the portability of these models and to explore their potential clinical applications.

Keywords: Schizophrenia; Genomic data; Machine learning; Prediction; Artificial intelligence; Classification techniques; Psychiatry; Mental Health; Genes

Introduction

Schizophrenia is a complex mental health disorder that can have a significant negative impact on patients' resilience, quality of life and self-esteem [1]. Considering the heterogenous nature of schizophrenia, several fields of research such as genomics also use the terminology psychotic disorder spectrum to refer to schizophrenia-like disorders [2]. Furthermore, while untreated, this mental health condition can lead to violence and violent offending [3]. A recent review of the literature estimated that schizophrenia has the highest societal cost amongst all mental health diseases. Indeed, reports from 10 countries estimated schizophrenia-related costs per person per year to be around $2,004-94,229 US dollars, with considerable variability amongst countries [4]. Despite several treatments being available such as antipsychotics (dopamine receptor antagonists and partial agonists), up to 20-30% patients will remain treatment-resistant and further approaches such as cognitive behavioral therapy will be used as adjuncts [5-7]. Various studies have explored the diverging clinical presentations of schizophrenia patients and developed complexity estimators to aid

clinicians in understanding the neuropathological processes involved in this complex illness [8,9]. Amongst recent research, several key factors have been identified as being linked to the development of the disorder such as the length of the first psychotic episode, hormonal variations, as well as the presence of negative symptoms [10]. Despite the current knowledge that early interventions can help in the prognosis of patients diagnosed with schizophrenia, no prediction model is used in clinical practice as they usually do not account for variance between individuals [11].

To account for this variance and the dimensional aspects of schizophrenia, there has been tremendous efforts to gather genomic data and in-depth knowledge of neurobiological aspects of this disorder [12]. The entirety of the genetic information contained in an organism's deoxyribonucleic acid (DNA) is referred to as genomic data [13-15]. This comprises details on gene structure, function, and variation in addition to the nucleotide sequence (adenine, thymine, cytosine, and guanine) found in the genome [16]. Genomic data is used to research the genetic contributions to traits, diseases, and biological processes [17]. It includes a variety of genetic information such as single-nucleotide polymorphisms (SNPs), copy number variations (CNVs), and gene expression patterns [18]. Worldwide collaborations have resulted in genome-wide association studies (GWAS) in over 56,000 schizophrenia cases and 78,000 controls, which identified 270 distinct genetic loci and polygenic risk scores can currently explain around 7.7% of the variance in schizophrenia case-control status [19]. Despite over 300 studies on gene expression in schizophrenia conducted over the past 15 years, none has consistently identified specific genes that contribute to schizophrenia risk [20]. Due to the complexity of schizophrenia, novel approaches are essential to better understand its neurobiological basis and improve outcome predictions, as it involves a network of genetic, neural, behavioral, and environmental factors [21].

Amongst novel approaches, machine learning has been increasingly used in the latest decade for various applications in medicine [22]. Machine learning is a branch of artificial intelligence that deals with teaching computers how to learn from and make predictions or judgments based on data via the use of statistical models and algorithms [23,24]. It focuses on creating systems that, through experience, may naturally perform better on a given task without having to be specifically designed to do so [25]. Data used by machine learning algorithms are referred to as model features [26]. Recent advancements in the field of data science have demonstrated that precision and genomic medicine combined with artificial intelligence have the potential to improve patient healthcare [27]. Examples of such advancements are the possibility of conducting variant calling, genome annotation and variant classification, and phenotype-to-genotype correspondence by using machine learning algorithms [28]. While existing literature reviews have explored specific applications of machine learning using genomic data for schizophrenia, none, to our knowledge, have comprehensively examined the diverse uses of machine learning at the intersection of these three fields which could enhance the understanding of schizophrenia, thereby justifying the necessity for a thorough literature review. [29,30].

This study aims to identify the various applications of machine learning algorithms utilizing genomic data in the field of schizophrenia. By examining these approaches, this research offers an initial exploration into the methods being investigated to address the complexities of schizophrenia, a significant yet challenging mental illness. Therefore, this scoping review aimed to provide a comprehensive overview of these applications, highlighting key areas for future development at the intersection of machine learning, genomic data, and schizophrenia, with the potential to enhance clinical approaches.

Methods

Search strategies

A comprehensive scoping search was conducted to identify recent studies across several electronic databases, including Medline (PubMed), Web of Science, PsycNet (PsycINFO), and Google Scholar, covering the period from 2013 to 2024. The review was conducted using the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines for scoping reviews. The search strategy employed both text words and MeSH terms, focusing on schizophrenia (e.g., schizophrenia, schizophrenic), genomic data (e.g., genes, genetic, genomic), and machine learning (e.g., artificial intelligence, machine learning). These topics were selected to align with the study's objectives. Detailed search strategies are provided in Multimedia Appendix 1. The search methodology was developed by the corresponding authors, with searches executed by AH and cross-validated by MB. No restrictions were applied regarding setting, or geography.

Study Eligibility

Studies were included based on the following criteria: (1) the population of interest consisted of patients diagnosed with schizophrenia or the study of schizophrenia; (2) the study employed a machine learning approach; and (3) the machine learning model incorporated genomic data features to find specific outcomes. Studies were included regardless of if they used a single algorithm or multiple algorithms. Excluded from consideration were unpublished literature and studies using artificial intelligence algorithms outside the scope of machine learning. Studies that used machine learning solely to reduce data from genomic datasets were excluded. The search was limited to sources in English and French.

Data Extraction

Data extraction was performed using a standardized form in Microsoft Excel (version Microsoft 365, EULA license, USA) and was independently counter verified for consistency and integrity by two authors (AH, MB). Any disagreements regarding the inclusion or exclusion of a study were mutually resolved by the authors. The systematically extracted information included: authors, population (sample), primary uses (or intent) of the machine learning algorithm(s), types of genomic data, type(s) of machine learning algorithm used, main model performances, and key outcomes identified.

Quality Assessment

The quality of the identified studies was evaluated using the Newcastle-Ottawa Scale for non-randomized controlled studies and the Cochrane Risk of Bias Tool for randomized controlled trials [31,32]. The Newcastle-Ottawa Scale is a tool used for assessing the quality of cohort and case-control studies. It evaluates studies based on three main domains: selection of study groups, comparability of groups, and ascertainment of exposure or outcome [31]. Each domain includes specific criteria, and studies are awarded stars for meeting these criteria, with a maximum of nine stars indicating the highest quality [31]. The Cochrane Risk of Bias Tool is a comprehensive framework used to assess the risk of bias in randomized controlled trials [32]. It evaluates seven specific domains: random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting, and other

potential sources of bias [32]. Each domain is rated as having a low, high, or unclear risk of bias based on predefined criteria [32]. In this scoping review, studies with 1-4 stars in the Newcastle-Ottawa Scale or a high risk of bias by the Cochrane Risk of Bias Tool will be identified as low in quality, 4-6 stars as moderate and 7-9 stars (or low risk of bias) as high.

Results

Description of studies

The scoping review evaluated studies at the intersection of schizophrenia, genomic data, and machine learning. Initially, the literature search identified 2437 eligible articles after removing duplicates. A total of 814 studies were excluded based on a first analysis of the titles and abstract. Following a second round of abstract screening, 143 full-text articles were thoroughly assessed, with 122 subsequently excluded. This left 21 studies for detailed analysis. A flowchart illustrating the inclusion process is provided in Figure 1, and the specific details of the included studies are available in Multimedia Appendix 2. The studies meeting the inclusion criteria included various algorithms for different tasks. The most common application of machine learning was predicting schizophrenia using genomic data (n=10), followed by identifying features to enhance the understanding of schizophrenia (n=6), drug discovery for schizophrenia patients (n=2), classifying schizophrenia amongst other mental health disorders (n=2), and predicting the quality of life and global functioning of schizophrenia patients (n=1).

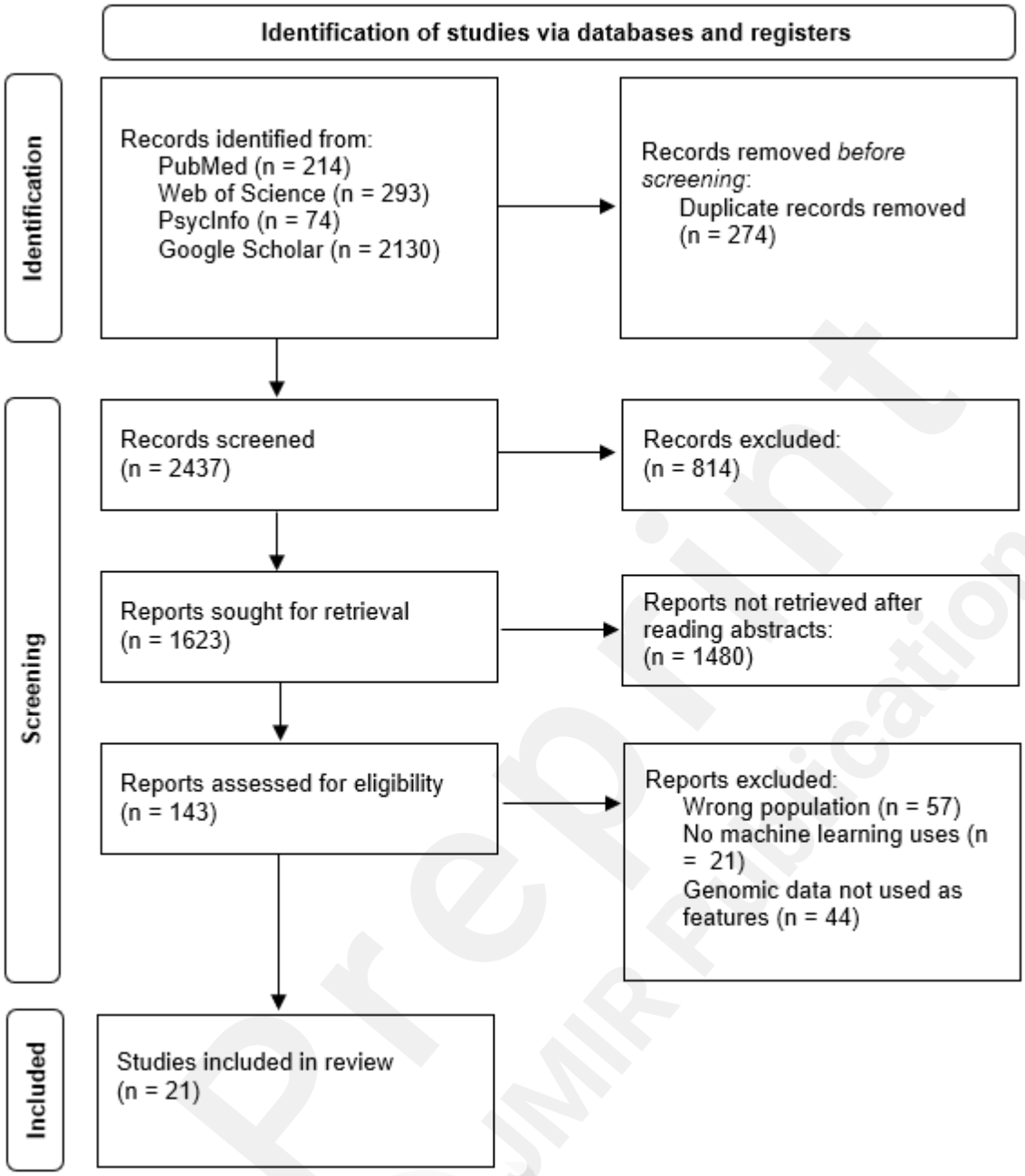


Figure 1. PRISMA flowchart for the inclusion of studies.

Algorithms Used

Several algorithms have been identified in the 21 included studies. The algorithms the most frequently used were support vector machine classifiers (SVM) (n = 14), random forest (RF) algorithms (n = 9), various implementations of neural networks (NN) (n=7) and eXtreme Gradient Boosting (XGboost) (n = 5). Definitions of these popular algorithms are listed below:

- RF: It constitutes an ensemble learning technique. During training, it creates several decision trees and outputs the class that is the average of the classes of each individual tree [33]. By merging the predictions of several trees, each trained on a different sample of the data, this

method increases accuracy and helps avoid overfitting [33].

- SVM: It is an algorithm for supervised machine learning that is applied to regression and classification problems [34]. Finding the ideal hyperplane to divide the data into distinct classes is the fundamental notion behind SVM [34]. Different kernels (function that quantifies similarity between a pair of data points) can be used to enhance the performance of the SVM to best-fit the data points [35].
- NN: These algorithms are modeled after the composition and operations of the human brain [36]. They are made up of networked layers of nodes, also called neurons, that process and change incoming data to create outputs [36].
- XGboost: It is founded on the gradient boosting principle, which entails building an ensemble of weak learners (usually decision trees) in a stepwise manner [37]. Every new tree seeks to fix the mistakes committed by the ones that came before it [37].

The remaining algorithms can be found in Multimedia Appendix 2.

Predicting Schizophrenia

Prediction of schizophrenia was identified as the main objective of 10 studies, all of which were deemed of high quality as per the Newcastle-Ottawa Scale ratings. The data used in these studies included: differentially expressed genes, polygenic risk scores, genotype and human leukocyte antigen alleles, gene expression microarray data, single nucleotide polymorphisms, long non-coding ribonucleic acids (RNAs), DNA methylation in blood, exomes, and G72 protein levels.

Li and colleagues utilized differentially expressed genes data from the Gene Expression Omnibus database, applying RF and SVM algorithms, and identified 15 key genes correlated with immune cell infiltration, achieving high diagnostic accuracy for schizophrenia with an area under the curve (AUC) of 0.77 in their test set [38]. Another study, by Bracher-Smith and colleagues, used data from the UK Biobank, applied machine learning algorithms such as least absolute shrinkage and selection operator, ridge-penalized logistic regression, SVM, RF, XGboost, NN, and stacked models, and found that while machine learning models incorporating polygenic risk scores and demographic factors showed good discrimination (AUC = 0.71), they did not significantly outperform logistic regression in predicting schizophrenia [39]. However, they reported that permutation features importance identified PRS-SZ as the most important predictor of schizophrenia [39].

Using data from the iPSYCH2012 case cohort, another study integrated genetics and registry data with a deep learning approach to stratify 19,636 patients with schizophrenia and/or major depressive disorder into clinically distinct subgroups, characterized by unique disorder severities and comorbidity signatures, with predictive models achieving AUCs of 0.55 to 0.97, and therefore emphasized the importance of data-driven stratification for improving psychiatric diagnosis and prognosis [40]. Similarly, Qi and colleagues analyzed gene expression datasets from untreated schizophrenia patients and controls, identified 14 key gene probes, and used artificial NN to achieve diagnostic accuracy of 91.2% in training and 87.9% in testing and highlighted the potential of machine learning in identifying clinically useful biomarkers for schizophrenia [41]. Another study introduced a sparse deep NN approach for identifying interpretable features for schizophrenia case–control classification using gray matter volume and single nucleotide polymorphism data, demonstrating slightly improved performance over traditional methods, and highlighting key brain regions related to the schizophrenia [42].

Studies with smaller sample sizes also reported several genomic data enhanced methodologies to predict schizophrenia. Zhu and colleagues demonstrated that a machine learning model using the

expression levels of six genes (GNAI1, FYN, PRKCA, YWHAZ, PRKCB, and LYN) in peripheral blood effectively distinguishes schizophrenia patients from healthy controls, with the SVM model achieving the highest accuracy (AUC = 0.993) [43]. Another study also reported the importance of long non-coding RNAs as they provided higher accuracy than coding genes in distinguishing schizophrenia from healthy controls [44].

Also focusing on predicting schizophrenia, a machine learning classifier based on DNA methylation in blood, specifically using CoRSIV regions and SPLS-DA, effectively distinguishes schizophrenia patients from controls with a highly positive predictive value (PPV) of 80%, outperforming models based on polygenic risk scores (PRS) [45]. Another machine learning implementation used whole exome sequencing data to identify individuals at high risk for schizophrenia, achieving the accuracy of 85.7% with the XGBoost algorithm and providing further insights into the genetic basis of the disorder [46]. Finally, the last identified study used machine learning algorithms to demonstrate that G72 protein levels alone, without incorporating G72 genetic variations, are effective in distinguishing schizophrenia patients from healthy controls with high specificity (0.9503) and sensitivity (0.8765) [47].

Identifying Features of Schizophrenia

A total of x included studies aimed at identifying features of schizophrenia or phenotyping using machine learning and genomic data, all of which were assessed as being of high quality. Feng and colleagues identified six candidate genes (SFN, KDM5B, MYLK, IRF3, IRF7, and ID1) with diagnostic significance for schizophrenia using machine learning on gene expression data. [48]. Another study by Zhu and colleagues attempted to identify immune-related biomarkers in peripheral blood in patients diagnosed with schizophrenia and reported that the messenger RNA (mRNA) expression of CLIC3 was significantly decreased in the schizophrenia samples compared to the healthy controls [49]. By employing machine learning methods to analyze RNA sequencing data from the dorsolateral prefrontal cortex and amygdala in a postmortem investigation, Liu and colleagues aimed to identify driving biological signals representing schizophrenia. In doing so, they identified 18 genes added to known schizophrenia-associated pathways and expanded the gene network. These results provide a more comprehensive understanding of schizophrenia pathogenesis [50].

De Rosa and colleagues identified biological signals representing schizophrenia in brain tissues of dorsolateral prefrontal cortex and hippocampus samples from post-mortem brains of non-psychiatric controls and schizophrenia patients [51]. Using a RF approach, they found 103 additional gene interactions were expanded to schizophrenia-associated networks, which were shared amongst both the dorsolateral prefrontal cortex and amygdala regions [51]. Another study by Feng and Shen used neural networks using programmed cell-death-related genes as features and found 10 candidate hub genes (DPF2, ATG7, GSK3A, TFDP2, ACVR1, CX3CR1, AP4M1, DEPDC5, NR4A2, and IKBKB) [52]. Finally, a study on fresh frozen post-mortem brain tissue aimed to identify DNA methylation patterns specific to schizophrenia patients.

A cohort of 73 subjects diagnosed with schizophrenia and 52 control samples was analyzed using an unsupervised machine learning approach achieving. As results were not convincing, the authors reported that, if there are methylation changes associated with schizophrenia, they are diverse, complex, and have a small effect size [53].

Drug Discovery

Two studies reported the use of machine learning specifically for drug discovery (or related issues) for patients diagnosed with schizophrenia. Both or them were deemed of high quality. A first study focusing on 2307 patients with schizophrenia from the Chinese Antipsychotics Pharmacogenomics Consortium, 1379 from the Chinese Antipsychotics Pharmacogenetics Consortium, 275 healthy controls used several SVM and RF implementations and identified six risk genes for schizophrenia (LINC01795, DDHD2, SBNO1, KCNG2, SEMA7A, and RUFY1), which are involved in cortical morphology and were identified as having genetic-epigenetic interactions linked to treatment response [54]. The other study, by Zhao and So, used the expression database ConnectivityMap that contains transcriptomic changes for HL60, PC3, MCF over several machine learning implementations and reported that the predictive performance of their five approaches in cross validation did not differ substantially, with SVM slightly outperforming the others while stating that repositioning hits are enriched for psychiatric medications considered in clinical trials [55].

Classifying schizophrenia amongst other mental health disorders

Two studies aiming at classifying schizophrenia amongst other mental disorders using machine learning were identified.

The first study by Yang and colleagues aimed at distinguishing schizophrenia from individuals with bipolar disorder, major depressive disorders, and healthy controls. To do so, the authors used differentially expressed genes from 268 individuals (67 patients with schizophrenia, 40 patients with bipolar disorder, 57 patients with major depressive disorders, and 104 healthy controls) over an SVM implementation that achieved an AUC of 0.96 for the schizophrenia group and of 0.71 for the independent set of the classification model. They reported that their model has a strong capacity to classify samples amongst multiple groups of mental illnesses [56]. Considering the opacity of the implementation, the quality was assessed as moderate for this study.

The other study, by Saardar and colleagues, used the dbGaP database (schizophrenia) and the NDAR database (autism spectrum disorder) to compare whole exomes to differentiate between schizophrenia and autism using an XGboost model. They achieved an average validation accuracy over five folds was 88% for both the singe nucleotide variants-based model and gene-based model and reported that the ion transmembrane transport, neurotransmitter transport, and microtubule/cytoskeleton processes were of importance for schizophrenia [57]. The quality of this study was determined to be high as per our assessment.

Predicting Quality-of-Life and Global Functioning

Only one of the included studies was focusing on predicting quality of life and global functioning of patients diagnosed with schizophrenia. This study was of high quality as per quality assessment. Using data from 302 schizophrenia patients in the Taiwanese population, Lin and colleagues compared a bagged ensemble of several machine learning algorithms to different permutations of these algorithms to predict functional outcomes of patients with schizophrenia [58]. Their analysis revealed that the bagging ensemble algorithm with feature selection outperformed other predictive algorithms in forecasting the quality-of-life functional outcome of schizophrenia using the G72 rs2391191 and MET rs2237717 SNPs [58].

Discussion

Principal Results

This scoping review aimed to identify the different ways machine learning algorithms can be applied to genomic data in the study of schizophrenia. A total of 21 studies were fully analyzed and five uses of machine learning algorithms on genomic data were identified: predicting schizophrenia, identifying features of schizophrenia, drug discovery, classifying schizophrenia amongst other mental health disorders and predicting quality-of-life and global functioning. The studies were overall of high quality.

Comparison With Prior Work

The application of predictive models to forecast mental health disorders, such as schizophrenia, is gaining importance in medical research [59]. These models hold potential to significantly assist clinicians in patient evaluation, particularly given the heterogeneity inherent to schizophrenia [60]. However, as observed in the identified studies, these models vary greatly in their implementation with diverging accuracy and validation methodologies. It is important to consider the implementation of these models as well as their accuracy and the techniques used to cross-validate the model, especially when using genomic data as this could hinder their external validity [61]. The results found in the identified studies reinforce the premise that genetic architecture of schizophrenia has proven to be very complex, heterogeneous, and polygenic and that a vast array of features could be integrated to improve predictive models [62]. Similarly, finding genomic related risk factors of schizophrenia in such model, could help in distinguishing between this disease and other mental disorders which may explain why classifying schizophrenia amongst other mental health disorders was one of the identified uses.

It is unsurprising that machine learning has been used to identify features of schizophrenia as this has been done in other medical fields. Using candidate genes, it can be possible for clinicians to better understand common diseases and complex traits [63]. In psychiatry, psychiatric genomics is a rapidly advancing field that shows great promise for enhancing risk prediction, prevention, diagnosis, treatment selection, and the understanding of the pathogenesis of patients' symptoms [64]. As an example, some genes and functional genomic data linked to complex features of schizophrenia demonstrated that specific alleles may confer risk to the disorder by directly affecting synaptic function in adulthood [65].

As for drug discovery, literature reviews on the subject support that machine learning techniques can improve the decision-making in pharmaceutical data across various applications [66, 67]. It is also reported that combining machine learning techniques to genomic data has the potential to speed up the process and reduce failure rates in drug discovery and development [67]. This may explain why two studies focused specifically on schizophrenia in the context of drug discovery were identified. There is an increasing effort to develop pharmaceutical treatments, given the 20-30% rate of treatment resistance observed in patients with this disorder [4].

Finally, quality-of-life assessment and functioning of patients suffering from schizophrenia is trending in this field, which may explain why this use was identified in one study [68, 69]. Another recent study on quality of life and genome-wide analyses of quality of life in psychosis, which used linear regression on 3,684 participants (including 1,119 psychosis patients), reported that numerous clinical and genetic associations with quality-of-life can be utilized in the daily care of these patients and enhance their overall well-being. These findings support the idea that there should be more work conducted in this area in the future [70].

Limitations of the Present Study

This scoping review highlighted the various applications of machine learning algorithms utilizing genomic data in the field of schizophrenia. Despite the relevance of this recension, it has a few limitations. The heterogeneity of diagnostic criteria for schizophrenia is a significant concern, as it is not addressed in half of the studies reviewed. Furthermore, the limited number of studies identified indicates the novelty of this field, necessitating future reviews to confirm findings. There is also a lack of external validation in samples differing from the training sample, such as those from different nationalities, raising questions about the generalizability of the results. Notably, no studies have concretely tested these algorithms in clinical settings, particularly for the prediction of schizophrenia, which remains an unmet need in the research. Due to the heterogeneity of the identified studies and the varying metrics used to assess precision and validate the machine learning models, performance comparisons were not conducted. Furthermore, studies on generic models using genomic data to predict overall mental health, rather than specifically focusing on schizophrenia, were excluded. This may have led to the omission of a small portion of relevant studies.

Conclusions

Considering the heterogeneity of clinical presentations observed in schizophrenia, genomic data combined to machine learning algorithms have been implemented to address several facets of this disorder. From the 21 studies analyzed, five main uses were identified: predicting schizophrenia, identifying schizophrenia features, discovering drugs, classifying schizophrenia amongst other mental health disorders, and predicting the quality-of-life of patients. These uses have potential implications as they could assist clinicians in providing a more personalized approach with their patients diagnosed with schizophrenia considering the complexity of this diagnosis. There is still a limited amount of literature on the subject, and this study provides a first overview of machine learning applications of genomic data for schizophrenia. Future research is essential to further evaluate the portability of the models identified and their potential clinical applications.

Conflicts of Interest

None declared

Abbreviations

CNVs: Copy Number Variations
DNA: Deoxyribonucleic Acid
GWAS: Genome-Wide Association Studies
MeSH: Medical Subject Headings
mRNA: messenger ribonucleic acid
NN: Neural networks
RF: Random forest
RNAs: ribonucleic acids
SNPs: Single Nucleotide Polymorphisms
SVM: Support vector machine
US: United States
XGboost: eXtreme Gradient Boosting

Multimedia Appendix 1

Multimedia Appendix 1. Electronic search strategy for the scoping review conducted.
Multimedia Appendix 2. Systematic review study selection detailed results.

References

1. Wartelsteiner F, Mizuno Y, Frajo-Apor B, et al. Quality of life in stabilized patients with schizophrenia is mainly associated with resilience and self-esteem. Acta Psychiatr Scand. 2016;134(4):360-367. doi:10.1111/acps.12628
2. Cuthbert BN, Morris SE. Evolving Concepts of the Schizophrenia Spectrum: A Research Domain Criteria Perspective. Front Psychiatry. 2021;12:641319. Published 2021 Feb 25. doi:10.3389/fpsyt.2021.641319
3. Fazel S, Gulati G, Linsell L, Geddes JR, Grann M. Schizophrenia and violence: systematic review and meta-analysis. PLoS Med. 2009;6(8):e1000120. doi:10.1371/journal.pmed.1000120
4. Kotzeva A, Mittal D, Desai S, Judge D, Samanta K. Socioeconomic burden of schizophrenia: a targeted literature review of types of costs and associated drivers across 10 countries. J Med Econ. 2023;26(1):70-83. doi:10.1080/13696998.2022.2157596
5. Efthimiou O, Taipale H, Radua J, et al. Efficacy and effectiveness of antipsychotics in schizophrenia: network meta-analyses combining evidence from randomised controlled trials and real-world data. Lancet Psychiatry. 2024;11(2):102-111. doi:10.1016/S2215-0366(23)00366-8
6. Bighelli I, Çıray O, Salahuddin NH, Leucht S. Cognitive behavioural therapy without medication for schizophrenia. Cochrane Database Syst Rev. 2024;2(2):CD015332. Published 2024 Feb 7. doi:10.1002/14651858.CD015332.pub2
7. Remington G, Hahn MK, Agarwal SM, Chintoh A, Agid O. Schizophrenia: Antipsychotics and drug development. Behav Brain Res. 2021;414:113507. doi:10.1016/j.bbr.2021.113507
8. Fernández A, Gómez C, Hornero R, López-Ibor JJ. Complexity and schizophrenia. Prog Neuropsychopharmacol Biol Psychiatry. 2013;45:267-276. doi:10.1016/j.pnpbp.2012.03.015
9. Bassett DS, Nelson BG, Mueller BA, Camchong J, Lim KO. Altered resting state complexity in schizophrenia. Neuroimage. 2012;59(3):2196-2207. doi:10.1016/j.neuroimage.2011.10.002
10. Häfner H. From Onset and Prodromal Stage to a Life-Long Course of Schizophrenia and Its

Symptom Dimensions: How Sex, Age, and Other Risk Factors Influence Incidence and Course of Illness. Psychiatry J. 2019;2019:9804836. Published 2019 Apr 16. doi:10.1155/2019/9804836
11. Lee R, Leighton SP, Thomas L, et al. Prediction models in first-episode psychosis: systematic review and critical appraisal. Br J Psychiatry. Published online January 24, 2022. doi:10.1192/bjp.2021.219
12. Owen MJ. Genomic insights into schizophrenia. R Soc Open Sci. 2023;10(2):230125. Published 2023 Feb 22. doi:10.1098/rsos.230125
13. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. Nature. 2024;627(8003):340-346. doi:10.1038/s41586-023-06957-x
14. Middleton A, Milne R, Almarri MA, et al. Global Public Perceptions of Genomic Data Sharing: What Shapes the Willingness to Donate DNA and Health Data?. Am J Hum Genet. 2020;107(4):743-752. doi:10.1016/j.ajhg.2020.08.023
15. Tatusova T. Update on Genomic Databases and Resources at the National Center for Biotechnology Information. Methods Mol Biol. 2016;1415:3-30. doi:10.1007/978-1-4939-3572-7_1
16. Eisenberg L. Are genes destiny? Have adenine, cytosine, guanine and thymine replaced Lachesis, Clotho and Atropos as the weavers of our fate?. World Psychiatry. 2005;4(1):3-8.
17. Daniels H, Jones KH, Heys S, Ford DV. Exploring the Use of Genomic and Routinely Collected Data: Narrative Literature Review and Interview Study. J Med Internet Res. 2021;23(9):e15739. Published 2021 Sep 24. doi:10.2196/15739
18. Liu J, Zhou Y, Liu S, et al. The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. Hum Genet. 2018;137(6-7):553-567. doi:10.1007/s00439-018-1910-3
19. Legge SE, Santoro ML, Periyasamy S, Okewole A, Arsalan A, Kowalec K. Genetic architecture of schizophrenia: a review of major advancements. Psychol Med. 2021;51(13):2168-2177. doi:10.1017/S0033291720005334
20. Merikangas AK, Shelly M, Knighton A, Kotler N, Tanenbaum N, Almasy L. What genes are differentially expressed in individuals with schizophrenia? A systematic review. Mol Psychiatry. 2022;27(3):1373-1383. doi:10.1038/s41380-021-01420-7
21. Haller CS, Padmanabhan JL, Lizano P, Torous J, Keshavan M. Recent advances in understanding schizophrenia. F1000Prime Rep. 2014;6:57. Published 2014 Jul 8. doi:10.12703/P6-57
22. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. Artificial Intelligence in Healthcare. 2020;25-60. doi:10.1016/B978-0-12-818438-7.00002-2
23. Lepakshi VA. Machine Learning and Deep Learning based AI Tools for Development of Diagnostic Tools. Computational Approaches for Novel Therapeutic and Diagnostic Designing to Mitigate SARS-CoV-2 Infection. 2022;399-420. doi:10.1016/B978-0-323-91172-6.00011-X
24. Lovis C. Unlocking the Power of Artificial Intelligence and Big Data in Medicine. J Med Internet Res. 2019;21(11):e16607. Published 2019 Nov 8. doi:10.2196/16607
25. Sarker IH. AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. SN Comput Sci. 2022;3(2):158. doi:10.1007/s42979-022-01043-x
26. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Front Bioinform. 2022;2:927312. Published 2022 Jun 27. doi:10.3389/fbinf.2022.927312
27. Quazi S. Artificial intelligence and machine learning in precision and genomic medicine. Med Oncol. 2022;39(8):120. Published 2022 Jun 15. doi:10.1007/s12032-022-01711-1

28. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med. 2019;11(1):70. Published 2019 Nov 19. doi:10.1186/s13073-019-0689-8
29. Bracher-Smith M, Crawford K, Escott-Price V. Machine learning for genetic prediction of psychiatric disorders: a systematic review. Mol Psychiatry. 2021;26(1):70-79. doi:10.1038/s41380-020-0825-2
30. Del Fabro L, Bondi E, Serio F, Maggioni E, D'Agostino A, Brambilla P. Machine learning methods to predict outcomes of pharmacological treatment in psychosis. Transl Psychiatry. 2023;13(1):75. Published 2023 Mar 2. doi:10.1038/s41398-023-02371-z
31. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol. 2010;25(9):603-605. doi:10.1007/s10654-010-9491-z
32. Higgins JP, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928. Published 2011 Oct 18. doi:10.1136/bmj.d5928
33. Rigatti SJ. Random Forest. J Insur Med. 2017;47(1):31-39. doi:10.17849/insm-47-01-31-39.1
34. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics. 2018;15(1):41-51. doi:10.21873/cgp.20063
35. Noble WS. What is a support vector machine?. Nat Biotechnol. 2006;24(12):1565-1567. doi:10.1038/nbt1206-1565
36. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. Transl Vis Sci Technol. 2020;9(2):14. Published 2020 Feb 27. doi:10.1167/tvst.9.2.14
37. Moore A, Bell M. XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study. Clin Med Insights Cardiol. 2022;16:11795468221133611. Published 2022 Nov 8. doi:10.1177/11795468221133611
38. Li Z, Li X, Jin M, et al. Identification of potential biomarkers and their correlation with immune infiltration cells in schizophrenia using combinative bioinformatics strategy. Psychiatry Res. 2022;314:114658. doi:10.1016/j.psychres.2022.114658
39. Bracher-Smith M, Rees E, Menzies G, et al. Machine learning for prediction of schizophrenia using genetic and demographic factors in the UK biobank. Schizophr Res. 2022;246:156-164. doi:10.1016/j.schres.2022.06.006
40. Allesøe RL, Nudel R, Thompson WK, et al. Deep learning-based integration of genetics with registry data for stratification of schizophrenia and depression. Sci Adv. 2022;8(26):eabi7293. doi:10.1126/sciadv.abi7293
41. Qi B, Boscenco S, Ramamurthy J, Trakadis YJ. Transcriptomics and machine learning to advance schizophrenia genetics: A case-control study using post-mortem brain data. Comput Methods Programs Biomed. 2022;214:106590. doi:10.1016/j.cmpb.2021.106590
42. Chen J, Li X, Calhoun VD, et al. Sparse deep neural networks on imaging genetics for schizophrenia case-control classification. Hum Brain Mapp. 2021;42(8):2556-2568. doi:10.1002/hbm.25387
43. Zhu L, Wu X, Xu B, et al. The machine learning algorithm for the diagnosis of schizophrenia on the basis of gene expression in peripheral blood. Neurosci Lett. 2021;745:135596. doi:10.1016/j.neulet.2020.135596
44. Liu Y, Qu HQ, Chang X, et al. Machine Learning Reduced Gene/Non-Coding RNA Features That Classify Schizophrenia Patients Accurately and Highlight Insightful Gene Clusters. Int J Mol Sci. 2021;22(7):3364. Published 2021 Mar 25. doi:10.3390/ijms22073364
45. Gunasekara CJ, Hannon E, MacKay H, et al. A machine learning case-control classifier for schizophrenia based on DNA methylation in blood. Transl Psychiatry. 2021;11(1):412.

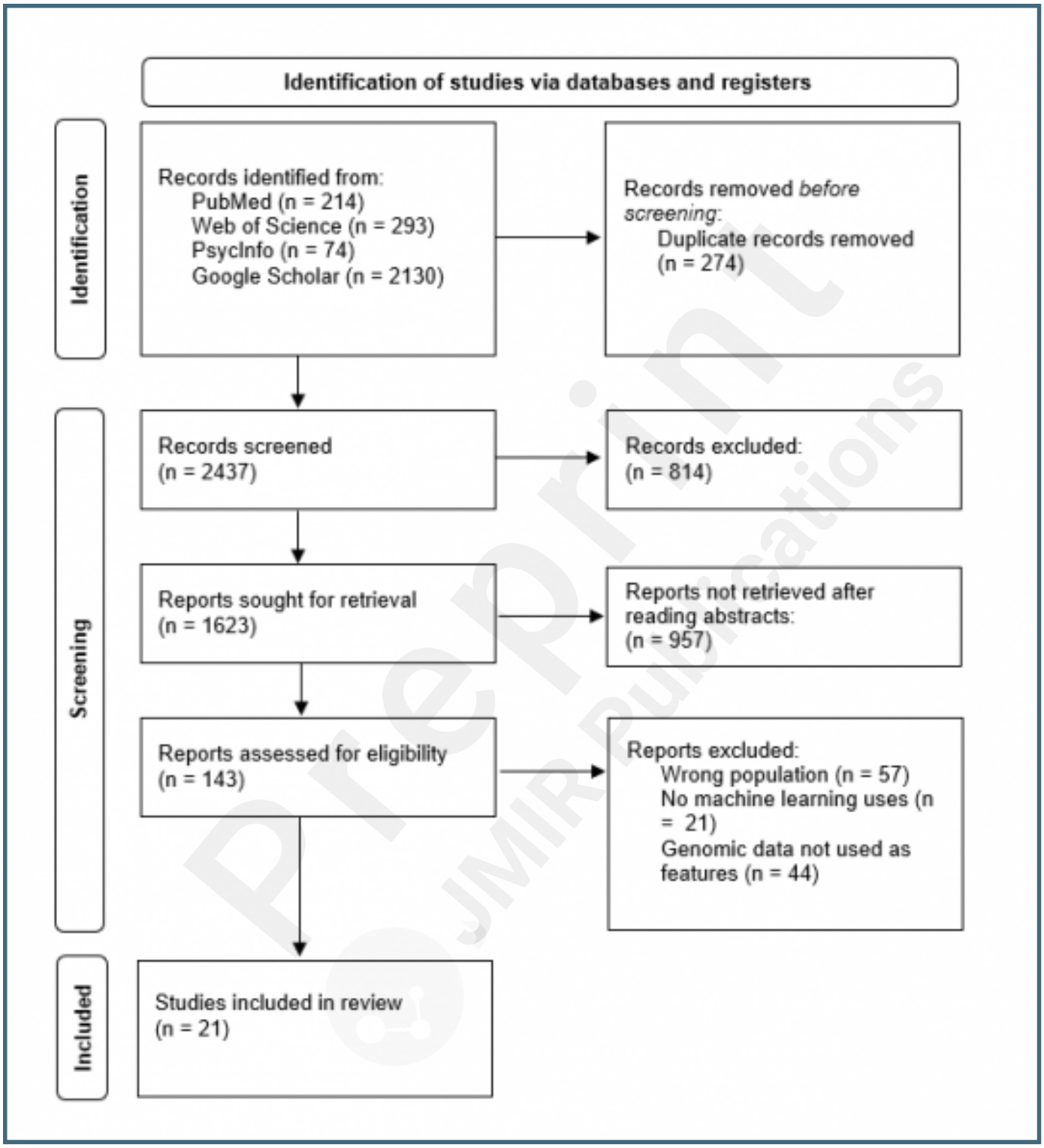
Published 2021 Aug 3. doi:10.1038/s41398-021-01496-3
46. Trakadis YJ, Sardaar S, Chen A, Fulginiti V, Krishnan A. Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes. Am J Med Genet B Neuropsychiatr Genet. 2019;180(2):103-112. doi:10.1002/ajmg.b.32638
47. Lin E, Lin CH, Lai YL, Huang CH, Huang YJ, Lane HY. Combination of G72 Genetic Variation and G72 Protein Level to Detect Schizophrenia: Machine Learning Approaches. Front Psychiatry. 2018;9:566. Published 2018 Nov 6. doi:10.3389/fpsyt.2018.00566
48. Feng Y, Shen J, He J, Lu M. Schizophrenia and cell senescence candidate genes screening, machine learning, diagnostic models, and drug prediction. Front Psychiatry. 2023;14:1105987. Published 2023 Apr 11. doi:10.3389/fpsyt.2023.1105987
49. Zhu X, Wang CL, Yu JF, et al. Identification of immune-related biomarkers in peripheral blood of schizophrenia using bioinformatic methods and machine learning algorithms. Front Cell Neurosci. 2023;17:1256184. Published 2023 Sep 28. doi:10.3389/fncel.2023.1256184
50. Liu Y, Qu HQ, Chang X, et al. Expansion of Schizophrenia Gene Network Knowledge Using Machine Learning Selected Signals From Dorsolateral Prefrontal Cortex and Amygdala RNA-seq Data. Front Psychiatry. 2022;13:797329. Published 2022 Mar 21. doi:10.3389/fpsyt.2022.797329
51. De Rosa A, Fontana A, Nuzzo T, et al. Machine Learning algorithm unveils glutamatergic alterations in the post-mortem schizophrenia brain. Schizophrenia (Heidelb). 2022;8(1):8. Published 2022 Feb 25. doi:10.1038/s41537-022-00231-1
52. Feng Y, Shen J. Machine learning-based predictive models and drug prediction for schizophrenia in multiple programmed cell death patterns. Front Mol Neurosci. 2023;16:1123708. Published 2023 Mar 13. doi:10.3389/fnmol.2023.1123708
53. Torabi Moghadam B, Etemadikhah M, Rajkowska G, et al. Analyzing DNA methylation patterns in subjects diagnosed with schizophrenia using machine learning methods. J Psychiatr Res. 2019;114:41-47. doi:10.1016/j.jpsychires.2019.04.001
54. Guo LK, Su Y, Zhang YY, et al. Prediction of treatment response to antipsychotic drugs for precision medicine approach to schizophrenia: randomized trials and multiomics analysis. Mil Med Res. 2023;10(1):24. Published 2023 Jun 2. doi:10.1186/s40779-023-00459-7
55. Zhao K, So HC. Drug Repositioning for Schizophrenia and Depression/Anxiety Disorders: A Machine Learning Approach Leveraging Expression Data. IEEE J Biomed Health Inform. 2019;23(3):1304-1315. doi:10.1109/JBHI.2018.2856535
56. Yang Q, Xing Q, Yang Q, Gong Y. Classification for psychiatric disorders including schizophrenia, bipolar disorder, and major depressive disorder using machine learning. Comput Struct Biotechnol J. 2022;20:5054-5064. Published 2022 Sep 12. doi:10.1016/j.csbj.2022.09.014
57. Sardaar S, Qi B, Dionne-Laporte A, Rouleau GA, Rabbany R, Trakadis YJ. Machine learning analysis of exome trios to contrast the genomic architecture of autism and schizophrenia. BMC Psychiatry. 2020;20(1):92. Published 2020 Feb 28. doi:10.1186/s12888-020-02503-5
58. Lin E, Lin CH, Lane HY. Prediction of functional outcomes of schizophrenia with genetic biomarkers using a bagging ensemble machine learning method with feature selection. Sci Rep. 2021;11(1):10179. Published 2021 May 13. doi:10.1038/s41598-021-89540-6
59. Montazeri M, Montazeri M, Bahaadinbeigy K, Montazeri M, Afraz A. Application of machine learning methods in predicting schizophrenia and bipolar disorders: A systematic review. Health Sci Rep. 2022;6(1):e962. Published 2022 Dec 28. doi:10.1002/hsr2.962
60. Gashkarimov VR, Sultanova RI, Efremov IS, Asadullin AR. Machine learning techniques in diagnostics and prediction of the clinical features of schizophrenia: a narrative review. Consort Psychiatr. 2023;4(3):43-53. Published 2023 Sep 29. doi:10.17816/CP11030
61. McGaugh SE, Lorenz AJ, Flagel LE. The utility of genomic prediction models in

evolutionary genetics. Proc Biol Sci. 2021;288(1956):20210693. doi:10.1098/rspb.2021.0693
62. Henriksen MG, Nordgaard J, Jansson LB. Genetics of Schizophrenia: Overview of Methods, Findings and Limitations. Front Hum Neurosci. 2017;11:322. Published 2017 Jun 22. doi:10.3389/fnhum.2017.00322
63. Hirschhorn JN. Genetic approaches to studying common diseases and complex traits. Pediatr Res. 2005;57(5 Pt 2):74R-77R. doi:10.1203/01.PDR.0000159574.98964.87
64. Ward ET, Kostick KM, Lázaro-Muñoz G. Integrating Genomics into Psychiatric Practice: Ethical and Legal Challenges for Clinicians. Harv Rev Psychiatry. 2019;27(1):53-64. doi:10.1097/HRP.0000000000000203
65. Hall J, Bray NJ. Schizophrenia Genomics: Convergence on Synaptic Development, Adult Synaptic Plasticity, or Both?. Biol Psychiatry. 2022;91(8):709-717. doi:10.1016/j.biopsych.2021.10.018
66. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine Learning in Drug Discovery: A Review. Artif Intell Rev. 2022;55(3):1947-1999. doi:10.1007/s10462-021-10058-4
67. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-477. doi:10.1038/s41573-019-0024-5
68. de Souza LA, Coutinho ES. The quality of life of people with schizophrenia living in community in Rio de Janeiro, Brazil. Soc Psychiatry Psychiatr Epidemiol. 2006;41(5):347-356. doi:10.1007/s00127-006-0042-6
69. Beaudoin M, Hudon A, Giguère CE, Potvin S, Dumais A. Prediction of quality of life in schizophrenia using machine learning models on data from Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) schizophrenia trial. Schizophrenia (Heidelb). 2022;8(1):29. Published 2022 Mar 21. doi:10.1038/s41537-022-00236-w
70. Pazoki R, Lin BD, van Eijk KR, et al. Phenome-wide and genome-wide analyses of quality of life in schizophrenia [published correction appears in BJPsych Open. 2021 Mar 26;7(2):e73]. BJPsych Open. 2020;7(1):e13. Published 2020 Dec 9. doi:10.1192/bjo.2020.140

Supplementary Files

Figures

PRISMA flowchart for the inclusion of studies.



Multimedia Appendixes

Electronic search strategy for the scoping review conducted.
URL: http://asset.jmir.pub/assets/771cb828322593c55b856628ba9a1c25.docx

Systematic review study selection detailed results.
URL: http://asset.jmir.pub/assets/f16d2b35625a03574de9cb693b678666.docx

CONSORT (or other) checklists

Prisma for scoping review checklist.
URL: http://asset.jmir.pub/assets/4518b8ae082ff9f7238578d357ce15aa.pdf