UNIVERSIDADE DE LISBOA

Faculdade de Medicina

Onset probability prediction of schizophrenia based on a multimodal approach

Vânia Sofia Santos Tavares

Orientadores: Prof. Doutora Diana Maria Pinto Prata
Prof. Doutor Hugo Alexandre Teixeira Duarte Ferreira

Tese especialmente elaborada para obtenção do grau de Doutor em Ciências Biomédicas (Neurociências)

2022

UNIVERSIDADE DE LISBOA

Faculdade de Medicina

Onset probability prediction of schizophrenia based on a multimodal approach

Vânia Sofia Santos Tavares

Orientadores: Prof. Doutora Diana Maria Pinto Prata
Prof. Doutor Hugo Alexandre Teixeira Duarte Ferreira

Tese especialmente elaborada para obtenção do grau de Doutor em Ciências Biomédicas (Neurociências)

Júri:
Presidente: Doutora Helena Maria Ramos Marques Coelho Cortez Pinto, Professora Catedrática e Presidente do Conselho Científico da Faculdade de Medicina da Universidade de Lisboa
Vogais:
- Doutora Sandra Marisa Mendes Vieira, Postdoctoral Fellow do King's College London (United Kingdom);
- Doutor Miguel Sá Sousa Castelo Branco, Professor Catedrático da Faculdade de Medicina da Universidade de Coimbra;
- Doutora Diana Maria Pinto Prata, Professora Auxiliar Convidada da Faculdade de Ciências da Universidade de Lisboa (Orientadora)
- Doutora Ana Maria Ferreira de Sousa Sebastião, Professora Catedrática da Faculdade de Medicina da Universidade de Lisboa;
Doutor Tiago Vaz Maia, Professor Associado da Faculdade de Medicina da Universidade de Lisboa;
- Doutor João Carlos Pereira Gama Marques, Professor Auxiliar Convidado da Faculdade de Medicina da Universidade de Lisboa.

Financiada pela Fundação para a Ciência e Tecnologia no âmbito da bolsa de doutoramento PD/BD/114460/2016

2022

A impressão desta tese foi aprovada pelo Conselho Científico da Faculdade de Medicina de Lisboa em reunião de 25 de Janeiro de 2022.

As opiniões expressas nesta publicação são da exclusiva responsabilidade do seu autor.

Contents

Index of figures

Index of tables

Abbreviations

ARMS: at-risk mental state

ARMS-NT: individuals at an ARMS who did not transitioned to psychosis

ARMS-T: individuals at an ARMS who later transitioned to psychosis

APS: attenuated psychotic symptoms

BAC: balanced accuracy

BLIP: brief limited intermittent psychotic

BD: bipolar disorder

CAARMS: comprehensive assessment of at-risk mental state

CNV: copy number variation

CV: cross-validation

ERS: environmental risk score

DNN: deep neural networks

DOR: diagnostic odds ratio

GAF: global assessment of functioning

eQTL: expression quantitative trait loci

GWAS: genome-wide association study

ICR: inhomogeneity contrast ratio

IQR: weighted average image quality rating

LOP: leave-one per group-out

LDA: linear discriminant analysis

LSO: leave-one scan protocol-out

ML: machine learning

MRI: magnetic resonance imaging

NCR: noise contrast ratio

NLR: negative likelihood ratio

PLR: positive likelihood ratio

PRS: polygenic risk score

QC: quality control

RES: root-mean-squared resolution

ROIGM: regional-based grey matter volume

ROISurface: surface- and regional-based cortical thickness, and gyrification, sulci, and complexity indexes

ROIWM: regional-based white matter volume

SCZ: schizophrenia

SE: sensitivity

SIPS: structured interview for prodromal syndromes

SOPS: scale of prodromal symptoms

SNP: single nucleotide polymorphism

SP: specificity

SVM: support vector machines

VBGM: voxel-based gray matter maps

VBWM: voxel-based white matter maps

TWAS: transcriptome-wide association study

Acknowledgments

First and foremost, I would like to thank my supervisor, Dr. Diana Prata, for her continuous support and her endless patience during my PhD study. Thank you for having me in your lab and for giving me the opportunity to learn, to grow both professionally and, specially, personally and to meet so many smart and kind people. I would also like to thank Dr. Hugo Ferreira, my co-supervisor, who has inspired me to follow my academic dreams since the beginning and who I have always seen as a mentor. Thank you for all the support and shared knowledge.

I would like to thank all the people that has given me advice during this PhD study: Dr. Evangelous Vassos for all the wise advice and data regarding genetics and environment; Dr. Andrea Marquand for his advice on machine learning applied to neuroimaging; Prof. Dr. Nikolaos Koutsouleris for his advice on machine learning and the collaboration for the replication of the psychosis prediction using neuroimaging; Dr. Joanne Knight for her advice on methods to predict gene expression from genotypes; Dr. James Stone for hosting me in the Centre for Neuroimaging Sciences at the Institute of Psychiatry, Psychology & Neuroscience, King's College London, United Kingdom during my summer visit at the institute; Dr. Gareth Barker and Dr. Matthew Kempton for their advice on neuroimaging data processing; Dr. Maria João for her advice on machine learning applied to neuroimaging; and Dr. Marília Antunes for her endless availability for helping me with statistics.

I have also to thank to all the people that have collected the data used in this PhD study: Dr. Evangelos Vassos, Dr. James Stone, Dr. Isabel Valli, Dr. James Woolley, Dr. Paolo Fusar-Poli, Dr. Ceri Jones, Dr. Maria Calem, Dr. Christopher Chaddock, and Dr. Diana Prata.

I would like to thank to all my lab colleagues that somehow have helped me during this journey. To Joana Monteiro, who contributed to the development of the eGenScore tool and to Vasco Diogo who had a short term mission at Dr. Nikolaos Koutsouleris's lab to establish the collaboration between that lab and Dr. Diana Prata's lab.

A special thanks to Beatriz Simões, Mónica Costa, Inês Casqueiro, and Carina Mendes for the long smart chats, for the late evenings without filter, and for the cider during the sunsets. I cannot forget Daniel Martins and Duarte Ferreira with whom we closed the first team circle! A special thanks to Katja Brodmann who I have always seen as a reference, a tutor,

and a friend – thank you for all the smart, supportive, and fun evening chats, without them I would have lost the strength.

Finally, I would like to express my gratitude to my family, my inner circle, my superheroes team: my parents, my brothers, my sister from the heart, my life partner and my little princess. Without their tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

This research was supported by Fundação para a Ciência e Tecnologia (FCT) under the PhD fellowship (PD/BD/114460/2016) and the grant (DSAIPA/DS/0065/2018).

Abstract

Psychosis is a severe mental condition characterized by a complex set of disturbances of thinking, perception, affect and social behaviour. It is usually preceded by a prodromal phase lasting months to years and in which patients are clinically identified has being 'At Risk Mental State' (ARMS). Retrospective studies have showed that ARMS individuals have a 30% risk of transition to psychosis within the first 2 years after presentation to clinical services. Moreover, several neuroimaging, genetic and environmental biomarkers have been independently associated with the onset of psychosis in the ARMS. However, at present there is yet no established method for predicting which individuals will develop the illness and which will not – which would allow cost-efficient targeting of early intervention therapies. Furthermore, a few studies have demonstrated the feasibility to predict psychosis transition from an ARMS using structural magnetic resonance imaging (sMRI) data and machine learning (ML) methods. However, the reliability of these findings is unclear due to possible sampling bias. Moreover, the value of genetic and environmental data in predicting transition to psychosis from an ARMS is yet to be explored. In this study I aimed at predicting transition to psychosis from an ARMS using ML and quantitative data – neuroimaging, genetics, and environment – as predictors.

I used several samples (one for each modality – neuroimaging, genetics or environment) drawn from a pool of 246 subjects identified as being at an ARMS when they first sought clinical help (i.e. at baseline). Subjects were clinically identified as transitioned to psychosis (ARMS-T, 60 subjects) if they later presented a first episode of psychosis (FEP) or as not-transitioned to psychosis (ARMS-NT, 186 subjects) if they did not present a FEP within at least a period of 2 years. Structural magnetic resonance imaging, genome-wide genotypes and environmental risk assessment data was collected from the ARMS subjects at baseline. Then, the modality-specific value in predicting transition to psychosis was evaluated using a) several feature types [regional and voxel-based grey matter and white matter volumes, and regional cortical thickness, and brain gyrification, sulci depth and complexity indexes (neuroimaging); a polygenic risk score (PRS) for schizophrenia, a list of psychosis-associated single nucleotide polymorphisms (SNP), and a list of psychosis-associated genes for which several brain tissue-specific expression quantitative trait loci (eQTL) scores were extracted (genetics); and an environmental risk score (ERS) for schizophrenia, and a list of environmental risks factors (environment)], b) several feature manipulation strategies

[feature dimensionality reduction through principal component analysis, no feature selection, and forward feature selection (neuroimaging), and embedded feature selection (genetics and environment)], c) several ML algorithms [linear support vector machines (neuroimaging), elastic-net and simple logistic regression (genetics and environment)], d) several cross-validation (CV) strategies [5-fold CV and leave-one scanning acquisition protocol-out (neuroimaging) and leave-one per group, i.e. 1 ARMS-T and 1 ARMS-NT,-out (neuroimaging, genetics and environment)], e) sample balancing, i.e. same number of ARMS-T and ARMS-NT subjects, and f) bootstrapping, i.e. 5 (neuroimaging) or 100 (genetics and environment) semi-random subsamples drawn from the original pool. Then, only the modalities whose classification models showed a balanced accuracy across bootstrapped samples statistically better than chance level were included in a multimodal classification model.

Overall, this study's results showed that only genetics, and when using a set of psychosis-associated SNPs, could predict the transition to psychosis from an ARMS marginally better than chance, albeit with no clinical significance, (balanced accuracy = 53%, diagnostic odds ratio = 3.3 – averaged across bootstrapped samples). Furthermore, the environmental and neuroimaging alone could not predict psychosis from an ARMS, statistically better than chance. Therefore, no multimodal classification model was trained/tested. Moreover, and unexpectedly, I could not replicate previous findings showing the usefulness of structural MRI in predicting transition to psychosis from an ARMS using ML. Therefore, my results suggest that: a) genetic data may be promising for predicting transition to psychosis from an ARMS; and b) the value of structural MRI data in predicting psychosis from an ARMS, as suggested by previous evidence, should be reconsidered. Finally, this study serves as a proof-of-concept on how multimodal quantitative data can be used to predict psychosis development already from a prodromal stage and should be replicated in larger ARMs samples.

Keywords: at-risk mental state; transition to psychosis prediction; structural neuroimaging; genetics; environmental risk assessment

Resumo

A psicose pode ser definida como uma condição psiquiátrica caracterizada por um conjunto complexo de distúrbios do pensamento, da perceção, do humor e do comportamento social. Esta pode ser desencadeada por uma doença mental, como, por exemplo, a esquizofrenia, por uma doença neurológica, como, por exemplo, a doença de Alzheimer, ou por uma outra condição médica, como, por exemplo, ser induzida por fármacos ou substância ilícitas. A doença psicótica, quando despontada por perturbação psiquiátrica, é normalmente precedida por uma fase prodromal que pode ter uma duração de alguns meses a alguns anos. Durante esta fase os doentes podem apresentar sintomas psicóticos ligeiros, que normalmente se resolvem sem intervenção clínica e num curto espaço de tempo. Os doentes que procuram apoio clínico durante este período são normalmente identificados como estando em risco de desenvolver doença psicótica (ARMS, do inglês *at-risk mental state*). Estudos retrospetivos demonstraram que indivíduos identificados como estando em ARMS têm um risco de 30% de transitarem para uma doença psicótica nos primeiros 2 anos após a sua primeira apresentação nos serviços clínicos.

Vários biomarcadores de neuroimagem, genética e ambiente têm sido associados ao início da doença psicótica de forma consistente. Entre estes estão, por exemplo, a) o volume da substância cinzenta em diversas regiões cerebrais (incluindo, o cíngulo anterior, o cerebelo e os córtices frontal, temporal, parietal e insular) que foi demonstrado estar reduzido nos doentes ARMS que transitam mais tarde para doença psicótica em comparação com os que não transitam; b) mutações de nucleótico único (SNP, do inglês *single nucleotide polymorphim*) em genes codificantes de proteínas envolvidas na formação de sinapses e cujo um dos alelos confere risco para esquizofrenia quando comparado com indivíduos saudáveis, e c) a exposição a eventos traumáticos durante a infância, nomeadamente ao sofrimento por abuso sexual, que confere um maior risco de transição para doença psicótica comparativamente à não exposição. Contudo, atualmente não há um método estabelecido e amplamente aceite pela comunidade científica e clínica para a previsão de quais os indivíduos que irão efetivamente desenvolver doença psicótica e quais não a irão desenvolver. Alguns estudos têm demonstrado de forma consistente a viabilidade de prever a transição para doença psicótica a partir do ARMS utilizando imagiologia de ressonância magnética estrutural (sMRI) e métodos de *machine learning* (ML). No entanto, a viabilidade destes resultados necessita de validação, pois estão possivelmente sob a influência de viés de amostragem, nomeadamente, do pequeno tamanho da amostra e da proveniência dos

dados de uma única fonte). Além disso, o valor de dados genéticos e de ambiente na previsão da transição para doença psicótica a partir de ARMS está ainda por explorar. Portanto, o principal objetivo deste estudo é prever a transição para doença psicótica a partir de ARMS utilizando métodos de ML e dados quantitativos, nomeadamente neuroimagiológicos, genéticos e ambientais, como preditores.

Para este estudo uma amostra por cada modalidade, isto é, uma para a neuroimagiologia, outra para a genética e outra para o ambiente, foram definidas a partir de um *pool* de dados de 246 indivíduos que foram clinicamente avaliados como estando em ARMS aquando da primeira vez que estes procuraram ajuda clínica. Neste ponto de amostragem temporal define-se a *baseline* (isto é, o ponto de referência). Numa segunda avaliação clínica, os indivíduos em ARMS foram identificados como tendo transitado para doença psicótica (ARMS-T, 60 indivíduos) se estes apresentaram um primeiro episódio psicótico (FEP, do inglês *first episode of psychosis*) ou como não tendo transitado para doença psicótica (ARMS-NT, 186 indivíduos) caso não tenham apresentado um FEP dentro de um período de pelo menos 2 anos. No momento da *baseline* imagens de sMRI, genótipos de cobertura genómica (do inglês *genome-wide*) e avaliação de fatores de risco ambiental dos indivíduos em ARMS foram recolhidos. O valor, avaliado com medidas de performance como a acurácia da previsão, destes dados em prever a transição para doença psicótica a partir de ARMS foi avaliado da seguinte forma. Primeiro procedeu-se à avaliação da performance de modelos preditivos – isto é, modelos de classificação – treinados com dados de cada uma das modalidades individualmente e utilizando a) vários tipos de *features*, nomeadamente neuroimagiológicas [volumes de substância cerebral cinzenta e branca (regional e mapas de vóxeis), espessura cortical regional, e índices regionais de girificação, da profundidade dos sulcos e de complexidade], genéticos [score de risco poligénico (PRS, do inglês *polygenic risk score*) para a esquizofrenia, uma lista de SNPs associadas à esquizofrenia, uma lista de genes associados à esquizofrenia para os quais uma score de loci de traço quantitativo de expressão (eQTL, do inglês *expression quantitative trait loci*) de vários tecidos cerebrais] e ambientais [score de risco ambiental (ERS, do inglês *environmental risk score*) e uma lista de fatores de risco ambiental]; b) várias estratégias de manipulação de *features*, nomeadamente, redução da dimensionalidade das *features* através da análise de componentes principais, sem seleção de *features* e com seleção de *features* progressiva (para os dados neuroimagiológicos) e seleção de *features* embutida no algoritmo de ML (para os dados genéticos e ambientais); c) vários algoritmos de ML, nomeadamente, máquinas de

vetores de suporte (*support vector machines* do inglês) lineares (para os dados neuroimagiológicos), e regressão logística simples e de *elastic-net* (para os dados genéticos e ambientais); d) várias estratégias de validação cruzada (CV, do inglês *cross-validation*), nomeadamente, 5-*fold* CV e *leave*-um protocolo de aquisição de imagens sMRI-*out* (para os dados neuroimagiológicos) e *leave*-um por grupo, i.e. 1 ARMS-T and 1 ARMS-NT,-*out* (para os dados neuroimagiológicos, genéticos e ambientais); e) balanceamento da amostra, i.e. o mesmo número de indivíduos ARMS-T e ARMS-NT; e f) *bootstrapping*, i.e. 5 (para os dados neuroimagiológicos) e 100 (para os dados genéticos e neuroimagiológicos) subamostras definidas de forma semi-aleatória a partir da amostra original. Segundo, apenas as modalidades cujos modelos de classificação mostraram uma acurácia balanceada, calculada como a acurácia mediana ao longo das amostras *bootstrapped*, estatisticamente superior à chance foram incluídas num modelo de classificação multimodal, isto é, treinados com os dados das três modalidades em conjunto.

Globalmente, os resultados demonstraram que apenas a genética, na forma de uma lista de SNPs previamente associadas à presença de esquizofrenia, conseguiu prever a transição para a psicose a partir do ARMS marginalmente melhor do que a chance. Estes modelos de classificação demonstraram uma acurácia média de 53% e uma razão de probabilidade de diagnóstico média de 3.3, o que significa que as probabilidades de obter um diagnóstico positivo verdadeiro, ou seja, de prever a transição para doença psicótica de um indivíduo ARMS que verdadeiramente desenvolveu a doença é, em média, 3 vezes maior do que obter um diagnóstico positivo falso, ou seja, de prever a transição para doença psicótica quando o indivíduo não desenvolveu a doença. Por outro lado, a neuroimagem e o ambiente, individualmente, não conseguiram prever o desenvolvimento de psicose a partir do ARMS, isto é, a acurácia média dos modelos de classificação treinados com dados de cada uma das modalidades não superou, estatisticamente, a chance. Por este motivo, nenhum modelo de classificação multimodal foi treinado e testado. Ademais, e contrariamente ao esperado, os estudos que anteriormente demonstraram a utilidade de dados de sMRI na previsão da transição para doença psicótica a partir do ARMS utilizando ML não foram replicados na amostra deste estudo.

Em suma, estes resultados sugerem que, em primeiro lugar, os dados genéticos poderão um preditor promissor da transição para doença psicótica a partir do ARMS, em segundo lugar, o valor dos dados sMRI em prever a psicose a partir de um estado prodromal da doença, tal como descrito na literatura anterior, deverá ser reconsiderado, e que, em terceiro lugar, os

dados ambientais e de neuroimagem, individualmente, não são suficientes para prever a transição para psicose melhor do que a chance. Este trabalho, no entanto, não está isento de limitações, sendo a mais evidente o tamanho pequeno da amostra de indivíduos em ARMS. Apesar de amostra inicial ter um tamanho considerável, as subamostras de indivíduos com dados de cada modalidade, isto é, neuroimagem, genética e ambiente, são substancialmente menores (máximo número de indivíduos nestas amostras é de 74 – para os dados ambientais). Para além do tamanho da amostra, a especificidade da mesma é outra limitação. Os indivíduos incluídos neste estudo foram clinicamente identificados como estando em ARMS com base na avaliação de sintomas essencialmente positivos. No entanto, está já amplamente evidenciado que estes são, na verdade, a última tipologia de sintomas a aparecer no estado prodromal da doença, sendo estes precedidos pelos sintomas básicos e negativos. Desta forma, é imperativo que este estudo seja replicado em amostras maiores e mais inclusivas, isto é, que recrutem indivíduos em ARMS que sejam identificados tanto com sintomas positivos, como com sintomas básicos e negativos.

Palavras-chave: estado mental de risco para psicose; predição da transição para doença psicótica; neuroimagem estrutural; genética; avaliação de risco ambiental

1. General introduction

1.1. Psychosis

Psychosis is a severe condition usually triggered by a mental disorder (e.g. schizophrenia spectrum disorder), a neurological disease (e.g. Alzheimer's disease) or other medical condition (e.g. induced by drugs or illicit substances) characterized by a disentanglement from the reality (Schrimpf et al., 2018). Hallucinations (auditory, visual or sensorial perceptions that have no correspondence with real stimuli), delusions (false beliefs), disorganized speech, behavior (including motor behavior, e.g. catatonia) and thinking, and lack of concentration are positive symptoms usually exhibited by an individual with a psychotic illness (Schrimpf et al., 2018).

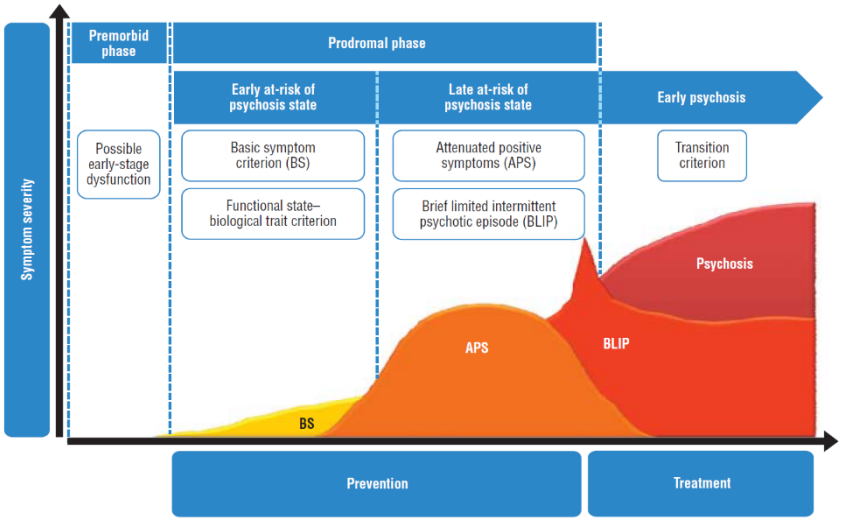


Figure 1. Model of psychosis onset from the clinical high-risk state. The higher the line on the y-axis, the higher the symptom severity. Image from (Paolo Fusar-Poli, Borgwardt, et al., 2013).

The onset of psychosis, when triggered by a mental disorder, is typically preceded by a prodromal phase that lasts months to years [usually starts early in the adolescence and precedes the onset of the psychotic symptoms by 10 or more years (Kahn et al., 2015)] and which can be characterized by two sub-phases: the early and late prodromal phases [**Figure 1**; (Paolo Fusar-Poli, Borgwardt, et al., 2013)]. In the very early prodromal phase subtle and subjectively experienced disturbances in mental processes start to emerge. Disturbed mental processes include perception (body and sensory), attention, thinking, speech, stress tolerance, drive, affect and central-vegetative functions (e.g., eating, sleeping, bladder activity). These are called basic symptoms as they are the first manifestations of the

neurobiological processes underlying psychosis and are mainly distinguished from other symptoms (i.e. positive or negative symptoms) by their self-experience nature, i.e. they are usually not evident to others until one is no longer able to cope with their basic symptoms (Schultze-Lutter et al., 2012). Therefore, as the course of the psychotic illness evolves more disabling behavior and cognitive symptoms start to emerge. In particular, a reduction or complete absence of normal behaviors related to motivation or expression, generally called negative symptoms, appear during the early phase of the prodrome (Correll & Schooler, 2020). These symptoms include alogia (poverty of speech), avolition (inability to initiate and persist in goal-directed activities), affective flattening, anhedonia (inability to experience pleasure), and asociality (inability to engage in social interaction). Moreover, individuals in a psychosis prodrome progressively show cognitive deficits in attention, memory, reasoning, and executive functioning (Sheffield et al., 2018). The late phase of the psychosis prodrome is mainly dominated by positive symptoms.

One can be clinically identified has being at a late prodromal phase of psychosis or at an 'At Risk Mental State' (hereinafter: ARMS; also known as 'Clinical High Risk' or 'Ultra High Risk') if one presents a functional decline in association with one or more of the following criteria: 1) attenuated psychotic symptoms (APS), such as delusions, hallucinations, or disorganized speech with a frequency of at least once per week in the past month; 2) a brief limited intermittent psychotic (BLIP) episode lasting less than one week which resolves without antipsychotic medication; or 3) a genetic liability to psychosis or schizotypal traits, i.e. having either a first-degree relative with psychosis or a schizotypal personality disorder (Paolo Fusar-Poli, Borgwardt, et al., 2013). APS and BLIP criteria can be assessed using several psychometric instruments, such as, the Comprehensive Assessment of ARMS [CAARMS; (Yung et al., 2005)], the Structured Interview for Prodromal Syndromes [SIPS; (T. J. Miller et al., 2003)], the Scale of Prodromal Symptoms [SOPS; (T. J. Miller et al., 2003)], or the Basel Screening Instruments for Psychosis [BSIP; (Riecher-Rössler et al., 2008)]; whereas the Global Assessment of Functioning [GAF; (American Psychiatric Association, 2000)] rates subjectively the social, occupational, and psychological functioning of the individual. Furthermore, as the ARMS criteria rely on young help-seeking individuals aged 8 to 40 years, the truly prevalence of the ARMS in the general population is currently unknown (Paolo Fusar-Poli, Borgwardt, et al., 2013).

Transition to psychosis from an ARMS is currently evaluated using the above-mentioned psychometric tools and is based on the severity, frequency, and total duration of the

psychotic symptoms, i.e. when the subject experiences a first episode of psychosis (FEP). CAARMS, for example, requires that one experiences at least one fully developed positive symptom several times a week for over 1 week to be considered as transitioned to a FEP (Yung et al., 2005). Moreover, subjects at an ARMS and seeking-help have a transition rate to psychosis of about 18% in the first 6 months and 32% in the first 3 years (Paolo Fusar-Poli, Borgwardt, et al., 2013) and, in particular, an increased risk of transition to schizophrenia of 15.7% within an average period of 2.35 years, as shown by a meta-analysis (Paolo Fusar-Poli, Bechdolf, et al., 2013). This means that most of the people at an ARMS that later develop a psychotic illness will be diagnosed with Schizophrenia.

Schizophrenia is the most disabling, chronic and severe psychotic illness affecting with a lifetime prevalence of about 1% of the world population (Kahn et al., 2015). Mortality rate in schizophrenia is high when compared with healthy population – a patient with schizophrenia dies, on average, 20 years earlier than a healthy person (Seeman, 2019). A diagnosis of schizophrenia as defined by the Diagnosis of and Statistical Manual of Mental Disorders, Fifth Edition [DSM-5; (American Psychiatric Association, 2013)] requires a) at least one positive (delusions, hallucinations, disorganized speech, grossly disorganized or catatonic behavior) and other positive or negative (diminished emotion expression or avolition) symptom. These two symptoms must be present for a minimum period of 1 month (or less if successfully treated). b) a decline in functioning (in work, interpersonal relations or self-care) since the onset of the illness; and c) a total illness duration of at least 6 months. Current pharmacological treatment for schizophrenia are limited to treat positive symptoms (i.e. to help reduce their intensity and frequency) by administering antipsychotics and their efficacy has been reported to be poor in the majority of chronic patients – with only about one quarter showing a 50% reduction on positive symptoms – as by a recent meta-analysis (Leucht et al., 2017). Negative and cognitive symptoms are generally ameliorated by psychosocial interventions, including cognitive behavioral therapy (Nowak et al., 2016), behavioral skills training (Turner et al., 2018), and family interventions (Pharoah et al., 2010). Given the very socioeconomic burden that a schizophrenia diagnosis has on patients, their families and on the health systems that take care of them and the substantial poor functioning (i.e. social and cognitive) throughout their lifetime, efforts have been put forward to explore preventive approaches to this illness (Millan et al., 2016).

Current preventive treatment is given to subjects at an ARMS, including psychosocial interventions, and the administration of omega-3 poly-unsaturated fatty acids and

antipsychotics (Millan et al., 2016). Treatment effectiveness at the prodromal phase of psychosis has been showed to lower the risk of developing psychosis by decreasing the transition rate from 32% in subjects at an ARMS without treatment to 11% in subjects at an ARMS with treatment at 12 months (Preti & Cella, 2010), irrespective of the treatment type. This finding has been confirmed by recent meta-analyses of randomized controlled treatment trials in the ARMS population (Paolo Fusar-Poli, Borgwardt, et al., 2013; Stafford et al., 2013; Van Der Gaag et al., 2013). However, there is no conclusive evidence of the benefits of any specific treatment nor whether the lower risk of transition to psychosis translates to a delay of the onset of the illness or to a successful prevention of the illness (Millan et al., 2016). Furthermore, stigmatization (Shrivastava et al., 2013) and adverse effects (Siafis et al., 2017) of antipsychotics are well-known and a concern for many young people at a prodromal phase of psychosis. This is even more worrisome given that about 70% of subjects diagnosed with an ARMS do not even develop any full-blown psychotic illness (Paolo Fusar-Poli, Bechdolf, et al., 2013), i.e. these people may suggestively benefit from a less aggressive treatment to ameliorate their symptoms or need no treatment at all. Such increase in treatment cost-effectiveness would represent a substantial decrease in healthcare costs in personal and medication, and treatment burden to patients, including pharmacological side-effects. However, there is no established method to distinguish between subjects that being already diagnosed with an ARMS will further develop a psychotic illness from those that will not – i.e. all the clinical assessments throughout the development of the prodromal phase of psychosis are, to date, subjective. Therefore, there is an urgent need for effective, precise, and quantitative tools for prediction of transition to psychosis from an ARMS (i.e. from a prodromal phase of psychosis).

1.2. Machine learning

In the last decades efforts have been put forward to make diagnosis and prognosis in psychiatry more objective, instead of being based on subjective clinical assessments (Dwyer et al., 2018; Sanfelici et al., 2020; Tai et al., 2019). The advent of machine learning (ML) enabled one to make individual (i.e. subject-level) predictions, as opposite to classical statistical inference (i.e. group-level; (Bzdok et al., 2018)). One main advantage of using ML in psychiatry is that it considers the individual differences inherent to psychiatry disorders, which are not represented when making group-level inferences. Thus, ML models learn

individual patient characteristics – i.e. as the psychiatric disorder-specific pattern – and then, use this pattern to make single-subject predictions.

ML models can be generally divided into supervised and unsupervised categories. The main difference among the two are the *a priori* knowledge of the data used to train the models: in supervised ML the labels of the training dataset are known, whereas in unsupervised ML they are not. Taking the example of schizophrenia diagnosis, a supervised ML model would learn the relationship between subjects' characteristics (e.g. neuroimaging data) and the labels (i.e. diagnosis: schizophrenia *versus* healthy), with these labels being assigned by human experts (i.e. by clinicians using standard subjects clinical assessments such as DSM-5). On the other hand, an unsupervised ML model would explore the subject's characteristics, with no data labeling, in search for hidden structure in the data. Whereas in the first example one would use the ML model to predict the diagnosis (i.e. schizophrenia *versus* healthy) to a previously unseen subject's data, in the second example one would use the model to explore how subjects can be sub-grouped based on their brain signatures. It is worth to note that when using human-based labels (as it is the case with the clinical psychiatric diagnoses) in supervised ML, the expected performance (i.e. accuracy) of the models is upper limited by how accurate the labeling by the human expert is (Vu et al., 2018).

A brief overview of the supervised ML model training and evaluation is given below (albeit some of the discussed aspects would also apply to unsupervised ML – given that in this project only supervised ML was used this will be the focus of this dissertation hereinafter). The development of a valid and translational ML model should essentially take into account its accuracy and generalizability (Dwyer et al., 2018). These two concepts are highly dependent on the sample type and size used to train them. Generally, small sample sizes tend to be clinically more homogeneous (i.e. less representative of the real-world population of interest), whereas bigger sample sizes usually integrate several cohorts and, therefore, are more heterogenous regarding its demographic, geographic and phenotypic characteristics (Schnack & Kahn, 2016). Moreover, it has been shown that ML models have lower performances when trained with larger sample sizes, as compared with the ones trained with smaller samples (Kambeitz et al., 2018). In summary, a ML model can be highly accurate, but poorly generalizable (especially when trained with small samples sizes – i.e. overestimation problem) or poorly accurate, but highly generalizable (usually trained with higher sample sizes – e.g. from multisites/cohorts); with generalizability here meaning the

extent to which a ML model trained in one given sample performs accurately in an independent one.

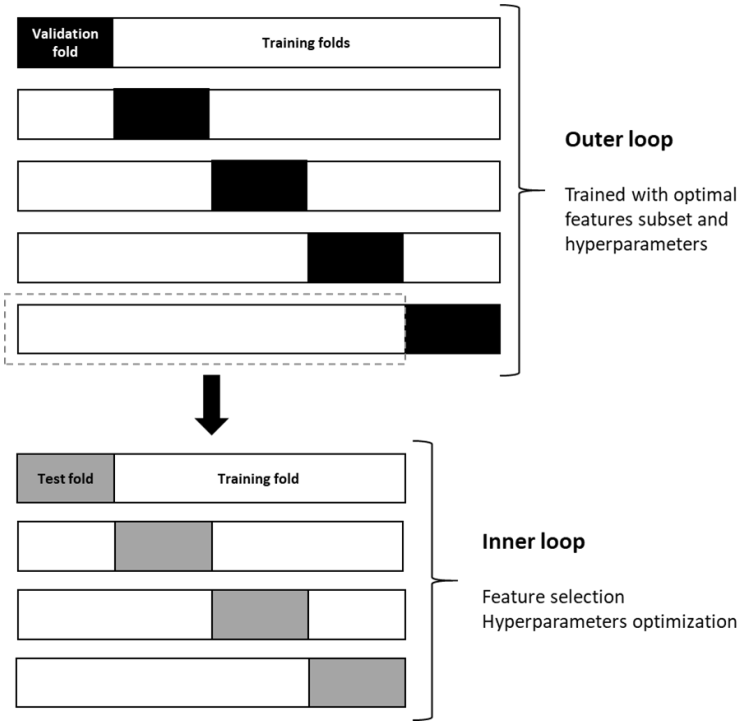


Figure 2. Nested cross-validation scheme.

One way to circumvent the overestimation problem and to make our ML model generalizable is to apply a well outlined train/test/validation approach to it – i.e. through data re-sampling. Ideally, if more than one sample, i.e. drawn from two independent populations, are available, one would use one for training and testing the ML model (i.e. performing feature selection and hyperparameters optimization) and the other for validating the model (i.e. to estimate the generalizability of the model). Additionally, if the sample size is big enough, one would even split the original first sample into two smaller subsamples: one for training the ML model, and the other for testing the ML model. This is the gold standard in all translational science (Dwyer et al., 2018). However, when sample sizes are limited generalizability can be estimated through more robust and complex data re-sampling methods – one of those being the nested cross-validation (CV) (**Figure 2**). In a CV scheme, the sample is divided into subsets of individuals, also known as folds (e.g. in 10 folds). Then the ML model is trained on all folds except one [i.e. (n – 1) folds] and tested on the left-out one. This process is then repeated for all the remaining folds and the out-of-sample generalizability of the ML model is computed as the average accuracy of the all the left-out test folds. The sample can

ultimately be divided into many folds as the number of individuals in the sample. This approach is called leave-one-out CV and is usually not recommended due to its unstable and possibly biased results (Varoquaux et al., 2017). Moreover, the current gold-standard CV scheme is to separate the data into train, test and validation subsamples (i.e. to avoid double dipping the data) in a nested CV scheme. First, the model is optimized (i.e. feature- and hyperparameter-wise) in an inner CV loop – i.e. a model is trained for each feature/hyperparameter combination, its accuracy is tested and the combination yielding the best accuracy is selected. Then, the optimized model (i.e. with the best combination of features/hyperparameters) is trained and validated in an outer CV loop – i.e. the generalizability is estimated in the completely left-out validation sample (i.e. not used in the model optimization inner CV loop). A more detailed description of nested CV scheme, feature selection methods and hyperparameters optimization can be found in the section **3.2.5. Machine learning approach** in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**.

In the context of translational ML, another key characteristic of the model is its clinical meaningfulness – i.e. being able to interpret the model by understanding which constellation of features lead to the model predictions. This task is highly dependent on the type of ML algorithm used – i.e. the more complex the algorithm is, the harder is to pinpoint the constellation of features contributing to the prediction. This may be one of the reasons why most of the ML studies in psychiatry have been using linear algorithms (Arbabshirani et al., 2017; Kambeitz et al., 2015), which are reasonably easy to interpret – e.g. regularized linear or logistic regression (Zou & Hastie, 2005); and linear support vector machines [SVM; (Cortes & Vapnik, 1995; B Scholkopf et al., 2003)]. A more detailed description of the SVM and regularized logistic regression algorithms can be found in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**, and **5. Individual prediction of transition to psychosis using genetics and machine learning**, respectively**.**

Finally, a review of the studies employing machine learning to diagnose psychosis (as opposed to healthy controls) using as training data 1) structural neuroimaging will be described in the section **1.3.2. Individual prediction of psychosis using structural neuroimaging and machine learning**; 2) genetics will be described in the section **1.4.2. Individual prediction of psychosis using genetics and machine learning**; and 3) environmental data will be described in the section **1.5.2. Individual prediction of**

psychosis using environmental risk factors and machine learning. A review of the studies employing machine learning to predict transition to psychosis (as opposed to no transition) from an ARMS using as training data 1) structural neuroimaging will be described in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**; 2) genetics will be described in **5. Individual prediction of transition to psychosis using genetics and machine learning**; and 3) environmental data will be described in **6. Individual prediction of transition to psychosis using environment and machine learning**. Additionally, multimodal psychosis prediction studies using two or more modalities (being at least one structural neuroimaging, genetics, or environmental data) will be described in the section **1.6. Individual prediction of psychosis using multimodality and machine learning**.

1.3. Structural neuroimaging in psychosis

1.3.1. Potential structural neuroimaging biomarkers

Magnetic resonance imaging (MRI) has been extensively used to study potential neuroimaging, in particular brain structural, biomarkers in psychosis. MRI is a powerful non-invasive technique that produces three dimensional anatomical images with a very high resolution. It relies on mapping the distribution of protons (i.e., the nucleus of hydrogen atoms) as they have the highest abundance in the body (i.e., being present in water and fat). Brain tissues, i.e., grey matter (that consists mainly in neuronal cell bodies) and white matter (that consists mainly in myelinated axons), and non-brain tissue, i.e. cerebrospinal fluid, are characterized has having a different concentration of protons and, therefore, they appear in a structural MRI image with different intensities. Brain volumetric studies take advantage of this technique to study brain morphological alterations, i.e., whole brain or regional alterations in a) grey or white matter volumes, b) cortical thickness, c) brain gyrification, or d) sulci depth in the presence of a given pathology, such as psychosis.

Over the past two decades, hundreds of studies have compared samples of various stages of psychosis and healthy controls [for a complete review see (Chung & Cannon, 2015; Shepherd et al., 2012); for a recent meta-analysis see (Kuo & Pogue-Geile, 2019)]. It has been consistently shown that, when compared to healthy controls, patients with chronic schizophrenia present smaller intracranial, total brain and total grey matter volumes (Chung & Cannon, 2015; Kuo & Pogue-Geile, 2019). Furthermore, the reduction of grey and white matter volumes are usually accompanied by an enlargement of the ventricles (Kuo & Pogue-

Geile, 2019), hypothesized as a mechanism aimed to preserve the intracranial volume (Chung et al., 2017). Moreover, the mean decrease in the intracranial volume and the mean enlargement of the ventricles were shown to be accompanied by an increase in variability in patients with schizophrenia when compared to healthy controls (Kuo & Pogue-Geile, 2019).

More specifically, the gray matter volume has been shown to be decreased in schizophrenia patients, when compared to healthy controls, in the prefrontal, medial and superior temporal cortices and in the hippocampus. Volume loss in the superior temporal cortex, in particular, in the superior temporal gyrus, has been suggested to be associated with psychotic symptoms (i.e., hallucinations, delusions and formal thought disorder) as this region mediates auditory-language processing (Jung et al., 2019; Kasai et al., 2003). Moreover, the loss of volume in the prefrontal and temporal cortices seem to progress with the development of psychosis after its onset [as shown by longitudinal studies comparing FEP patients with healthy controls (Ho et al., 2003; Kasai et al., 2003; M. Nakamura et al., 2007)]. Another key brain structure which volume has been consistently shown to be reduced in the presence of psychosis is the hippocampus, i.e. when comparing patients with schizophrenia (Kuo & Pogue-Geile, 2019) or patients with schizophrenia or bipolar disorder with psychotic symptoms (Haukvik et al., 2018) with healthy controls. Although the role of this structure in psychotic disorders is not yet known, it has been suggested that it has a role in the formation of psychotic symptoms (Haukvik et al., 2018).

One key potential confounder in brain volumetric studies analyzing patients with an already established psychotic disorder is whether the patients are under antipsychotic drugs. The administration of these type of drugs has shown to influence brain imaging measures [e.g. volume measures; (Navari & Dazzan, 2009; Voineskos et al., 2020)]. Therefore, the volumetric alterations seen in patients may be directly associated with the pathophysiology of the disorder or with the accumulated impact of medication. Such a potential confounder can be overcome by searching for neuroimaging biomarkers around the time of the disorder's onset, i.e. comparing individuals at an ARMS with healthy controls and between individuals at an ARMS that later transitioned to psychosis with those that did not [for a recent and complete review see (Andreou & Borgwardt, 2020)]. In summary, when compared to healthy controls, individuals at an ARMS have consistently showed a grey matter volume reduction in the hippocampus, parahippocampus, cingulate, medial and lateral frontal, and parietal cortices (Bois et al., 2015; P. Fusar-Poli et al., 2011). Moreover, individuals at an ARMS who later transitioned to psychosis have shown a reduction of the gray matter volume of the

anterior cingulate, cerebellum, and frontal, temporal, parietal, and insular cortices, and an increase of the pituitary volume when compared to those who did not (Borgwardt et al., 2008; Dazzan et al., 2012; Pantelis et al., 2003; Takahashi et al., 2009). Furthermore, reduction of the insular cortex volume was found to be associated with both negative (Smieskova et al., 2012; Takahashi et al., 2009) and positive symptoms (Smieskova et al., 2012). In addition, the above reported gray matter volume reduction has shown a steeper rate in individuals at ARMS that later transitioned to psychosis [*vs.* the ones that did not; (Borgwardt et al., 2008; Pantelis et al., 2003; Takahashi et al., 2009)]. Furthermore, the steeper rate of gray matter volume reduction has been associated with higher level of unusual thought content only in individual at an ARMS who later transitioned to psychosis [i.e. the association was not found in individuals at an ARMS who did not developed psychosis; (Chung et al., 2015)]. Lastly, these findings are consistent with the notion of neurodevelopmental abnormalities leading to the emergence of psychotic disorders. Additionally, although one cannot exclude the influence of antipsychotic drugs administration in the above reported findings (as this effect was not directly addressed by those studies), there are some studies showing a differential grey matter loss among individuals at an ARMS who later transitioned to psychosis regardless of drug administration (T. D. Cannon et al., 2015) and in antipsychotic naïve FEP patients compared to healthy controls (Haijma et al., 2013), suggesting that at least some of the grey matter loss is independent of the antipsychotics exposure.

The volume differences in white matter have also been explored in psychosis, although to a lower extent. When compared to healthy controls, patients with chronic schizophrenia have shown reduced white matter volume in the prefrontal cortex (Breier et al., 1992; Haijma et al., 2013; Hulshoff Pol et al., 2002; Suzuki et al., 2002; Wagner et al., 2013) and in the internal capsule (Suzuki et al., 2002; Zhou et al., 2003), a brain structure that contains projections from the medial dorsal thalamus to the frontal lobes. Furthermore, the maturation of white matter tracts (i.e. the myelinization of axons), especially in the prefrontal cortex, has been shown to be associated with the development of cognitive functions (Nagy et al., 2004). Additionally, reduction in these structures' white matter volume have been suggested to underlie the cognitive impairments seen in patients (Nagy et al., 2004; Suzuki et al., 2002; Wagner et al., 2013; Zhou et al., 2003). Moreover, changes in white matter volume seem to appear already at the prodromal phase of psychosis. Individuals with ARMS who later transitioned to psychosis have showed a) reduced white matter volume in the right superior

temporal lobe, when compared with healthy controls (Witthaus et al., 2008); and b) larger volume of white matter in the left frontal lobe, when compared with individuals at an ARMS who do not transitioned to psychosis (Walterfang et al., 2008). The white matter abnormalities seem to progress with the development of the psychotic disorder, with FEP patients showing a widespread reduction in white matter volume when compared to individuals at an ARMS (Witthaus et al., 2008). Furthermore, longitudinal reduction in the white matter volume of the left fronto-occipital fasciculus has also been observed in individuals with ARMS who later transitioned to psychosis, but not in those who did not – suggesting that this tract may be associated with the onset of psychosis (Walterfang et al., 2008).

Moreover, cortical abnormalities in psychosis have also been explored through cortical thickness measurements. Generally, cortical thinning has been observed in both FEP (Sprooten et al., 2013) and schizophrenia patients (Van Haren et al., 2011), compared to healthy controls. Moreover, this cortical thinning has been shown to increase with the disease progression, particularly, in the temporal and frontal cortices (Van Haren et al., 2011). Furthermore, the reduction in cortical thickness, when compared to controls, has been reported to be already present at an ARMS, particularly in a) the temporal lobe (Sprooten et al., 2013); specifically in b) the superior temporal gyrus (Benetti et al., 2013) – a result also reported in FEP patients; and c) the parahippocampal gyrus (S Tognin et al., 2014). No significant differences in cortical thickness between ARMS who later transition to psychosis and those who do not have been consistently reported (Sprooten et al., 2013; S Tognin et al., 2014).

Gyrification of the brain refers to the level of cortical folding, usually measured as a gyrification index, and has been recently proposed as a potential biomarker for psychosis (White & Hilgetag, 2011). When compared to controls, patients with chronic schizophrenia have shown a) lower gyrification index in the frontal lobe [i.e. prefrontal (Bonnici et al., 2007; McIntosh et al., 2009), anterior (Nanda et al., 2014) and posterior (Nanda et al., 2014; Wheeler & Harper, 2007) cingulate cortices; middle frontal sulcus (Cachia et al., 2008); inferior frontal gyrus (Madeira et al., 2020); Broca's area (Cachia et al., 2008)], in the temporal lobe [i.e. superior temporal sulcus (Cachia et al., 2008)], and in the parietal lobe [i.e. supramarginal gyrus (Madeira et al., 2020)]; and b) greater gyrification index in the frontal cortex (Falkai et al., 2007) and in the frontal and temporal poles and in the insula (Spalthoff et al., 2018). Furthermore, reduction in total gyrification index has been

associated with greater disorganized and negative symptoms in patients with chronic schizophrenia (Sallet et al., 2003). Moreover, abnormalities in cortical gyrification have also been reported in early stages of psychosis, with FEP patients presenting higher temporal lobe gyrification index than healthy controls (Harris et al., 2004). Higher gyrification index has also been shown in individuals at an ARMS in the frontal cortex, when compared with healthy controls (Jou et al., 2005). Moreover, the transition to psychosis has been shown to be associated to a higher gyrification index in the prefrontal cortex (Harris et al., 2007).

On a final note, one study has explored differences in the sulci depth between ARMS subjects and healthy controls showing reversed hemispheric asymmetry in their cingulate sulci depth (Goghari et al., 2007).

1.3.2. Individual prediction of psychosis using structural neuroimaging and machine learning

Brain structural neuroimaging has been widely used with ML methods to make individual predictions in psychosis. Three main prediction problems in the context of psychosis are describe in this section: a) classifying patients in a prodromal phase (i.e., ARMS) or with a recent-onset (i.e., FEP) or with a chronic (i.e., schizophrenia) psychotic disorder versus healthy controls; b) classifying patients in distinct stages of a psychotic disorder (i.e., at ARMS, FEP, or schizophrenia); and c) prediction of disease progression (**Tables 1-3**). The literature review regarding the prediction of disease progression from an ARMS will be described in section **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**.

Among the most common extracted features and ML algorithms in these studies are brain volume and cortical thickness from structural MRI (sMRI) and SVM (Castellani et al., 2012; de Wit et al., 2017; Fan et al., 2005, 2007; Iwabuchi et al., 2013; Koutsouleris et al., 2009; Lu et al., 2016; Mourao-Miranda et al., 2012; Nieuwenhuis et al., 2012; Pettersson-Yeo, Benetti, Marquand, Dell'Acqua, et al., 2013; Pina-Camacho et al., 2015; Pinaya et al., 2016; Rozycki et al., 2018; Schnack et al., 2014; Sui et al., 2013; Valli et al., 2016; Vieira et al., 2020; Xiao et al., 2019; Zanetti et al., 2013; Zhang & Davatzikos, 2013) and linear discriminant analysis (LDA; particularly in older studies; (Csernansky et al., 2004; Janousova et al., 2015; Karageorgiou et al., 2011; Kasparek et al., 2011; Kawasaki et al., 2007; K. Nakamura et al., 2004; Ota et al., 2012; Radulescu et al., 2014; Takayanagi et al.,

2011)), respectively. SVM works by estimating an optimal hyperplane, whereas LDA finds a linear combination of features that maximizes the separation between two classes (e.g., schizophrenia patients from healthy controls). A few recent studies have started to employ deep neural networks (DNN) for classification (Pinaya et al., 2016, 2019; Vieira et al., 2020). DNN are hierarchical structures composed of various levels of complexity able to apply nonlinear transformations to the input data (e.g., brain sMRI images). These transformations result in increasingly higher levels of abstraction, where higher-level features are more invariant to the noise present in the input data than lower-level ones. Furthermore, the overall accuracy reported in these studies seems to decrease with sample size, which poses a concern to the generalizability of the studies with small sample sizes. Moreover, several of these studies did not employed a rigorous validation strategy (i.e. holding out part of the sample for validation – that is not used to train the ML model) (Castellani et al., 2012; Kawasaki et al., 2007; K. Nakamura et al., 2004; Sui et al., 2013), which may rise the concern of overfitting and, therefore, overoptimistic results.

Chronic schizophrenia patients have been successfully classified against healthy controls with overall accuracies between 52% and 90% (Castellani et al., 2012; Csernansky et al., 2004; Davatzikos et al., 2005; Greenstein et al., 2012; Iwabuchi et al., 2013; Kawasaki et al., 2007; Lu et al., 2016; Nieuwenhuis et al., 2012; Pinaya et al., 2016; Radulescu et al., 2014; Rozycki et al., 2018; Salvador et al., 2017; Schnack et al., 2014; Sui et al., 2013; Zhang & Davatzikos, 2013). A few studies have explored sex-specific schizophrenia diagnosis classification (Castellani et al., 2012; Fan et al., 2005, 2007; K. Nakamura et al., 2004; Ota et al., 2012; Takayanagi et al., 2011), with female- and male- specific classifiers reaching overall accuracies of 92% and 91%, respectively (Fan et al., 2007). Furthermore, patients with schizophrenia have been diagnosed (compared to healthy controls) with higher accuracies when training a classifier with an older sample ($\geq$ 40 years; as opposed to a classifier trained with a younger sample < 40 years) (Castellani et al., 2012). Moreover, one study have reached 74% accuracy in diagnosing patients with chronic and children-onset schizophrenia, against healthy controls (Greenstein et al., 2012). Additionally, Rozycki and colleagues used the largest (to the best of my knowledge) multi-site sample reported so far to classify schizophrenia patients, against healthy controls (Rozycki et al., 2018). They showed accuracies in the range of 52% to 81%, being the highest accuracies replicated across sites. Moreover, one study reported an accuracy of 72% in classifying patients diagnosed with a recent onset schizophrenia (i.e. experiencing no longer than 5 years of psychotic

symptoms with minimal antipsychotic exposure), as opposed to healthy controls (Karageorgiou et al., 2011). Finally, a study by Pinaya and colleagues showed that a classifier trained to diagnose patients with chronic schizophrenia against healthy controls was not able to classify FEP patients (showed an accuracy of only 56%) (Pinaya et al., 2016). This suggestively justifies the need for classifiers trained specifically for samples of patients in the early stages of a psychotic disorder (i.e., FEP patients and individuals at an ARMS).

FEP-specific prediction models trained to distinguish FEP patients from healthy controls have shown accuracies between 41% and 88% (Janousova et al., 2015; Kasparek et al., 2011; Mourao-Miranda et al., 2012; Pettersson-Yeo, Benetti, Marquand, Dell'Acqua, et al., 2013; Pina-Camacho et al., 2015; Sun et al., 2009; Vieira et al., 2020; Xiao et al., 2019; Zanetti et al., 2013). Janousava and colleagues trained a male-specific classifier to diagnose FEP patients against healthy controls using different combinations of features and ML algorithms, achieving an accuracy of 81% (Janousova et al., 2015). Furthermore, a few studies have explored the prediction of future illness course from FEP. Patients (at their FEP) with a continuous psychosis at follow-up were already distinguishable from both healthy controls and patients with an episodic psychosis at follow-up (albeit this ones could not be distinguished from healthy controls) with an accuracy of 67% and 70%, respectively (Mourao-Miranda et al., 2012). Moreover, Zanetti and colleagues showed that sMRI data from patients at their FEP could not predict the course of the illness (i.e., remitting *vs.* non-remitting their psychotic symptoms) better than chance level [albeit the sample size used to train and test this classifier was small – 15 subjects per group; (Zanetti et al., 2013)]. Additionally, Pina-Camacho and colleagues showed that FEP patients who developed schizophrenia could be distinguished from healthy controls with an accuracy of 88%, but not from those patients who later developed psychosis other than schizophrenia (i.e. in this case the classification accuracy was not better than chance level) (Pina-Camacho et al., 2015). Lastly, Vieira and colleagues conducted the largest multi-site study (to the best of my knowledge) to distinguish FEP patients from healthy controls (Vieira et al., 2020). In this study, several types of features extracted from sMRI data and several ML algorithms were used to train the FEP classifier, showing overall accuracies between 50% and 70%. Furthermore, the classification models with the highest accuracies showed poor generalization to other sites. Taken together, these findings suggest that development of classification models for the early stages of psychosis is more challenging than originally thought (Vieira et al., 2020).

In the last decade, a few studies have started to explore the potential of using sMRI together with ML to detect individuals at the prodromal phase of psychosis, i.e., individuals at an ARMS. For instance, individuals at an ARMS have been distinguished form FEP patients with an accuracy of 77% (Pettersson-Yeo, Benetti, Marquand, Dell'acqua, et al., 2013). Koutsouleris and colleagues reported an accuracy of 87% or 78% in distinguishing healthy controls from individuals at an ARMS in the early or in the late phase of the prodrome, respectively (Koutsouleris et al., 2009). Furthermore, individuals at an ARMS in different stages of the prodrome (i.e., in the early or in the late phase of the prodrome) could be distinguished with an accuracy of 82% (Koutsouleris et al., 2009). Interestingly, individuals at an ARMS that have later transitioned to psychosis could be distinguished from healthy controls with a remarkable accuracy of 94% (Koutsouleris et al., 2009). Valli and colleagues using a similar sample size reported a lower accuracy in distinguishing individuals at an ARMs from healthy controls (accuracy = 72%) (Valli et al., 2016). The above studies have used a small sample size (30 to 50 subjects in total), which poses a question of the reliability of these models. For instance, De Wit and colleagues, using a modest sample size of 126 subjects, reported an accuracy of only 64% in distinguishing individuals at an ARMS from healthy controls.

Table 1. Studies using structural neuroimaging and machine learning methods to classify schizophrenia patients (against healthy controls; i.e. with diagnosis as the outcome measure).

Study	**Sample** (site/project)	**Data** (features)	**ML approach** (algorithm validation statistical significance)	**Performance metrics**
(K. Nakamura et al., 2004)	Males: 25 HC vs. 30 SCZ Females: 22 HC vs. 27 SCZ (Department of Neuropsychiatry, Toyama University Hospital)	1.5T sMRI (8 regional GMV, WMV, and CSFV)	LDA	ACC = 80% SE = 80%, SP = 80% ACC = 82% SE = 78%, SP = 86%
(Csernansky et al., 2004)	65 HC vs. 52 SCZ	sMRI (thalamic and hippocampal shapes)	LDA LOO CV	ACC = 79% SE = 73%, SP = 83%
(Davatzikos et al., 2005)	79 HC vs. 69 SCZ (SRC)	1.5T sMRI (voxel-based GMD, WMD, and CSFD maps)	nonlinear classifier LOO 1 184 permutations	ACC = 81 % SE = 87, SP = 74 % $p < .001$***
(Fan et al., 2005)	Females: 38 HC vs. 23 SCZ (SCR)	1.5T sMRI (voxel-based GMD, WMD, and CSFD maps)	non-linear SVM LOO CV	ACC = 92%

(Kawasaki et al., 2007)	30 Hc vs. 30 SCZ	1.5T sMRI (voxel-based GMD maps)	LDA	ACC = 90% SE = 90%, SP = 90%
	16 Hc vs. 16 SC (validation sample) (Department of Neuropsychiatry, Toyama University Hospital)			ACC = 84% SE = 88%, SP =81%
(Fan et al., 2007)	Females: 38 HC vs. 23 SCZ	1.5T sMRI (voxel-based brain tissue density maps)	non-linear SVM LOO CV	ACC = 92% SE = 87%, SP = 95%
	Males: 41 HC vs. 46 SCZ (SRC)			ACC = 91% SE = 96%, SP = 85%
(Takayanagi et al., 2011)	Males: 22 HC vs. 29 SCZ	1.5T sMRI (66 regional cortical volumes + 62 regional cortical thickness)	LDA 70%/30% sample (train/test)	ACC = 87% SE = 89%, SP = 83%
	Females: 18 HC vs. 23 SCZ (TMMH)			ACC = 81% SE = 67%, SP = 100%
(Karageorgiou et al., 2011)	47 HC vs. 28 SCZ-recent onset (MCIC)	1.5T (3 sites) and 3T (1 site) sMRI (95 regional whole-brain volumes)	LDA LOO CV χ^2 test	ACC = 72% SE = 64%, SP = 77% $p < .001$***
(Ota et al., 2012)	Females:	1.5T sMRI	LDA	ACC = 73%

	105 HC vs. 38 SCZ	(4 regional GMV and CSFV)		SE = 76 %, SP = 72 %
	Females: 23 HC vs. 23 SCZ (validation sample)			ACC = 72% SE = 74%, SP = 70%
(Nieuwenhuis et al., 2012)	111 HC vs. 128 SCZ (USP)	1.5T sMRI (voxel-based GMD maps)	linear SVM LOO CV 10 000 permutations	ACC = 71% SE = 73%, SP = 69% p < .001***
	122 HC vs. 155 SCZ (validation sample) (GROUP)			ACC = 70% SE = 67%, SP = 74%
(Greenstein et al., 2012)	99 HC vs. 98 SCZ-children onset (NIMH)	1.5T sMRI (68 regional cortical thickness and 6 regional brain volumes)	1 000 Random forests 66%/33% sample (train/test)	ACC = 74%
(Castellani et al., 2012)	54 HC vs. 54 SCZ	1.5T sMRI (visual words extracted from the left DLPFC by SIFT and clustered by k-means)	SVM with local kernel	ACC = 75%
		(visual words extracted from the right DLPFC by SIFT and clustered by k-means)		ACC = 66%
	Females: 25 HC vs. 19 SCZ	(visual words extracted from the left DLPFC by SIFT and clustered by k-means)		ACC = 84%
		(visual words extracted from the right DLPFC by SIFT and clustered by k-means)		ACC = 77%
	Males:	(visual words extracted from the left DLPFC by SIFT and clustered by k-means)		ACC = 60%

	29 HC vs. 35 SCZ			
		(visual words extracted from the right DLPFC by SIFT and clustered by k-means)		ACC = 68%
	Seniors: 23 HC vs. 25 SCZ	(visual words extracted from the left DLPFC by SIFT and clustered by k-means)		ACC = 82%
		(visual words extracted from the right DLPFC by SIFT and clustered by k-means)		ACC = 71%
	Juniors: 31 HC vs. 29 SCZ	(visual words extracted from the left DLPFC by SIFT and clustered by k-means)		ACC = 72%
	(South Verona Psychiatric Case Register)	(visual words extracted from the right DLPFC by SIFT and clustered by k-means)		ACC = 63%
(Zhang & Davatzikos, 2013)	79 HC vs. 69 SCZ (SCR)	1.5T sMRI (voxel-based GMD maps)	ODVBA-TFCE + linear SVM 5-fold nested-CV	ACC = 71%
(Iwabuchi et al., 2013)	20 HC vs. 19 SCZ (Nottinghamshire,UK)	3T sMRI (voxel-based GMV maps)	linear SVM LOO CV 1 000 permutations	ACC = 67% SE = 63%, SP = 70% p = .018*
		(voxel-based WMV maps)		ACC = 64% SE = 58%, SP = 70% p = .022*
		7T sMRI (voxel-based GMV maps)		ACC = 77% SE = 79%, SP = 75 p = .001***

		(voxel-based WMV maps)		ACC = 69% SE = 63%, SP = 75 p = .013*
(Sui et al., 2013)	28 HC vs. 35 SCZ (COBRE)	3T sMRI (voxel-based GMD maps)	linear SVM	ACC = 70%
(Schnack et al., 2014)	66 HC vs. 66 SCZ	1.5T sMRI (voxel-based GMD maps)	SVM 4-fold CV 1 000 permutations	ACC = 90% SE = 92%, SP = 88 p < .001***
	43 HC vs. 46 SCZ (validation sample) (NTR-UMC)	3T sMRI (voxel-based GMD maps)		ACC = 72% SE = 54%, SP = 91% p < .005**
(Radulescu et al., 2014)	20 HC vs. 20 SCZ	1.5T sMRI (voxel-based GM texture maps + regional GMV)	LDA LOO CV	ACC = 65% SE = 65%, SP = 65%
	4 HC vs. 7 SCZ (validation sample) (SlaM)			ACC = 73% SE = 71%, SP = 75%
(Lu et al., 2016)	42 HC vs. 41 SCZ (Guangzhou Brain Hospital)	3T sMRI (voxel-based GMV and WMV difference maps between HC and SCZ)	linear SVM LOO CV 1 000 permutations	ACC = 84% SE = 92%, SP = 84 p < .050*
		(1024 regional GMV and WMV differences between HC and SCZ)		ACC = 81% SE = 72%, SP = 88%

				p < .050*
(Pinaya et al., 2016)	83 HC vs. 143 SCZ (PROESQ)	1.5T sMRI (68 regional CT and 54 regional anatomical volumes)	linear SVM 3-fold CV	BAC = 68% SE = 77%, SP = 59
			DBN-DNN 3-fold CV	BAC = 74% SE = 76%, SP = 71%
(Salvador et al., 2017)	127 HC vs. 128 SCZ (Benito Menni CASM and Mare de De´u de la Mercè hospitals, Spain)	1.5T sMRI (voxel-based GMD maps)	Ridge regression 10-fold CV	ACC = 76%
			Lasso regression 10-fold CV	ACC = 74%
			Elastic net regularization 10-fold CV	ACC = 76%
			L0-norm regularization 10-fold CV	ACC = 75%
			SVM 10-fold CV	ACC = 77%
			Regularized LDA 10-fold CV	ACC = 75%
			Gaussian 48onsensos 10-fold CV	ACC = 76%
			Random forests 10-fold CV	ACC = 75%

(Rozycki et al., 2018)	170 HC vs. 155 SCZ (dataset1)	3T sMRI	ensemble linear SVM (combining a feature-specific classifier by 49onsensos-voting) 10-fold nested-CV	ACC = 70% AUC = .77
	105 HC vs. 75 SCZ (validation sample) (dataset2)	3T sMRI		ACC = 69% AUC = .77
	173 HC vs. 157 SCZ (validation sample) (dataset3)	1.5T sMRI		ACC = 71% AUC = .80
	24 HC vs. 18 SCZ (validation sample) (dataset4)	3T sMRI		ACC = 64% AUC = .59
	29 HC vs. 35 SCZ (validation sample) (dataset5)	3T sMRI (259 regional anatomical measures – intensity, shape, size, and texture)		ACC = 76% AUC = .76
	105 HC vs. 75 SCZ (dataset2)			ACC = 81% AUC = .87
	170 HC vs. 155 SCZ (validation sample) (dataset1)			ACC = 72% AUC = .77
	173 HC vs. 157 SCZ (validation sample) (dataset3)			ACC = 71% AUC = .79
	24 HC vs. 18 SCZ (validation sample)			ACC = 81%

(dataset4)		AUC = .83
29 HC vs. 35 SCZ (validation sample) (dataset5)		ACC = 81% AUC = .93
173 HC vs. 157 SCZ (dataset3)		ACC = 73% AUC = .80
170 HC vs. 155 SCZ (validation sample) (dataset1)		ACC = 70% AUC = .78
105 HC vs. 75 SCZ (validation sample) (dataset2)		ACC = 71% AUC = .79
24 HC vs. 18 SCZ (validation sample) (dataset4)		ACC = 52% AUC = .64
29 HC vs. 35 SCZ (validation sample) (dataset5)		ACC = 81% AUC = 88
24 HC vs. 18 SCZ (validation sample) (dataset4)	using model trained with dataset1 + dataset2 + dataset3	ACC = 74% AUC = .73
29 HC vs. 35 SCZ (validation sample) (dataset5)	using model trained with dataset1 + dataset2 + dataset3	ACC = 77% AUC = .91

(Pinaya et al., 2019)	40 HC vs. 35 SCZ (NUSDAST)	sMRI (68 regional CT and 36 regional anatomical volumes)	deep auto-encoder (trained with 1113 HC from HCP) Mann–Whitney U test (SCZ vs. HC)	mean deviance = 1.14 ± 0.28 (SCZ) p = .001***

Abbreviations: ACC: accuracy; AUC: area under the roc curve; BAC: balanced accuracy; BP bipolar disorder; COBRE: Center for Biomedical Research Excellence; CT: cortical thickness; CV: cross-validation; DBN-DNN: deep neural network initialized by a deep belief network; DLPFC: dorso-lateral prefrontal cortex; GMD: grey matter density; GMV: gray matter volume; GROUP: Genetic Risk and Outcome of Psychosis; HC: healthy controls; LDA: linear discriminant analysis; LOO: leave-one-out; NIMH: National Institute of Mental Health; NUSDAST: Northwestern University Schizophrenia Data and Software Tool; ODVBA-TFCE: Optimally-Discriminative Voxel-Based Analysis based Threshold-Free Cluster Enhancement; PROESQ: Schizophrenia Program at the Universidade Federal de São Paulo; SCZ: schizophrenia patients; SE: sensitivity; SIFT: scale invariance feature transform; SlaM: South London and Maudsley National Health System Trust, London; sMRI: structural magnetic resonance imaging; SP: specificity; SRC: Schizophrenia Research Center, University of Pennsylvania Medical Center; SVM: support vector machines; TMMH: Tokyo Metropolitan Matsuzawa Hospital; WMD: white matter density; WMV: white matter volume; * $p < .050$; ** $p < .010$; *** $p < .001$.

Table 2. Studies using structural neuroimaging and machine learning methods to classify first episode psychosis patients (against healthy controls; i.e. with diagnosis as the outcome measure).

Study	**Sample** (site/project)	**Data** (features)	**ML approach** (algorithm validation statistical significance)	**Performance metrics**
(Sun et al., 2009)	36 HC vs. 36 FEP (ABBRC)	1.5T sMRI (cortical GMD maps)	Sparse multinominal logistic regression LOO CV 1 000 permutations	ACC = 86% SE = 86%, SP = 86% p < .001***
(Kasparek et al., 2011)	39 HC vs. 38 FEP (Department of Psychiatry, Masaryk University in Brno)	1.5T sMRI (voxel-based whole-brain intensity maps)	maximum-uncertainty LDA LOO CV	ACC = 72% SE 67%, SP = 77%
(Mourao-Miranda et al., 2012)	28 HC vs. 28 FEP-C	1.5T sMRI (voxel-based GMV maps)	linear SVM LPO nested-CV 1 000 permutations	ACC = 67% SE = 71%, SP = 61% p = .010*
	28 HC vs. 28 FEP-E			ACC = 54% SE = 64%, SP = 43 p = .300
	28 FEP-C vs. 28 FEP-E (AESOP)			ACC = 70% SE = 71%, SP = 68% p = .004**
(Pettersson-Yeo, Benetti, Marquand, Dell'Acqua, et al., 2013)	19 HC vs. 19 FEP (OASIS)	3T sMRI (voxel-based GMV maps)	linear SVM LOO CV 1 000 permutations	ACC = 63% SE = 58%, SP = 68% p = .066
(Zanetti et al., 2013)	62 HC vs. 62 FEP	1.5T sMRI (voxel-based GMD, WMD, and CSFD maps)	non-linear SVM LOO CV	ACC = 73% SE = 79%, SP = 68%
	57 HC vs. 44 FEP-NR (São Paulo, Brasil)			ACC = 64% SE = 52%, SP = 74%
(Pina-Camacho et al., 2015)	34 HC vs. 39 FEP-SSD	1.5T sMRI (221 regional GMV, WMV, and CSFV)	linear SVM LOO CV permutations	ACC = 88% SE = 87%, SP = 88% p < .010**
			linear SVM Jacknife	ACC = 67% SE = 65%, SP = 69%

			permutations	p < .010**
	8 HC vs. 10 FEP-SSD (validation sample)			ACC = 61% SE = 40%, SP = 88%
	34 HC vs. 26 FEP-non-SSD		linear SVM LOO CV permutations	ACC = 87% SE = 88%, SP = 85% p > .050
			linear SVM Jacknife Permutations	ACC = 59% SE = 49%, SP = 50% p > .050
	8 HC vs. 6 FEP-non-SSD (validation sample)			ACC = 61% SE = 60%, SP = 63%
	39 FEP-SSD vs. 26 FEP-non-SSD		linear SVM LOO CV permutations	ACC = 60% SE = 67%, SP = 50% p > .050
			linear SVM Jacknife Permutations	ACC = 60% SE = 67%, SP = 50% p > .050
	10 FEP-SSD vs. 6 FEP-non-SSD (validation sample) (CAFEPS)			ACC = 50% SE = 40%, SP = 67%
(Janousova et al., 2015)	Males: 49 HC vs. 49 FEP (Department of Psychiatry, Masaryk University in Brno)	1.5T sMRI (voxel-based whole-brain intensity maps)	maximum-uncertainty LDA LOO CV	ACC = 65% SE = 80%, SP = 51%
			centroid method LOO CV	ACC = 61% SE = 59%, SP = 63%
			average linkage LOO CV	ACC = 51% SE = 18, SP = 90
		(voxel-based GMD maps)	maximum-uncertainty LDA LOO CV	ACC = 79% SE = 82%, SP = 76%
			centroid method LOO CV	ACC = 64% SE = 63%, SP = 65%
			average linkage LOO CV	ACC = 69% SE = 49%, SP = 90%
		(voxel-based deformation maps)	maximum-uncertainty LDA LOO CV	ACC = 77% SE = 86%, SP = 67%
			centroid method LOO CV	ACC = 81% SE = 71%, SP = 90%

			average linkage LOO CV	ACC = 67% SE = 35%, SP = 100%
		(voxel-based whole-brain intensity, GMD, deformation maps)	maximum-uncertainty LDA, centroid method and average linkage (different combinations of features and ML algorithms by majority voting; 256 different combinations tested) LOO CV	average ACC = 73% [54% 82%] average SE = 65% [18% 86%], average SP = 82% [51% 100%]
(Xiao et al., 2019)	163 HC vs. 163 FEP	3T sMRI (68 regional CT)	SVM 10-fold CV 1 000 permutations	ACC = 82% SE = 77%, SP = 85% p < .001***
		3T sMRI (68 regional surface area)	SVM 10-fold CV 1 000 permutations	ACC = 85% SE = 83%, SP = 87% p < .001***
(Vieira et al., 2020)	111 HC vs. 111 FEP (Chengdu, China)	3T sMRI (voxel-based GMV maps)	nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 61% SE = 50%, SP = 72% p = .001**
			elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 60% SE = 63%, SP = 59% p = .003**
			linear SVM 10-fold nested-CV 1 000 permutations	BAC = 61% SE = 63%, SP = 59% p = .004**
			DNN 10-fold nested-CV 1 000 permutations	BAC = 58% SE = 60%, SP = 56% p = .008**
		3T sMRI (voxel-based CT maps)	nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 62% SE = 72%, SP = 52% p = .002**
			elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 67% SE = 66%, SP = 69% p = .001**
			linear SVM 10-fold nested-CV 1 000 permutations	BAC = 53% SE = 25%, SP = 97% p = .013*
			DNN	BAC = 66%

		10-fold nested-CV 1 000 permutations	SE = 64%, SP = 68% p = .001**
	3T sMRI (169 surfaced-based regional volumes and CT)	nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 61% SE = 74%, SP = 47% p = .003**
		elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 62% SE = 65%, SP = 60% p = .003**
		linear SVM 10-fold nested-CV 1 000 permutations	BAC = 61% SE = 66%, SP = 56% p = .003**
		DNN 10-fold nested-CV 1 000 permutations	BAC = 71% SE = 72%, SP = 69% p = .001**
71 HC vs. 71 FEP (GAP)	3T sMRI (voxel-based GMV maps)	nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 44% SE = 34%, SP = 54% p = .891
		elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 52% SE = 54%, SP = 50% p = .346
		linear SVM 10-fold nested-CV 1 000 permutations	BAC = 54% SE = 53%, SP = 54% p = .207
		DNN 10-fold nested-CV 1 000 permutations	BAC = 41% SE = 47%, SP = 34% p = .593
	3T sMRI (voxel-based CT maps)	nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 54% SE = 38%, SP = 69% p = .200
		elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 62% SE = 63%, SP = 60% p = .009**
		linear SVM 10-fold nested-CV 1 000 permutations	BAC = 51% SE = 96%, SP = 6% p = .450
		DNN 10-fold nested-CV 1 000 permutations	BAC = 53% SE = 52%, SP = 55% p = .265

		3T sMRI (169 surfaced-based regional volumes and CT)	nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 57% SE = 51%, SP = 63% p = .083
			elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 52% SE = 45%, SP = 58% p = .381
			linear SVM 10-fold nested-CV 1 000 permutations	BAC = 46% SE = 49%, SP = 43% p = .757
			DNN 10-fold nested-CV 1 000 permutations	BAC = 59% SE = 50%, SP = 68% p = .014*
110 HC vs. 110 FEP (Santander A, Spain)		3T sMRI (voxel-based GMV maps)	nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 51% SE = 32%, SP = 69% p = .444
			elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 63% SE = 64%, SP = 63% p = .002**
			linear SVM 10-fold nested-CV 1 000 permutations	BAC = 66% SE = 68%, SP = 64% p = .001**
			DNN 10-fold nested-CV 1000 permutations	BAC = 50% SE = 53%, SP = 64% p = .448
		3T sMRI (voxel-based CT maps)	nearest neighbors 10-fold nested-CV 1000 permutations	BAC = 58% SE = 50%, SP = 66% p = .011*
			elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 59% SE = 58%, SP = 60% p = .013*
			linear SVM 10-fold nested-CV 1 000 permutations	BAC = 52% SE = 91%, SP = 13% p = .036*
			DNN 10-fold nested-CV 1 000 permutations	BAC = 60% SE = 60%, SP = 59% p = .010*
		3T sMRI (169 surfaced-based regional volumes and CT)	nearest neighbors 10-fold nested-CV	BAC = 60% SE = 46%, SP = 74%

		1 000 permutations	p = .004**
70 HC vs. 140 FEP (Santander B, Spain)	1.5T sMRI (voxel-based GMV maps)	elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 59% SE = 58%, SP = 59% p = .021*
		linear SVM 10-fold nested-CV 1 000 permutations	BAC = 61% SE = 62%, SP = 59% p = .005**
		DNN 10-fold nested-CV 1 000 permutations	BAC = 70% SE = 70%, SP = 70% p = .001**
		nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 59% SE = 71%, SP = 47% p = .003**
		elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 60% SE = 58%, SP = 61% p = .012*
		linear SVM 10-fold nested-CV 1 000 permutations	BAC = 57% SE = 72%, SP = 43% p = .032*
		DNN 10-fold nested-CV 1 000 permutations	BAC = 59% SE = 81%, SP = 37% p = .014*
	1.5T sMRI (voxel-based CT maps)	nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 60% SE = 68%, SP = 51% p = .028*
		elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 63% SE = 57%, SP = 62% p = .001**
		linear SVM 10-fold nested-CV 1 000 permutations	BAC = 58% SE = 72%, SP = 53% p = .030*
		DNN 10-fold nested-CV 1 000 permutations	BAC = 59% SE = 62%, SP = 53% p = .002**
	1.5T sMRI (169 surfaced-based regional volumes and CT)	nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 57% SE = 92%, SP = 21% p = .041*

		elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 55% SE = 74%, SP = 36% p = .129
		linear SVM 10-fold nested-CV 1 000 permutations	BAC = 56% SE = 65%, SP = 47% p = .081
		DNN 10-fold nested-CV 1 000 permutations	BAC = 62% SE = 77%, SP = 47% p = .001**
81 HC vs. 81 FEP (GROUP)	1.5T sMRI (voxel-based GMV maps)	nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 55% SE = 34%, SP = 75% p = .163
		elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 61% SE = 57%, SP = 66% p = .003**
		linear SVM 10-fold nested-CV 1 000 permutations	BAC = 62% SE = 63%, SP = 62% p = .004**
		DNN 10-fold nested-CV 1 000 permutations	BAC = 58% SE = 58%, SP =58% p = .010*
	1.5T sMRI (voxel-based CT maps)	nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 52% SE = 37%, SP = 68% p = .262
		elastic net regularization 10-fold nested-CV 1 000 permutations	BAC = 61% SE = 61%, SP = 60% p = .007**
		linear SVM 10-fold nested-CV 1 000 permutations	BAC = 56% SE = 51%, SP = 61% p = .408
		DNN 10-fold nested-CV 1 000 permutations	BAC = 60% SE = 56%, SP = 64% p = .008**
	1.5T sMRI (169 surfaced-based regional volumes and CT)	nearest neighbors 10-fold nested-CV 1 000 permutations	BAC = 53% SE = 54%, SP = 52% p = .237
		elastic net regularization	BAC = 59%

	10-fold nested-CV 1 000 permutations	SE = 62%, SP = 55% p = .033*
	linear SVM 10-fold nested-CV 1 000 permutations	BAC = 61% SE = 60%, SP = 62% p = .007**
	DNN 10-fold nested-CV 1 000 permutations	BAC = 55% SE = 59%, SP = 52% p = .108

Abbreviations: ABBRC: Aftercare Research Program and Adolescent Brain-Behavior Research Clinic; ACC: accuracy; BAC: balanced accuracy; CSFD: cerebrospinal fluid density; CSFV: cerebrospinal fluid volume; CT: cortical thickness; CV: cross-validation; DNN: deep neural network; FEP: first episode of psychosis; FEP-C: FEP patients with continuous psychosis at follow-up; FEP-E: FEP patients with episodic psychosis at follow-up; FEP-non-SSD: FEP patients that developed psychosis other than the schizophrenia spectrum disorders; FEP-NR: FEP patients that did not remitted their symptoms at follow-up; FEP-R: FEP patients that remitted their symptoms at follow-up; FEP-SSD: FEP patients that developed psychosis in the schizophrenia spectrum disorders; GAP: Genetic and Psychosis study; GMD: grey matter density; GMV: gray matter volume; GROUP: Genetic Risk and Outcome of Psychosis; HC: healthy controls; LDA: linear discriminant analysis; LOO: leave-one-out; LPO: leave-one-pair-out; SE: sensitivity; sMRI: structural magnetic resonance imaging; SP: specificity; SVM: support vector machines; WMD: white matter density; WMV: white matter volume; * $p < .050$; ** $p < .010$; *** $p < .001$.

Table 3. Studies using structural neuroimaging and machine learning methods to classify patients at an at-risk mental state (against healthy controls, first episode of psychosis patients or in different stages of the prodrome; i.e. with diagnosis as the outcome measure).

Study	**Sample** (site/project)	**Data** (features)	**ML approach** (algorithm validation statistical significance)	**Performance metrics**
(Koutsouleris et al., 2009)	25 HC vs. 20 ARMS-E vs. 25 ARMS-L	1.5T sMRI (voxel-based GMV maps)	non-linear SVM 5-fold CV 5 000 permutations	ACC = 86% p<.001***
	25 HC vs. 20 ARMS-E			ACC = 87% SE = 95%, SP = 80% p < .001***
	25 HC vs. 25 ARMS-L			ACC = 78% SE = 76%, SP = 80% p < .001***
	20 ARMS-E vs. 25 ARMS-L			ACC = 82% SE = 84%, SP = 80% p < .001***
	17 HC vs. 15 ARMS-T vs. 18 ARMS-NT			ACC = 82% p < .001***
	17 HC vs. 15 ARMS-T			ACC = 94%, SE = 100%, SP = 88% p < .001***
	17 HC vs. 18 ARMS-NT (FETZ)			ACC = 86% SE = 78%, SP = 94% p < .001***
(Pettersson-Yeo, Benetti, Marquand, Dell'Acqua, et al., 2013)	15 ARMS vs. 15 FEP (OASIS)	3T sMRI (voxel-based GMV maps)	linear SVM LOO CV 1 000 permutations	ACC = 77% SE = 80%, SP = 73% p = .001**
(Valli et al., 2016)	25 HC vs. 25 ARMS (OASIS)	1.5T sMRI (voxel-based GMV maps)	linear SVM LOO CV 1 000 permutations	ACC = 72% SE = 68%, SP = 76% p < .001***
(de Wit et al., 2017)	62 HC vs. 64 ARMS (BCRM)	1.5T sMRI (30 regional subcortical GMV)	linear SVM LOO CV 1 000 permutations	ACC = 64% SE = 59%, SP = 68% p = .018*
		1.5T sMRI (70 regional surface areas)		ACC = 61% SE = 66%, SP = 56%

		p = .005** (note: n.s. when matched for age and gender)
	1.5T sMRI (68 regional CT)	ACC = 60% SE = 56%, SP = 63% p = .007** (note: n.s. when matched for age and gender)
	1.5T sMRI (75 regional cortical GMV)	ACC = 55% SE = 53%, SP = 56% p > .050
	1.5T sMRI (68 regional gyrification indexes)	ACC = 58% SE = 70%, SP = 46% p > .050

Abbreviations: ACC: accuracy; ARMS: at-risk mental state; ARMS-E: individual at ARMS in the early phase of the prodrome; ARMS-L: individual at ARMS in the late phase of the prodrome; ARMS-NT: individual at an ARMS that did not transitioned to psychosis; ARMS-T: individual at ARMS that transitioned to psychosis; BCRM: Department of Psychiatry at the University Medical Center Utrecht, Brain Center Rudolf Magnus in the Netherlands; CT: cortical thickness; CV: cross-validation; FEP: first episode of psychosis; FETZ: Early Detection and Intervention Centre for Mental Crises, Ludwig-Maximilians-University; GMV: gray matter volume; HC: healthy controls; LOO: leave-one-out; OASIS: Outreach and Support in South London service in London, United Kingdom; SE: sensitivity; sMRI: structural magnetic resonance imaging; SP: specificity; SVM: support vector machines; * $p < .050$; ** $p < .010$; *** $p < .001$.

1.4. Genetics in psychosis

1.4.1. Potential genetic biomarkers

The heritability (i.e. the degree of variation in a phenotypic trait in a population attributable to genetic variation among that population) of psychotic disorders is estimated to be around 80% (Cardno et al., 1999; Sullivan et al., 2003), meaning that genetic variation may play an important role to the development of psychosis. It has been extensively shown that psychotic disorders are multifactorial and stem from multiple genetic variants, such as single nucleotide polymorphisms (SNPs) and copy number variations (CNVs). SNPs are common variants (i.e., with an observed frequency > 1% in the population) of individual nucleotide sequence. CNVs are rare variants (i.e. with an observed frequency < 1% in the population) involving alterations in the number of copies – i.e. deletions or duplications – of specific regions (i.e. sets of nucleotides) of the Deoxyribonucleic Acid (DNA) chain.

Several SNPs have already been identified as conferring risk to the development of schizophrenia (the most common psychotic disorder) through genome-wide association studies (GWASs). GWAS is an observational study that look for associations between genetic variants, usually SNPs, throughout the whole-genome and a phenotypic trait by assessing the prevalence of those variants in cases compared to controls. Then, genes that co-segregate with the SNPs identified by GWAS are putatively associated with the trait. GWAS represents a powerful and a hypothesis-free approach to assess the genetic architecture of complex traits. However, each of these common variants confer low relative risk of having the trait of interest and, therefore, to achieve enough statistical power to find those variants large sample sizes are a requirement in GWAS.

The Psychiatric Genome Wide Association Study Consortium (PGC) was the first largest and most innovative consortium created to bring larger genetic samples (in the order of thousands) by joining forces of many specialists worldwide (Ripke et al., 2011). In 2014, the Schizophrenia Working Group of PGC found 108 independent SNPs (from 128 genome-wide significant associations) with genome-wide relevance for schizophrenia using a sample of 36 989 cases and 113 075 controls. An enrichment analysis of the genes that were mapped to the 108 SNPs (i.e. the identification of classes of genes that are over-represented in the original set of genes and, therefore, may have an association with schizophrenia) found enrichment in brain tissue enhancers (i.e. a short sequence of DNA that increase the likelihood of the transcription of a particular gene) and in tissues with important roles in immunity. Furthermore, the most significant associations were found at promising targets

for treatment [e.g. DRD2 (Dopamine receptor D_2) – the site of action of antipsychotic drugs]; genes involved in glutamatergic neurotransmission [e.g. GRM3 (Metabotropic glutamate receptor 3), GRIN2A (Glutamate receptor subunit epsilon-1), SRR (serine racemase), CLCN3 (chloride channel 3), and GRIA1 (glutamate receptor 1)]; genes involving neuronal calcium signaling [e.g. CACNA1C (calcium channel, voltage-dependent, L type, alpha 1C subunit), CACNA1I (calcium channel, voltage-dependent, T type, alpha 1I subunit), CACNB2 (calcium channel, voltage-dependent, L-type, beta 2 subunit), and RIMS1 (regulating synaptic membrane exocytosis protein 1)]; genes involved in broader synaptic function [e.g. KCTD13 (BTB/POZ domain-containing protein KCTD13), CNTN4 (contactin-4), and PAK6 (serine/threonine-protein kinase PAK 6)]; and genes encoding proteins that play a central role in the immune system by presenting antigen-derived peptides for recognition by CD4+ T lymphocytes [MHC (major histocompatibility complex)] (Foley et al., 2017). More recently, a meta-analysis of the CLOZUK (genome-wide genotype information for schizophrenia cases from the UK) and independent PGC datasets (with the exclusion of related and overlapping samples; in a total of 40 675 cases and 64 643 controls) identified 179 independent genome-wide significant SNPs mapping to 145 independent loci (Pardiñas et al., 2018). Ninety-three of the 145 loci were also found to be genome-wide significant in the previous PGC study (Ripke et al., 2014), with the majority showing a strengthened association. Additionally, gene enrichment analysis has shown that mutation-intolerant genes (i.e., loss-of-function genes) are more enriched for these associations with schizophrenia compared to all other annotated genes. Furthermore, these genes were shown to account for 30% of the SNP-based heritability for schizophrenia (Pardiñas et al., 2018). Moreover, in a recent systematic review of all independent GWAS i.e. non-overlapping samples within each article, between articles, and with those of the previous review (Lee et al., 2012), the following genes were found to be associated with schizophrenia in at least two GWASs: two genes widely expressed in the brain [AMBRA1 (autophagy/beclin-1 regulator), and UGT1A1 (UDP-glucuronosyltransferase 1-1)]; genes from the MHC locus; a gene which is part of the genetic network that maintains circadian rhythm [ARNTL (aryl hydrocarbon receptor nuclear translocator-like protein 1)]; a gene part of the mechanisms involving neuronal calcium signaling [EFHD1 (EF-hand domain family, member D1)]; and two genes involved in broader synaptic function [CDH13 (T-Cadherin), and PLXNA2 (Plexin-A2)].

Despite the large sample sizes used in the large-scale GWASs conducted thus far, the case control numbers seems to remain too low to encompass the entire amount of genetic variability that accounts for the schizophrenia heritability (Cardno et al., 1999; Sullivan et al., 2003). In specific, the statistical power of GWAS is still not enough to catch genome-wide significant SNP associations that are less frequent and with very small effect sizes. Moreover, the genome-wide SNP heritability has been shown to be uniform with a likelihood of more than 71% of genomic regions of 1 mega bases range containing at least one risk variant for schizophrenia (Loh et al., 2015). CNVs, on the other hand, are highly penetrant, usually affecting a single gene or a defined subset of genes. The association of these variants with the phenotypic trait are usually very strong (i.e. with high effect sizes) and, therefore, are hypothesized to have an important contribution to the heritability puzzle of schizophrenia. One major problem of this hypothesis is that CNVs are rare variations in the genome, i.e. very small cases of patients with schizophrenia have indeed these variations, and, therefore, they cannot fully explain its heritability. The most common and most penetrant CNV that has been associated with schizophrenia is the 22q11.2 deletion (Kirov et al., 2014; Murphy et al., 1999). This deletion encompasses approximately 3 million bases and causes velocardiofacial syndrome – i.e. with pharyngeal dysfunction, cardiac anomaly, and dysmorphic facies – or Di George syndrome – i.e. with cardiac anomalies, hypoparathyroidism, and immunodeficiency (Kobrynski & Sullivan, 2007). Furthermore, approximately 30% of individuals with this deletion have been shown to meet the criteria for diagnosis of schizophrenia in early adulthood (Murphy et al., 1999) and 0.5% of individuals with schizophrenia have been shown to have a 22q11.2 deletion (Kirov et al., 2014). Additionally, the frequency of a few more CNVs [e.g. deletions: NRXN1 (neurexin-1-alpha), 15q11.2, 1q21.1, 15q13.3; duplications: CHRNA7 (neuronal acetylcholine receptor subunit alpha-7), 16p13.11, 16p11.2] have been found to be higher in patients with schizophrenia than in controls, but with a smaller effect size on the risk of having the disease (Kirov et al., 2014).

It is evident that schizophrenia has a polygenic architecture, with possibly thousands of genomic variants collectively contributing to the disease risk (Gratten et al., 2014). The polygenic risk score (PRS) is an analytical approach that tries to achieve a more meaningful prediction of the risk by collectively summarizing all the common variants that have been previously associated with the disease in a GWAS. In particular, it is computed as the sum of all alleles that confer risk to the disease weighted by their effect size (i.e. odds rations).

The number of variants included in the PRS depend on the *p*-value threshold applied to the statistical significance of each association in the GWAS, i.e. for higher *p*-value thresholds, the number of variants included in the PRS decreases (Wray et al., 2014). PRS have been shown to distinguish controls from schizophrenia (Ripke et al., 2014) and FEP patients (Vassos et al., 2017), with PRS explaining between 3% and 18% or 2% and 9% of the variance, respectively, depending on which *p*-value threshold is used. Furthermore, among FEP patients, PRS was able to distinguish those who further developed schizophrenia from those who developed other psychotic disorders (with 9% of the variance explained) (Vassos et al., 2017). Moreover, the majority of the PRS studies are based on samples of European ancestry and the reliability of the findings for other ancestries are not yet ascertained (Curtis, 2018; Vassos et al., 2017).

Transcriptome-wide association studies (TWASs) are a recent approach in psychiatric research (Hernandez et al., 2020). They integrate GWAS and gene expression [as measured by levels of messenger ribonucleic acid (mRNA)] datasets to identify gene-trait associations. Loci identified by GWAS are mostly localized in noncoding regions (i.e. regions of the DNA that do not encode protein sequences) with unknown functional role. TWAS tackles this issue by mapping genomic loci to the expression of a given gene (i.e. selecting loci that have a functional effect on gene expression levels), named expression quantitative trait loci (eQTL). This approach requires large datasets of concomitant genomic and transcriptomic information and works under the assumption that gene expression mediates the effect of genetic variation on a trait (Wainberg et al., 2019). TWAS have been used to prioritize candidate risk genes. For example, gene expression in the dorsolateral prefrontal cortex has been shown to be influenced by around 20% of schizophrenia-associated variants (Fromer et al., 2016) using data from the CommonMind Consortium. Fromer and colleagues further showed that a) altering expression in three genes whose expression was influenced by one single variant [FURIN, TSNARE1 (t-SNARE domain-containing protein 1), and CNTN4] changed neurodevelopment in zebrafish; and b) co-expressed genes belonged to networks involved in neurobiological functions previously suggested to be altered in schizophrenia (e.g. synaptic function, glutamatergic neurotransmission).

1.4.2. Individual prediction of psychosis using genetics and machine learning

Notwithstanding the large evidence that schizophrenia has a polygenetic architecture, with many genetic variants contributing to the risk development of the disease, only a handful of

studies have explored the potential of combining this modality with ML methods for individual-level predictions (Aguiar-Pulido et al., 2010; Antonucci et al., 2020; Chen et al., 2018; Pettersson-Yeo, Benetti, Marquand, Dell'Acqua, et al., 2013; Struyf et al., 2008; Vivian-Griffiths et al., 2019; Yang et al., 2010). All studies focused on building prediction models for classifying patients in a prodromal phase (i.e. ARMS) or with a recent-onset (i.e. FEP) or chronic (i.e. schizophrenia) psychotic disorder versus healthy controls (**Table 4**). The literature review regarding the prediction of disease progression from an ARMS will be described in **5. Individual prediction of transition to psychosis using genetics and machine learning**.

Gene expression profiling in the frontal brain region has been shown to distinguish chronic schizophrenia patients from healthy controls with an area under the receiver operating characteristic curve (AUC) of 0.91. This was achieved using a sample size of 112 healthy controls and 115 patients and SVM (Struyf et al., 2008). All the other studies have used a hypothesis-based selection of SNPs as features in their prediction models. For instance, Yang and colleagues reported an accuracy of 74% in distinguishing schizophrenia patients from healthy controls using a sample composed by 20 healthy controls and 20 patients, a set of 367 SNPs from 222 genes derived from six physiological systems (neurobiology, metabolism, cell proliferation, cardiovascular, inflammation, and cholesterol biochemistry) and SVM (Yang et al., 2010). However, smaller accuracies have been reported when larger sample sizes have been used (Aguiar-Pulido et al., 2010; Antonucci et al., 2020; Vivian-Griffiths et al., 2019), but the opposite has also been reported (Chen et al., 2018). Aguiar-Pulido and colleagues used a set of 48 SNPs from two previously psychosis-associated genes (DRD3 – Dopamine 3 receptors – and HTR2A – serotonin receptors) of a sample composed by 354 healthy controls and 260 patients with schizophrenia to train several classifiers with different ML algorithms. The reported accuracies were between 59% and 66%. Antonucci and colleagues computed risk scores based on the 116 psychosis-associated SNPs reported in the PGC2 from a sample of 337 healthy controls and 103 patients with schizophrenia and trained a classifier with SVM. They have reported an accuracy of 56%, but it was not replicated in an independent sample of 102 healthy controls and 75 patients. Moreover, one of the largest study exploring genetics and ML to classify schizophrenia reported AUC between 0.63 and 0.69 when testing several prediction models trained with different features sets (top 125 SNPs, top 4998 SNP or PRS for schizophrenia using the findings from the PGC2) and different ML algorithms (logistic regression, linear, or non-linear SVM) (Vivian-

Griffiths et al., 2019). The other largest study using genetics to distinguish patients with schizophrenia from healthy controls at an individual level was conducted by Chen and colleagues and, using a set of 116 PRSs for schizophrenia and comorbid conditions and traits and a combination of regularized regression (LASSO) and DNN algorithms, reported a model and a replication accuracies of 82% and 71%, respectively (Chen et al., 2018). Lastly, Pettersson-Yeo and colleagues reported an accuracy of 68% in distinguishing FEP patients from healthy controls using a small set of SNPs (around 20 previously psychosis-associated SNPs) from 19 healthy controls and 19 FEP patients and SVM. Prediction models were also trained to classify individuals at an ARMS (n = 19) against healthy controls or FEP patients, but their accuracy was not better than chance level.

Table 4. Studies using genetics and machine learning methods to classify individuals at an at-risk mental state, first episode psychosis or schizophrenia patients (against healthy controls) or individuals at an at-risk mental state against first episode of psychosis patients; i.e. with diagnosis as the outcome measure.

Study	**Sample** (site/project)	**Data** (features)	**ML approach** (algorithm validation statistical significance)	**Performance metrics**
(Struyf et al., 2008)	112 HC vs. 115 SCZ (Stanley Neuropathology Consortium)	gene expression from 22283 probe sets	SVM 10-fold CV	AUC = 0.91
			nearest shrunken centroids 10-fold CV	AUC = 0.71
			decision trees 10-fold CV	AUC = 0.64
			ensemble of voters 10-fold CV	AUC = 0.71
			naïve Bayes 10-fold CV	AUC = 0.71
			nearest neighbors 10-fold CV	AUC = 0.70
(Yang et al., 2010)	20 HC vs. 20 SCZ (Hartford hospital)	367 SNPs	linear SVM LOO CV	ACC = 74% SE = 72%, SP = 76%
(Aguiar-Pulido et al., 2010)	354 HC vs. 260 SCZ (Galicia, Spain)	48 SNPs from the DRD3 (17 SNPs) and HTR2A (31 SNPs) genes	linear neural networks 10-fold CV	ACC = 65%
			multilayer perceptron 10-fold CV	ACC = 65%
			radial base functions 10-fold CV	ACC = 63%
			evolutionary computation 10-fold CV	ACC = 59%
			multifactor dimensionality reduction 10-fold CV	ACC = 64%
			Bayes networks 10-fold CV	ACC = 61%
			naïve Bayes 10-fold CV	ACC = 65%
			SVM	ACC = 65%

			10-fold CV	
			decision tables 10-fold CV	ACC = 62%
			decision table I bayes hybrid classifier 10-fold CV	ACC = 66%
			best-first decision tree 10-fold CV	ACC = 63%
			adaptative boosting 10-fold CV	ACC = 66%
(Pettersson-Yeo, Benetti, Marquand, Dell'Acqua, et al., 2013)	19 HC vs. 19 FEP	20 SNPs	linear SVM LOO CV 1000 permutations	ACC = 68% SE = 71%, SP = 64% p = .031*
	19 HC vs. 19 ARMS	20 SNPs		ACC = 53% SE = 53%, SP = 53% p = .457
	15 ARMS vs. 15 FEP (OASIS)	19 SNPs		ACC = 33% SE = 42%, SP = 25% p = .927
(Chen et al., 2018)	6489 HC vs. 5576 SCZ (MGS + SSCSS)	116 PRS for SCZ and from genetically related conditions and traits	LASSO + DNN 90%/10% sample (train/test)	ACC = 82% AUC = 0.91
	751 HC vs. 741 SCZ (validation sample) (CATIE)			ACC = 71% AUC = 0.75
(Vivian-Griffiths et al., 2019)	6299 HC vs. 5554 SCZ (CLOZUK)	PRS for SCZ using top 125 SNPs from the PGC2	logistic regression 75%/25% sample (train/test)	AUC = 0.64
		top 125 SNPs from the PGC2	linear SVM 75%/25% sample (train/test)	AUC = 0.64
			non-linear SVM 75%/25% sample (train/test)	AUC = 0.63
		PRS for SCZ using top 4998 SNPs from the PGC2	logistic regression 75%/25% sample (train/test)	AUC = 0.69
		top 4998 SNPs from the PGC2	linear SVM 75%/25% sample (train/test)	AUC = 0.63
			non-linear SVM 75%/25% sample (train/test)	AUC = 0.66

(Antonucci et al., 2020)	337 HC vs. 103 SCZ	risk scores of each of the 116 SNPs reported in the PGC2	linear SVM 10-fold nested-CV 1000 permutations	BAC = 56% SE = 46%, SP = 65% p = .050*
	102 HC vs. 75 SCZ (validation sample)			BAC = 50% SE = 40%, SP = 60%

Abbreviations: ACC: accuracy; ARMS: at-risk mental state; AUC: area under de receiver operating characteristic curve; DNN: deep neural network; CATIE: clinical antipsychotic trials for intervention effectiveness; CV: cross-validation; FEP: first episode of psychosis; HC: healthy controls; LASSO: least absolute shrinkage and selection operator; LOO: leave-one-out; MGS: molecular genetics of schizophrenia; OASIS: Outreach and Support in South London service in London, United Kingdom; PGC2: psychiatric genomics consortium; PRS: polygenic risk score; SCZ: schizophrenia patients; SNP: single nucleotide polymorphism; SSCSS: Swedish schizophrenia case control study; SVM: support vector machines; * $p < .050$; ** $p < .010$; *** $p < .001$.

1.5. Environment in psychosis

1.5.1. Potential environmental biomarkers

Environmental risk factors have been extensively shown to contribute to the increase of the risk of schizophrenia (Jaaro-Peled & Sawa, 2020; Schmitt et al., 2014). Early life factors (obstetric complication, prenatal and postnatal infection, and childhood environment) have been proposed to lead to vulnerability to the disease at childhood, whereas later life factors (substance abuse, migration, urbanicity, ethnic minority, and stressful life events) may induce the onset of the disease (Keshavan & Hogarty, 1999).

Being born in winter or early spring has been shown to increase the risk of developing schizophrenia with a population attributable risk of 3.3% in Northern Hemispheres sites (Davies et al., 2003). One potential explanation for this association is the exposure to infectious agents in uterus, suggesting a role of the maternal immune system. Indeed, exposure to rubella, genital or reproductive infections, influenza during the first half of pregnancy, toxoplasmosis, respiratory infections or herpes simplex virus 1 have been shown to increase the risk for schizophrenia 1.5 to 5.3 times (Brown, 2011). Additionally, obstetric complications such as bleeding, preeclampsia, diabetes, rhesus compatibility, abnormal fetal growth (low birth weight, congenital malformations, and small head circumference), and complication in delivery (emergency cesarean section, uterine atony or asphyxia) have been shown to increase the risk of schizophrenia by 1.5 to 2.0 fold (M. Cannon et al., 2002). Furthermore, a history of obstetric complications has been shown to increase the transition to psychosis rate among individuals at an ARMS by 4.9 to 6.6 fold (Kotlicka-Antczak et al., 2018; Mittal et al., 2009). Moreover, advanced paternal age has been associated with an increased risk for schizophrenia by 1.7 fold (B. Miller et al., 2011). This association may be explained by the increase rate of the novo mutations with advanced paternal age (Crow, 1997).

Urbanicity, migration, childhood adversity and recent stressful events are potential stress-inducing factors, and all have been related to schizophrenia. In particular, living in a higher degree of urbanization (compared to living in rural areas) have been shown to increase the risk of developing schizophrenia with a dose-dependent relationship (Vassos et al., 2012). Migrants (both first- and second-generation migrants) and minority ethnic groups across all cultures appears to be increasingly at risk of schizophrenia by 2 to 4 fold (Bourque et al., 2011; Selten et al., 2019). Indeed, psychotic symptoms have been shown to be increased in these groups (Rapoport et al., 2012). Childhood adversity is another widely replicated risk

factor for psychosis. Overall, children that have been exposed to any adversity have a 2.8 times higher risk of developing psychosis compared to those that have not (Varese et al., 2012). Furthermore, the risk has been shown to vary with the type of adversity, from the least to the highest risk increase: sexual abuse, bullying, physical abuse, and emotional abuse (by 2.3-3.4 times) (Varese et al., 2012). Interestingly, among individuals at an ARMS, those with high sexual abuse scores (i.e. more severe trauma) have showed a transition to psychosis risk 2 to 4 times that of those with low scores (i.e. less severe trauma) (Thompson et al., 2014). Additionally, exposure to stressful life events in adolescence and adulthood has been associated with the onset of psychosis. Individuals with psychosis show 3-fold increased odds of experiencing stressful live events in the period prior to the psychosis onset, with more intrusive events being suggestively relevant to the development of psychosis (Beards et al., 2013). It has been hypothesized that repeated exposure to stressful life events may induce the dysregulation of the hypothalamic-pituitary-adrenal (HPA) axis by increasing baseline cortisol levels (Beards et al., 2013; Borges et al., 2013).

The association between tobacco and cannabis consumption and the risk of developing psychosis has been widely replicated (Gurillo et al., 2015; Hunter et al., 2018; Kraan et al., 2016; Marconi et al., 2016). Smokers have shown a 2 times increased risk of schizophrenia compared to non-smokers (Hunter et al., 2018). Furthermore, smokers have been shown to be 3.2 times more likely to develop a FEP than non-smokers (Gurillo et al., 2015). Moreover, cannabis use has been associated with increased risk for schizophrenia with a dose-dependent relationship: heaviest cannabis users have been shown to be 3.9 times more likely to develop psychosis than non-users (Marconi et al., 2016). Interestingly, individuals at ARMS with a diagnosed cannabis abuse or dependence have shown a transition to psychosis rate higher than non-cannabis-using individuals at an ARMS by 1.8 fold (Kraan et al., 2016).

Recently, it has been hypothesized that a more accurate estimate of the risk of psychosis can actually be given, in a similar approach to the PRS, by an aggregated score representing the loading for multiple environmental risk factors: the environmental risk score (ERS) (Oliver et al., 2019, 2020; Padmanabhan et al., 2017; Vassos et al., 2020). A few studies have used the ERS to assess the risk of psychosis showing a higher ERS in: a) individuals at an ARMS compared with healthy controls [ERS composed by a weighted sum of the exposure of risk and protective factors, i.e. childhood trauma; ethnicity; immigration; non-right-handedness; pollution; urbanicity; season of birth in northern hemisphere; paternal age; paternal socioeconomic status; parental severe mental illness; adult life events; tobacco and cannabis

use; hearing problems; trait anhedonia; being male and 25-35 years old (Oliver et al., 2020)]; and b) individuals at ARMS (i.e. in young, familial high-risk subjects) who later transitioned to psychosis to those who did not by 2 fold [ERS composed by a weighted sum of the exposure of risk factors, i.e. season of birth; urbanicity; cannabis use; paternal age; obstetric complications; childhood trauma – physical and sexual abuse, neglect and parental death; (Padmanabhan et al., 2017)].

1.5.2. Individual prediction of psychosis using environmental risk factors and machine learning

The prediction of psychosis at an individual level using the exposure of environmental risk factors and machine learning has been explored, to the best of my knowledge, only once. Antonucci and colleagues used a sample of 337 healthy controls and 103 patients with schizophrenia to train a linear SVM classifier fed with 5 environmental risk factors from the pre-birth period to the adolescence period: a) the presence of obstetric complication during pregnancy, labor and delivery; b) the parental age at the time of birth of the individual; c) the presence of developmental anomalies during the first 66 months of life; d) the level of urbanicity; and e) the parental socio-economic status (Antonucci et al., 2020). The reported balanced accuracy was 65% and, although it was not statistically better than chance level, the validation accuracy was very similar (66%, using an independent sample of 102 healthy controls and 75 patients).

1.6. Individual prediction of psychosis using multimodality and machine learning

In this section, studies combining more than one modality (being at least one of them structural neuroimaging, genetics, or environment) and machine learning for prediction of psychosis at an individual level will be described. Most of the studies combined, in their prediction models, sMRI with a) clinical assessments (de Wit et al., 2017; Koutsouleris et al., 2018); b) clinical and neuropsychological assessments (Zarogianni et al., 2017); c) neuropsychological assessments (Karageorgiou et al., 2011; Pardo et al., 2006); d) clinical and neuropsychological assessments and biochemical data (Pina-Camacho et al., 2015); and e) functional MRI and diffusing tensor imaging (Sui et al., 2013). Two studies combined genetics with a) functional MRI (Yang et al., 2010); and b) demographics and clinical assessments (Struyf et al., 2008). Finally, one study combined genetics, environment and

neuropsychological assessments to build a psychosis prediction model (Antonucci et al., 2020).

de Wit and colleagues showed that including disorganized symptoms scores together with brain gyrification indexes and subcortical volumes into the same prediction model improved the accuracy in predicting the functional outcome (i.e. at follow-up and comparing high functioning to low functioning individuals) of individuals at an ARMS (accuracy = 82%, sensitivity = 69%, specificity = 94%, p = .003), but not when using functioning as a continuous outcome (r = .33, p > .050; see also performance of the prediction models trained only with the sMRI measures for comparison in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**) (de Wit et al., 2017). Furthermore, adding baseline global functioning social and role scores to voxel-based gray matter volume maps improved the accuracy in predicting the both social and role functioning at follow-up of individuals at ARMS (social functioning: accuracy = 83%, sensitivity = 83%, specificity = 82%, p < .001; role functioning: accuracy = 65%, sensitivity = 59%, specificity = 70%, p < .001; see also performance of the prediction models trained only with the sMRI measures for comparison in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**) (Koutsouleris et al., 2018). Moreover, Zarogianni and colleagues reported an increased accuracy in predicting transition to psychosis from an ARMS when including in the same prediction model regional grey matter densities, a measure of schizotypy and schizotypal cognitions, and verbal memory and learning assessment (compared to a model with only the neuroimaging data; balanced accuracy = 94%, sensitivity = 76%, specificity = 100%; see also performance of the prediction models trained only with the sMRI measures for comparison in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**) (Zarogianni et al., 2017). Another study reported an accuracy of 96% in distinguishing 8 healthy controls, 10 patients with schizophrenia, and 10 patients with bipolar disorder using several neuropsychological assessments (evaluating maintaining set, abstract thinking, auditory arithmetic, cognitive flexibility, eye-hand coordination, general knowledge, performance intelligence quotient, verbal fluency, planning, processing speed, recall memory for content, receptive vocabulary, selective attention, semantic verbal fluency, verbal memory, visuospatial perception, and working memory with distraction) and regional grey and white matters and cerebrospinal volumes (Pardo et al., 2006). The combination of structural neuroimaging with neuropsychological assessments to classify

patients with schizophrenia versus healthy controls was also tested by Karageorgiou and colleagues using a sample of 28 patients with a recent onset and 47 healthy controls. They have trained a classifier with regional whole-brain volumes and neuropsychological tests evaluating handedness, fine motor dexterity and speed, spatial motor coordination, reading and potential or premorbid intelligence quotient, visual and working memory, verbal and linguistic skills, verbal abstraction and declarative memory and problem solving executive skills, reporting an improved accuracy of 76% (sensitivity = 64%, specificity = 83%, $p < .001$; see also performance of the prediction models trained only with the sMRI measures for comparison in section **1.3.2. Individual prediction of psychosis structural neuroimaging using machine learning**) (Karageorgiou et al., 2011). Additionally, Pina-Camacho and colleagues showed that adding several clinical and neuropsychological assessments and oxidative stress determinations from blood samples (please refer to the original manuscript for an extensive list of all the variables assessed (Pina-Camacho et al., 2015)) increased the accuracy of a) classifying FEP patients who developed schizophrenia and b) who developed psychosis other than schizophrenia versus healthy controls and c) distinguishing FEP patients who developed schizophrenia from those who developed psychosis other than schizophrenia to 99%, 99%, and 81%, respectively (see also performance of the prediction models trained only with the sMRI measures for comparison in section **1.3.2. Individual prediction of psychosis using structural neuroimaging and machine learning**). Other neuroimaging modalities, in particular, functional and diffusion tensor imaging, have been combined with structural neuroimaging showing an improved accuracy of 75% in distinguishing patients with schizophrenia from healthy controls (see also performance of the prediction model trained only with the sMRI measures for comparison in section **1.3.2. Individual prediction of psychosis using structural neuroimaging and machine learning**) (Sui et al., 2013).

The combination of genetics with other modalities to classify schizophrenia have been explored. Yang and colleagues showed that the accuracy of classifying patients versus healthy controls improved when they included in the model (i.e. added to the genetic data) functional neuroimaging (accuracy = 87%, sensitivity = 86%, specificity = 89%; see also performance of the prediction model trained only with the genetics for comparison in section **1.4.2. Individual prediction of psychosis using genetics and machine learning**) (Yang et al., 2010). Another study showed that by adding clinical (post-mortem interval between death and freezing of brain samples, brain pH, rate of death, antibodies for the herpes simplex

viruses), demographical (age, sex) and environmental (data smoking at time of dead, past or current use of alcohol and drug) information to the gene expression profiling in the frontal brain region the AUC of classifying patients with schizophrenia *vs.* healthy controls increased to 0.94 (see also performance of the prediction model trained only with the gene expression for comparison in section **1.4.2. Individual prediction of psychosis using genetics and machine learning**) (Struyf et al., 2008). These findings are, indeed, in line with the current understanding of the etiology of psychosis, i.e. that it arises from a gene-environment interplay (Zwicker et al., 2018). Antonucci and colleagues conducted the most recent study taking this interplay into account by training a patients with schizophrenia *vs.* healthy controls classifier based on neuropsychological assessments of subsamples generated based upon genetic and environmental stratification (Antonucci et al., 2020). They have showed that this stratification strategy increased the balanced accuracy of diagnosis schizophrenia to 89% compared to the individual classifiers (i.e. genetic and environmental, see also performance of the prediction models trained only with each of those modalities for comparison in section **1.4.2. Individual prediction of psychosis using genetics and machine learning**, and **1.5.2. Individual prediction of psychosis using environment and machine learning**, respectively).

To the best of my knowledge, the combination of structural neuroimaging, genetics and environmental data and machine learning has never been explored to predict psychosis, neither for diagnosis purposes (i.e., classifying patients with schizophrenia, FEP patients, or individuals at ARMS, versus healthy controls), nor for prognosis purposes (i.e., predicting the transition to psychosis from an ARMS).

1.7. Objectives

The main aim of this project is to predict transition to psychosis, at a 2-year follow-up date, from an ARMS using quantitative data – i.e., neuroimaging (sMRI), genetics (genome-wide genotypes), and environmental (environmental risk assessments) data – collected when the subjects first sought clinical help (i.e., at baseline) and were identified with an ARMS. Four main objectives have been set to fulfill this aim.

Objective 1: to predict transition to psychosis from an ARMS using neuroimaging (sMRI) data. First, brain volume and surface measures (neuroimaging features) will be extracted

from the sMRI data using a brain segmentation pipeline that is more updated and computationally efficient than a classic one (see **objective 1.1.** and **2. Comparing SPM12 with CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study**). Then, the transition to psychosis prediction will be tested using: 1) a previous model-based approach, and 2) a novel-based approach. In the first, the ARMS sample (i.e. our study sample) will be used as an independent validation sample to the previously reported classification models (Koutsouleris, Riecher-Rössler, et al., 2015). In the second, a new machine learning model will be trained using the neuroimaging features and linear support vector machines and considering: a) sample balancing between groups (i.e. same number of subjects who later transitioned to psychosis and those who did not, matched for possible confounders – i.e. age, sex and data acquisition protocols); b) generalizability of the ML model (i.e. applying a nested cross-validation approach and artificially improving data sampling through bootstrapping); and c) different combinations of neuroimaging features, feature manipulation approaches and CV schemes (see **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**).

> **Objective 1.1:** to compare two brain segmentation pipelines of sMRI – the classical unified segmentation integrated in the Statistical Parametric Mapping 12 tool (SPM12), a standard neuroimaging processing tool, and the more recent and advanced segmentation pipeline of the Computational Anatomy Toolbox (CAT12), an SPM12 add-on. Differences in brain volume measures (i.e., the total grey and white matter volumes, and the left and right hippocampi grey matter volumes) using the segmented images from both pipelines will be evaluated in two stages. First, the brain volume measures of healthy subjects will be regressed against age and were compared with those of patients with Alzheimer's disease. Second, the brain volume measures will be tested as predictors of the presence of an Alzheimer's disease diagnosis (see **2. Comparing SPM12 with CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study**). This data and design were chosen to compare preprocessing platforms, due to the following reasons: 1) I had access to the Open Access Series of Imaging Studies (OASIS) brains database (https://www.oasis-brains.org/) which has sMRI images from healthy controls and patients with Alzheimer's disease; and there is strong previous evidence that 2) brain volume decreases with age, and 3) patients with Alzheimer's disease show decreased

grey matter volume, in particular in the hippocampus, when compared to health controls. Although the main goal of this PhD project is to predict transition to psychosis, there is no strong evidence, as for Alzheimer's disease, of association of specific structural brain biomarkers with the development of psychosis or even schizophrenia. Therefore, for the sole purpose of methodological validation of the best procedure possible to conduct the preprocessing of the images used in the present thesis, associations using age and an Alzheimer's diagnosis (where hippocampus volume is a robust biomarker) have been used. The study was already published (Vânia Tavares et al., 2020).

Objective 2: to predict transition to psychosis from an ARMS using genetic (i.e. genome-wide genotypes) data. The transition to psychosis prediction will be tested using 1) a previous model-based approach, and 2) a novel model-based approach. In the previous model-based approach, a polygenic risk score for schizophrenia that has been previously shown to distinguish FEP patients from healthy controls (Vassos et al., 2016) will be used as a predictor in a simple logistic regression model. In the novel model-based model approach, first, a list of psychosis-associated SNPs and a list of psychosis-associated genes for which an eQTL score is extracted (see **Objective 2.1** and **4. Creating a brain-based gene expression quantitative trait loci score (eGenScore) tool**) will be selected (genetic features). Then, a ML model will be trained using the genetic features and a logistic regularized regression algorithm, i.e. the elastic-net. For both model approaches (i.e. previous and novel) and similar to what was described in the **Objective 1.**, the ML approach will take into account a) sample balancing between groups (i.e. same number of subjects who later transitioned to psychosis and those who did not); b) generalizability of the ML model, i.e. applying a simple (in the case of the previous model-based approach) or nested (in the case of the novel model-based approaches) cross-validation approach and artificially improving data sampling through bootstrapping (see **5. Individual prediction of transition to psychosis using genetics and machine learning**).

Objective 2.1: to build and validate the eGenScore, a tool that computes a brain-based gene eQTL score. A gene expression model per each brain tissue (i.e. frontal, temporal, and occipital cortices, putamen, substantia nigra, hippocampus,

cerebellum, and white matter) and per gene will be trained using a reference database [Brain expressive Quantitative Trait Loci (eQTL) Almanac (BrainEAC)] with simultaneous genotype (i.e. SNPs) and transcriptome (i.e. messenger RNA levels per transcript – which has a correspondence to a given gene and, therefore, can be thought of as a proxy for gene expression) data, i.e. both data are available for each subject in the database. First, the additive effect of each SNP in the vicinity of a gene will be statistically tested (i.e., using a linear regression) against the expression of that gene. Second, the gene expression model will be composed by a weighted (i.e., effect sizes corrected for the frequency of each SNP's allele in the general population) sum of the SNPs with a statistically significant association with gene expression (i.e., eQTLs). Third, each model will be internally (using a 5-fold CV – for all brain tissues) and externally [using CommonMind Consortium (CMC) dataset – also with genotype and transcriptome data – only for the frontal cortex] validated (see **4. Creating a brain-based gene expression quantitative trait loci score (eGenScore) tool**). The methods used in this study are described in the follow-up study that has been already published (Vânia Tavares et al., 2021).

Objective 3: to predict transition to psychosis from an ARMS using environmental data (i.e. environmental risk assessments). The transition to psychosis prediction will be tested using 1) a previous model-based approach, and 2) a novel model-based approach. In the previous model-based approach, an environmental risk score for schizophrenia (Vassos et al., 2020) will be used as a predictor in a simple logistic regression model. In the novel model-based model approach, a list of environmental risk factors (i.e. upbringing urbanicity, cannabis and tobacco consumption, season of birth, parental age, childhood trauma, migration, and ethnic minority) will be used as features. Then, a ML model will be trained using the genetic features and a logistic regularized regression algorithm, i.e. the elastic net. For both model approaches (i.e. previous and novel) and similar to what was described in the **Objective 1.**, the ML approach will take into account a) sample balancing between groups (i.e. same number of subjects who later transitioned to psychosis and those who did not); b) generalizability of the ML model, i.e. applying a simple (in the case of the previous model-based approach) or nested (in the case of the novel model-based approaches) cross-validation approach and artificially improving data sampling through bootstrapping (see **6. Individual prediction of transition to psychosis using environmental data and machine learning**).

Objective 4: to predict transition to psychosis from an ARMS using a multimodal approach. A multimodal (i.e., neuroimaging and/or genetics and/or environment) and multilevel ML model will be trained if the modality-specific ML models (i.e., ML trained with neuroimaging, genetic or environmental features; see **Objectives 1, 2** and **3**) show a statistically significant accuracy (i.e., better than chance level) in predicting the development of psychosis from an ARMS. First, a modality-specific ML model (bottom-level models) will be trained (see **Objective 1, 2** and **3**). Then, another ML model (top-level) will be trained using the predictions each modality-specific model (i.e., whether it predicts that a given tested subject will transition to psychosis or not) and a simple logistic regression algorithm considering a) sample balancing between groups (i.e., same number of subjects who later transitioned to psychosis and those who did not, matched for possible confounders – i.e. age and sex); b) generalizability of the ML model (i.e., applying a nested cross-validation approach and artificially improving data sampling through bootstrapping); and c) different combinations of features. Of note is that this multimodal approach is limited by the very small size of the sample of individuals with concomitant structural, genetic and environmental data available (only 6 ARMS-T and 23 ARMS-NT).

1.8. Collaborations

The data used in this work was collected by researchers from the Institute of Psychiatry, Psychology Neurosciences (IoPPN) in King's College London, United Kingdom. The study design, the subject recruitment, the data acquisition (neuroimaging, genetic and environmenta data) and the clinical assessment were done by Dr. Evangelos Vassos, Dr. James Stone, Dr. Isabel Valli, Dr. James Woolley, Dr. Paolo Fusar-Poli, Dr. Ceri Jones, Dr. Maria Calem, Dr. Christopher Chaddock, and Dr. Diana Prata. The processing of the genetic data and calculation of the PRSs was performed by Dr. Evangelos Vassos. The data processing, the data analysis, and the results interpretation were performed by me, under the supervision of Prof. Diana Prata and Prof. Hugo Ferreira (Instituto de Biofísica e Engenharia Biomédica, Faculdade de Ciências da Universidade de Lisboa). Additionally, Dr. Vasco Diogo had a short-term mission at Dr. Nikolaos Koutsouleris's lab to establish the collaboration between that lab and Dr. Diana Prata's lab. The main goal of this mission was to validate machine learning models previously shown to predict transition to psychosis from

ARMS with accuracies better than chance level (see also **2.5.1. Previous model-based approach** in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**). Finally, Dr. Joana Monteiro contributed to the development of the eGenScore tool, namely for code optimization and definition of tool validation strategies.

2. Comparing SPM12 with CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study

2.1. Introduction

Segmentation is a key neuroimaging processing step to study brain structure (Despotović et al., 2015). Brain tissue volumes, i.e., grey and white matter volumes, measured using structural magnetic resonance imaging (sMRI) have been widely used as biomarkers in the research context. It is well established that the whole brain volume starts declining in early adulthood and accelerates in advanced normal aging, driven essentially by the atrophy of grey matter (Fjell & Walhovd, 2010; Fotenos et al., 2005; Marcus et al., 2007; Toepper, 2017), with an accelerated degeneration, for example, of the hippocampi (Fraser et al., 2015). Understanding normal aging brain atrophy profiles is indeed a fundamental step to identify specific age-related pathologies, such as the case of Alzheimer's disease (AD) (Fjell & Walhovd, 2010). AD is the most common neurodegenerative disorder characterized by synapse and neuronal loss which leads to macroscopic brain atrophy visible with sMRI (Lane et al., 2018). It has been consistently shown that, when compared to healthy subjects, patients with AD show a global grey matter volume loss that is accelerated with disease progression, with the hippocampi showing an initial atrophy already in the prodromal phase of the disease (Toepper, 2017). In fact, the hippocampi volumetry is considered a key imaging biomarker for AD (Teipel et al., 2013). However, the lack of a standard brain segmentation protocol able to uniformize brain volume measures across research and clinical centers is preventing the transition of these biomarkers into clinical practice, i.e., usable to define the clinical diagnosis criteria. Therefore, there is a need for studies comparing brain segmentation pipelines.

Several well-established automatic preprocessing pipelines designed to process sMRI data and segment brain images are currently available. However, it is not always clear what features represent a real relative improvement, and findings from existing and future studies using different platforms cannot be interpreted reliably in comparison to each other. To inform future choice of pipeline by researchers and study comparability, in this study I have compared two: the classical unified segmentation integrated in the Statistical Parametric Mapping 12 tool (SPM12) (Ashburner & Friston, 2005), a standard neuroimaging processing tool, and the more recent and advanced segmentation pipeline of the Computational Anatomy Toolbox (CAT12; Christian Gaser 2018, http://www.neuro.uni-jena.de/cat/), an SPM12 add-on. The main difference between these two pipelines is in the way they initialize

and update their estimation models for brain tissue classification. SPM12 bases the image segmentation on Tissue Probability Maps (TPM), which represent the prior probability of an image unit (voxel) being either grey or white matter, or non-brain tissue. On the other hand, CAT12 uses TPM only to spatially normalize the image, to perform an initial skull-stripping, and to initialize the segmentation. Then, it uses an adaptive maximum *a posteriori* (AMAP) segmentation approach (hypothesis-free approach) (Rajapakse et al., 1997), for further update of the estimation models for brain tissue classification, accounting also for partial volume effects. This allows for a more precise segmentation than SPM12 by estimating the amount of each brain tissue type – grey or white matter or non-brain tissue – that is present in each image unit. Furthermore, the AMAP approach models the parameters of the estimation as slowly varying spatial functions, which accounts for local intensity variations in the original brain image, including subject specific biological variations of brain tissues. Additionally, CAT12 is also computationally less expensive than SPM12 due to its parallel processing algorithms.

In order to improve the comparability between studies and aid future pipeline choice in the research setting, herein, I processed 1.5T sMRI data from healthy controls and patients with AD using SPM12 and CAT12 segmentation pipelines. Differences in brain volume measures (i.e., the total grey and white matter volumes, and the left and right hippocampi grey matter volumes) using the segmented images from both pipelines were evaluated in two stages. First, the brain volume measures of healthy subjects were regressed against age and were compared with those of patients with AD. Second, the brain volume measures were tested as predictors of the presence of an AD diagnosis. I chose these brain measures to compare pipelines, given the robust existing evidence for expecting a statistically significant negative association between all brain measures with age in health (Fotenos et al., 2005; Fraser et al., 2015; Marcus et al., 2010) and between grey matter (total and of the hippocampi) with AD diagnosis (Toepper, 2017). Finally, I replicated the above described analyses in an independent 3T sMRI dataset to verify if the results were replicable across datasets from different magnetic field strengths.

2.2. Materials and methods

2.2.1. Sample description

Two datasets were selected and downloaded from the Open Access Series of Imaging Studies database (OASIS, https://www.oasis-brains.org/): "1.5T dataset" and "3T dataset". A detailed description of both datasets is given below.

2.2.1.1. 1.5T dataset

Two samples of subjects from a pool of 316 healthy subjects and 100 patients with AD (cross-sectional OASIS-1, (Marcus et al., 2007)) were defined as follows: "Sample1" with 238 healthy subjects and no statistically significant gender effect on age (2 independent sample *t*-test; *p*-value > .05); and "Sample2" with 100 patients with AD (all subjects available in the original pool) and 78 age- and gender-matched healthy subjects, so that no statistically significant gender and diagnosis effects on age (2 independent sample *t*-test; *p*-value > .05) or gender effect on diagnosis (Pearson χ^2 test; *p*-value > .05) were present. All subjects had at least one 1.5T sMRI scan available. In cases when there was more than one scan available from a single session, the first sMRI scan was chosen.

2.2.1.2. 3T dataset

An additional sample, "Sample3", was defined from a pool of 609 healthy subjects and 489 patients with AD (longitudinal OASIS-3, (Marcus et al., 2010)) with the following criteria (see also **Table 5** in the Supplementary Material for more details): 1) composed by 100 patients with AD and 78 age- and gender-matched healthy subjects (same number of subjects as in "Sample2"); 2) same number of females and males as in Sample2; 3) no statistically significant gender and diagnosis effects on age (2 independent sample *t*-test; *p*-value > .05) or gender effect on diagnosis (Pearson χ^2 test; *p*-value > .05) were present (similar to Sample2); and 4) only subjects with at least one 3T sMRI scan were included. In cases when there was more than one scan available from a single session, the first scan was chosen. No sample composed only by healthy subjects (similar to Sample1) was selected due to the difference in the subjects' age range between the OASIS-1 and OASIS-3 (18-96 years and 42-95 years, respectively). A direct comparison of the results obtained by the analysis of these two samples would be impractical.

Table 5. Age differences between Sample2 and Sample3 using a 2 independent sample t-test.

	Male	**Female**
Healthy subjects	t = 1.53, p = .133	t = 1.87, p = .066
Patients with Alzheimer's disease	t = 1.86, p = .066	t = 1.71, p = .091

The statistical testing described above was performed in R (R Core Team, 2018) and the results are shown in **Table 6**. The subjects' IDs included in each Sample are shown in **Table S1** in the Supplementary Material of the online version of this paper (Vânia Tavares et al., 2020). All subjects were right-handed. Patients were diagnosed with AD if they scored higher than zero in the Clinical Dementia Rating (CDR) (Morris, 1993). Information regarding age and gender is also shown in **Table 6**.

Table 6. Sample characteristics.

	Age, years (mean ± SD, [range])	**Female/Male**
Sample1		
Healthy Subjects[1] (n=238)	40.0 ± 22.2 [18, 90]	119/119
Sample2		
Healthy Subjects[2] (n=78)	76.3 ± 7.6 [62, 94]	54/24
Alzheimer's Disease[2] (n=100)	76.8 ± 7.1 [62, 96]	59/41
Sample3		
Healthy Subjects[3] (n=78)	73.5 ± 6.7 [62 89]	54/24
Alzheimer's Disease[3] (n=100)	74.4 ± 5.9 [64, 89]	59/41

1. Effect of gender on age: 2 independent sample t-test = - 0.05, p = .963
2. Effect of gender on age: 2 independent sample t-test = - 0.58; p = .592;
   Effect of diagnosis on age: 2 independent sample t-test = - 0.40; p = .690;
   Effect of gender on the diagnosis: χ^2 = 1.56, p = .211
3. Effect of gender on age: 2 independent sample t-test t = 1.2; p = .773;
   Effect of diagnosis on age: 2 independent sample t-test = - 1.88; p = .062;
   Effect of gender on the diagnosis: χ^2 = 0.01, p = .905.

2.2.2. Structural magnetic resonance imaging

Structural MRI scans were acquired with three different scanners using a structural T1-weighted (T1w) magnetization prepared rapid gradient-echo (MP-RAGE) protocol: 1.5T

Magnetom Vision scanner (Siemens, Erlangen, Germany, voxel size = 1.0 x 1.0 x 1.25 mm^3; field-of-view = 256 x 256 x 128; repetition/echo/inversion times = 9.7 msec/4.0 msec/20.0 msec; flip angle = 10°); 3T TIM Trio scanner (Siemens, Erlangen, Germany, voxel size = 1.0 x 1.0 x 1.0 mm^3; field-of-view = 176 x 256 x 256; repetition/echo/inversion times = 2.4 msec/3.2 msec/1.0 msec; flip angle = 8°; 125 scans); and 3T BioGraph mMR PET-MR scanner (Siemens, Erlangen, Germany, voxel size = 1.2 x 1.1 x 1.1 mm^3; field-of-view = 176 x 240 x 256; repetition/echo/inversion times = 2.3 msec/3.0 msec/0.9 msec; flip angle = 9°; 53 scans).

2.2.3. Image processing

T1w 1.5T MRI images were first reoriented to the anterior commissure – posterior commissure line using the Display tool of SPM12 (v6909, http://www.fil.ion.ucl.ac.uk/spm/). Then, both reoriented T1w 1.5T and T1w 3T images were processed, in parallel, with SPM12 and with CAT12 (v1092, http://www.neuro.uni-jena.de/cat/) using default settings and MATLAB (9.1). First, bias field inhomogeneity correction was performed using an algorithm shared by SPM12 and CAT12. Second, images were segmented into grey matter, white matter, and cerebrospinal fluid using a classical unified segmentation approach with SPM12 (Ashburner & Friston, 2005) and the AMAP approach with CAT12 (Rajapakse et al., 1997). Third, the images were spatially normalized to a template derived from 555 healthy subjects of the IXI-database (http://www.brain-development.org) using the Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL) algorithm (Ashburner, 2007), which is shared by SPM12 and CAT12. Finally, total relative volumes of grey and white matter were computed by dividing the total volume of grey and white matter (the sum of all voxels classified as grey or white matter, respectively) by the total intracranial volume (TIV; the sum of all voxels classified as grey or white matter or as cerebrospinal fluid). Additionally, the grey matter relative volume of the left and right hippocampi was obtained using the Hammers region-of-interest (ROIs) atlas (Hammers et al., 2003). In particular, the Hammers atlas, which was first normalized to the same template as the segmented grey matter images (i.e. to the IXI-database template) using DARTEL algorithm, was used as a binary mask to select the ROIs corresponding to the left and right hippocampi in the segmented grey matter image. Then, the gray matter relative volume was defined as the sum of all the voxels classified as grey matter inside the respective ROI and divided by the TIV. The segmentation results (i.e. the

total relative volume of grey and white matter and of left and right hippocampi's grey matter) are summarized in the **Table 7** and illustrated in **Figure 3**.

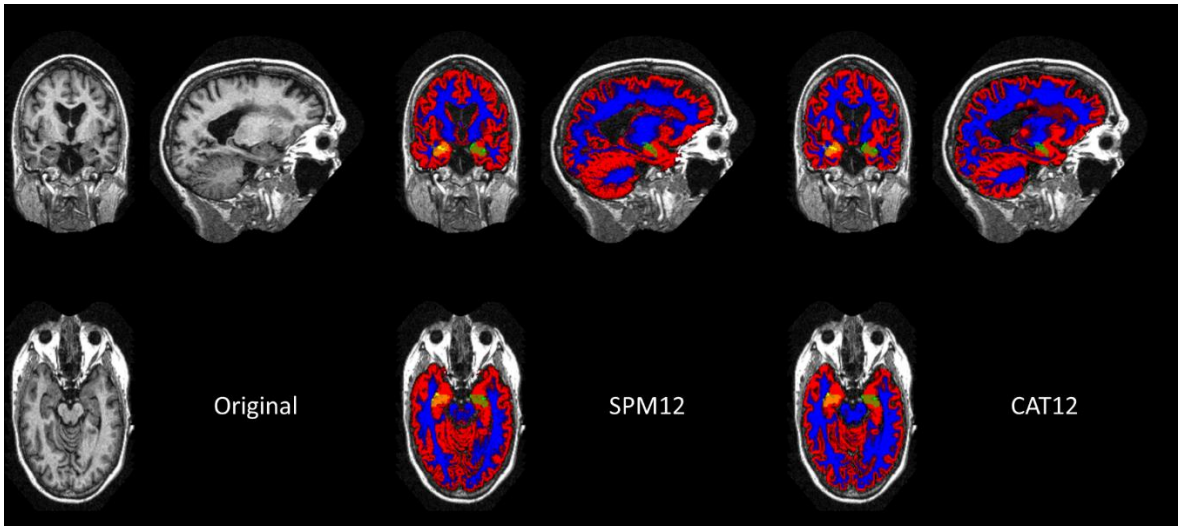


Figure 3. Segmented grey (in red) and white (in blue) brain tissues using SPM12 and CAT12. The hippocampi masks used to extract the grey matter from these regions are in green for the left hippocampus and in yellow for the right hippocampus. The original brain image belongs to a healthy subject (subject ID: OAS1, see also **Table S1**).

2.2.4. Statistical analysis

The statistical analysis was divided in 3 main parts. In the first part (Part 1 below), a correlation between pipelines (SPM12 or CAT12)'s brain volume measures was performed using linear regression. The second part (Part 2 below) examined the interaction of pipeline with age and AD diagnosis using linear mixed models (LMMs). In the third part (Part 3 below), using logistic regression, the predictive accuracy of each brain volume measure in detecting the presence of AD was estimated for each pipeline separately, and then compared between them.

2.2.4.1. Correlation between the pipelines' brain volume measures (Part 1)

The total grey and white matter relative volumes obtained with SPM12 were linearly regressed against the ones obtained with CAT12, using only the healthy subjects from Sample1 and Sample3, separately. Additionally, the coefficient of determination (R^2) was extracted.

Table 7. Mean relative volume (i.e. the volume divided by the total intracranial volume) of the total grey and white matters and the left and right hippocampi for healthy subjects from Sample1, healthy subjects and patients with Alzheimer's disease from Sample2 and from Sample3. The volumes were estimated using SPM12 and CAT12.

Subjects from Sample1		
	SPM12	**CAT12**
Grey matter	0.51 ± 0.05	0.46 ± 0.04
White matter	0.33 ± 0.02	0.35 ± 0.02
Left hippocampus	0.0014 ± 0.0001	0.0015 ± 0.0002
Right hippocampus	0.0015 ± 0.0001	0.0016 ± 0.0002

Subjects from Sample2				
	Healthy controls		**AD patients**	
	SPM12	**CAT12**	**SPM12**	**CAT12**
Grey matter	0.43 ± 0.04	0.39 ± 0.03	0.41 ± 0.04	0.37 ± 0.03
White matter	0.31 ± 0.03	0.33 ± 0.03	0.30 ± 0.03	0.32 ± 0.03
Left hippocampus	0.0013 ± 0.0001	0.0013 ± 0.0002	0.0011 ± 0.0002	0.0011 ± 0.0002
Right hippocampus	0.0014 ± 0.0001	0.0014 ± 0.0002	0.0012 ± 0.0003	0.0012 ± 0.0003

Subjects from Sample3				
	Healthy controls		**AD patients**	
	SPM12	**CAT12**	**SPM12**	**CAT12**
Grey matter	0.42 ± 0.04	0.40 ± 0.04	0.40 ± 0.05	0.38 ± 0.03
White matter	0.30 ± 0.03	0.33 ± 0.03	0.29 ± 0.03	0.32 ± 0.03
Left hippocampus	0.0013 ± 0.0001	0.0014 ± 0.0002	0.0012 ± 0.0002	0.0012 ± 0.0003
Right hippocampus	0.0014 ± 0.0001	0.0015 ± 0.0002	0.0013 ± 0.0002	0.0013 ± 0.0003

2.2.4.2. Interaction of pipeline with age and Alzheimer's diagnosis on volume measures (Part 2)

2.2.4.2.1. Effect of age on brain volume measures using SPM12 and CAT12 in healthy subjects (Part 2.1.)

The interaction of pipeline (SPM12, CAT12) with age on each brain volume measure (i.e. total grey and white matter relative volumes and grey matter relative volume of the right and left hippocampi) was tested with four LMMs in healthy subjects from Sample1, i.e. one for each brain volume measure. In these LMMs, age, pipeline, and 'age by pipeline' interaction were entered in the model as fixed effects, and subject as a random effect.

2.2.4.2.2. Effect of Alzheimer's disease diagnosis on brain volume measures using SPM12 and CAT12 (Part 2.2.)

The interaction of pipeline (SPM12, CAT12) with AD diagnosis (AD, healthy subjects) on each brain volume measure was also tested in subjects from Sample2 and Sample3, separately. Diagnosis, pipeline, and 'diagnosis by pipeline' interaction were entered in the LMM model as fixed effects, covarying for age, and subject as a random effect.

2.2.4.3. Detection of Alzheimer's disease diagnosis from brain volume measures using SPM12 and CAT12 (Part 3)

The predictive ability of each brain volume in detecting AD was tested for each pipeline separately in Sample2 and in Sample3 individually, by logistically regressing each brain volume measure against diagnosis, while controlling for age. The area under the receiver operating characteristic curve was then statistically compared between pipelines using the DeLong method (DeLong et al., 1988).

Overall statistical significance of each effect of interest (pipeline and 'age by pipeline' in Sample1, and 'diagnosis by pipeline' in Sample2 and in Sample3 - the main effect of pipeline was already tested in Sample1) was evaluated using the Log-Likelihood ratio test. All *p*-values were corrected for multiple testing using False Discovery Rate (FDR) and reported as statistically significant at an FDR-corrected p-value < .05. Cohen's f^2 (Selya et al., 2012) effect size was computed for all pipeline effects (i.e. the main effect and the 'age by pipeline'

and ‘diagnosis by pipeline’ interaction effects). Additionally, the beta coefficients for age in each pipeline is reported as the effect size for the ‘age by pipeline’ interaction and Cohen’s *d* (Gibbons et al., 1993) as the effect size for the pipeline and ‘diagnosis by pipeline’ contrasts (i.e. SPM12 *vs.* CAT12 and SPM12 *vs.* CAT12 in AD and in healthy subjects, respectively). Interpretation of Cohen’s *d* was performed using Kristoffer Magnusson web tool (Interpreting Cohen's *d* effect size, https://rpsychologist.com/d3/cohend/). All the statistical analysis described in this section was done using R (R Core Team, 2018), using the following R packages: 1) ‘stats’ (R Core Team, 2018) for simple linear and logistic regression, and FDR correction of the statistical testing *p*-values; ‘nlme’ (Pinheiro et al., 2018) for LMMs fitting and statistical testing; and ‘pROC’ (Robin et al., 2011) for the ROC curve analysis.

2.3. Results

2.3.1. Correlation between the pipelines’ brain volume measures (Part 1)

The linear regression analysis showed that the total grey and white matter relative volumes estimated with SPM12 and CAT12 are strongly correlated [Sample1 - grey matter relative volume: $p < .001$, $R^2 = 0.90$, **Figure 4A**, top left; and white matter relative volume: $p < .001$, $R^2 = 0.71$, **Figure 4A**, top right; Sample3 - grey matter relative volume: $p < .001$, $R^2 = 0.55$, **Figure 4B**, top left; and white matter relative volume: $p < .001$, $R^2 = 0.50$, **Figure 4B**, top right]. Grey matter relative volume estimations with SPM12 showed higher values than the ones with CAT12, with this difference increasing with higher average total grey matter relative volumes. The opposite effect was shown for white matter relative volume estimations (see Bland-Altman plots in **Figure 4A and B**, bottom).

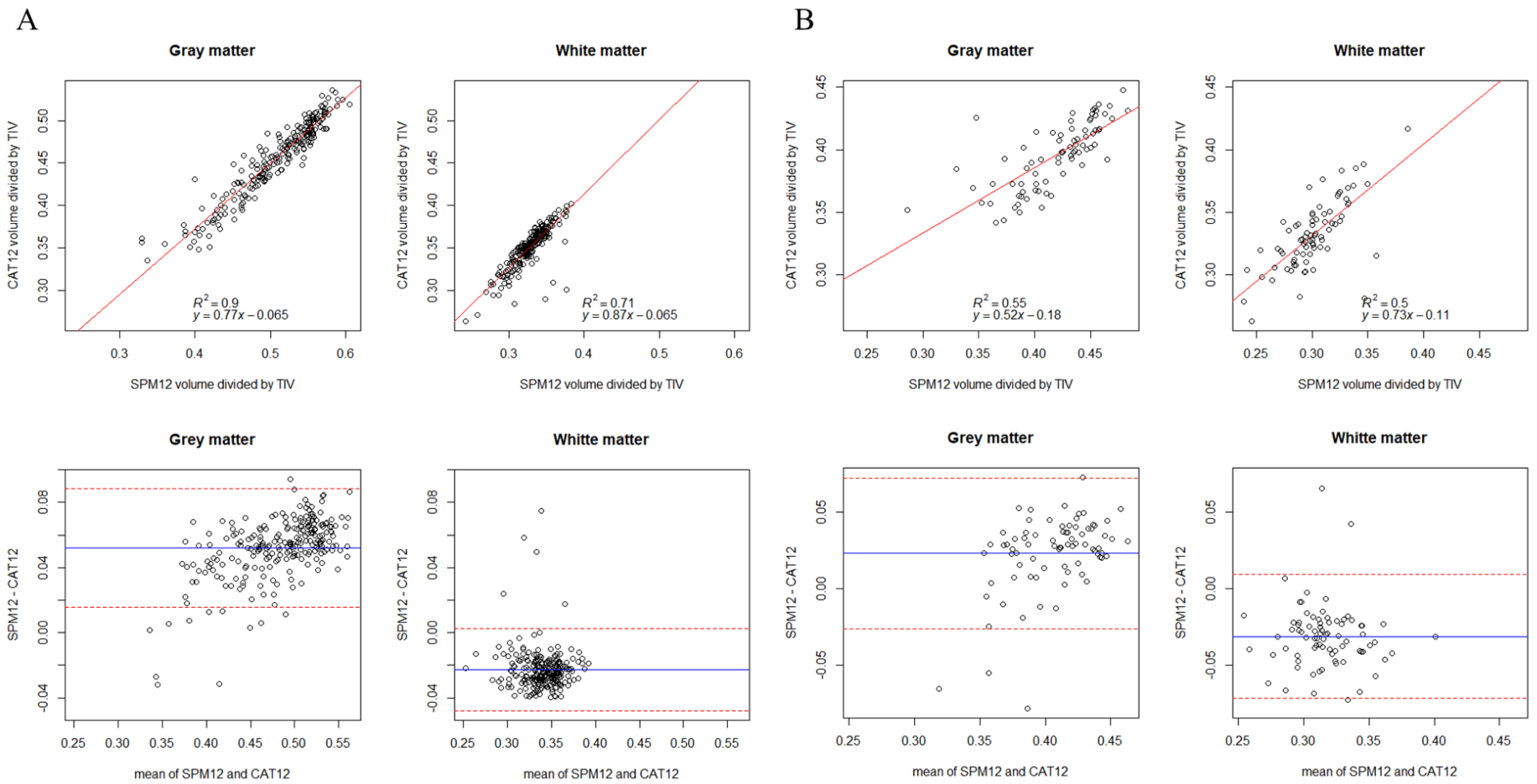


Figure 4. Top: linear regression analysis between total grey (**left**) and white (**right**) matter volumes obtained with SPM12 and CAT12 using only the healthy subjects from Sample1 (**A**) or Sample3 (**B**). The red line represents the fitted regression line (which equation is represented in the form of CAT12 volume (y) = slope * SPM12 volume (x) + intercept; and effect size is represented by the R^2). **Bottom**: Bland-Altman plots with limits (dashed red lines) of agreement for mean (continuous blue line) total grey (**left**) and white (**right**) matter volumes. TIV: total intracranial volume.

2.3.2. Interaction of pipeline with age and Alzheimer's disease diagnosis on volume measures (Part 2)

2.3.2.1. Effect of age on brain volume measures using SPM12 and CAT12 in healthy subjects (Part 2.1.)

I found a statistically significant effect of pipeline on all brain volume measures ($p < .001$), with SPM12 showing lower volume estimations in total white matter and grey matter of the hippocampi and higher volume estimations in total grey matter (grey matter: $d_{SPM12\text{-}CAT12}$ = 4.02; white matter: $d_{SPM12\text{-}CAT12}$ = -1.25; left hippocampus: $d_{SPM12\text{-}CAT12}$ = -1.56; right hippocampus: $d_{SPM12\text{-}CAT12}$ = -1.64; **Figure 5**). Most importantly, I found an interaction effect of age and pipeline on all brain measures ($p < .001$), with SPM12 showing a higher decrease in total grey matter relative volume with aging than CAT12 ($\beta_{age,CAT12}$ = -0.0018, $\beta_{age,SPM12}$ = -0.0022); and lower decrease in total white matter and both hippocampi grey matter relative volumes (white matter: $\beta_{age,CAT12}$ = -0.25, $\beta_{age,SPM12}$ = -0.17, $\beta_{age^2,CAT12}$ = -0.24, $\beta_{age^2,SPM12}$ = -0.17; left hippocampus: $\beta_{age,CAT12}$ = -4.7×10^{-6}, $\beta_{age,SPM12}$ = -3.1×10^{-6}; right hippocampus: $\beta_{age,CAT12}$ = -4.2×10^{-6}, $\beta_{age,SPM12}$ = -2.7×10^{-6}; **Figure 6**). See also **Table 8** for statistics and effect sizes.

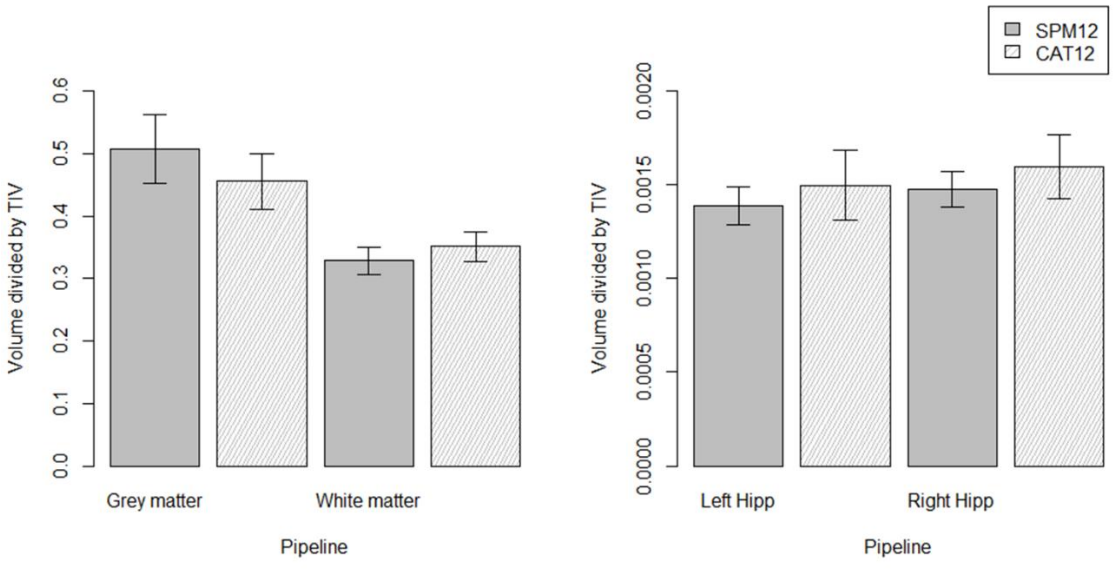


Figure 5. Mean total grey and white matters volume (**left**), and grey matter volume of hippocampi (**right**) estimated using SPM12 and CAT12 pipelines. Only the healthy subjects from Sample1 were used. Error bars represent the standard deviation from the mean volume. TIV: total intracranial volume.

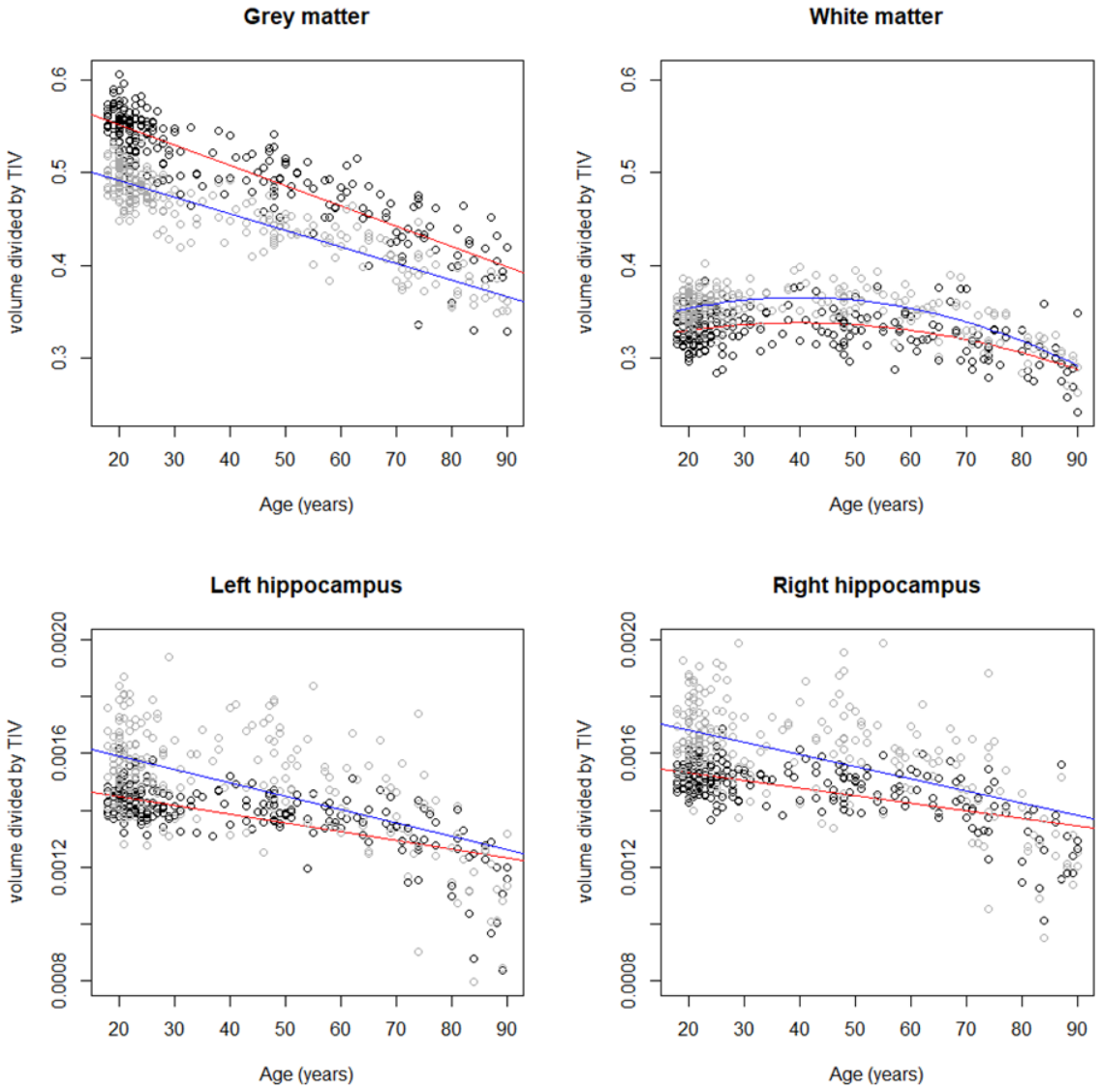


Figure 6. Scatter plot of the grey (**top left**) and white (**top right**) matter volume estimation and grey matter volume estimation of the left (**bottom left**) and right (**bottom right**) hippocampus using SPM12 (black dots) and CAT12 (grey dots). Fitted trend lines are in red (continuous line) for SPM12 and in blue for CAT12 (dashed line). Only the healthy subjects from Sample1 were used. TIV: total intracranial volume.

Table 8. Effect of age, pipeline, and interaction effect of 'age by pipeline' (SPM12 vs. CAT12) on brain volume measures using Sample1 and effect of diagnosis (patients with Alzheimer's disease (AD) vs. healthy subjects (HS)) and interaction effect of 'diagnosis by pipeline' on brain volume measures using Sample2 and Sample3. The overall statistical significance of the effect was tested with the Log-likelihood test (χ^2). Cohen's f^2 effect size was computed for all pipeline effects (i.e. 'age by pipeline' and 'diagnosis by pipeline' interaction effects). Additionally, the beta coefficients for age in each pipeline is reported as the effect size for the 'age by pipeline' interaction and Cohen's d as the effect size for the pipeline and 'diagnosis by pipeline' contrasts (i.e. SPM12 vs. CAT12 and SPM12 vs. CAT12 in AD and in HS, respectively). Statistically significant results are reported with a p-value < .05 (all p-values are FDR-corrected) and market with an asterisk.

	Grey matter			**White matter**			**Left hippocampus**			**Right hippocampus**		
Sample1	χ^2	*p*	**Effect size**	χ^2	*p*	*Effect size*	χ^2	*p*		χ^2	*p*	*Effect size*
pipeline	536.97	<.001*	f^2 = 0.39	337.09	<.001*	f^2 = 0.95	156.62	<.001*	f^2 = 0.53	178.03	<.001*	f^2 = 0.51
			$d_{SPM12\text{-}CAT12}$ = 4.02			$d_{SPM12\text{-}CAT12}$ = -1.25			$d_{SPM12\text{-}CAT12}$ = -1.56			$d_{SPM12\text{-}CAT12}$ = -1.64
age by pipeline	60.18	<.001*	f^2 = 0.03	39.32	<.001*	f^2 = 0.05	24.16	<.001*	f^2 = 0.09	23.75	<.001*	f^2 = 0.08
			$\beta_{ageCAT12}$ = -0.0018			$\beta_{ageCAT12}$ = -0.25			$\beta_{ageCAT12}$ = -4.7x10^{-6}			$\beta_{ageCAT12}$ = -4.2x10^{-6}
			$\beta_{ageSPM12}$ = -0.0022			$\beta_{ageSPM12}$ = -0.17			$\beta_{ageSPM12}$ = -3.1x10^{-6}			$\beta_{ageSPM12}$ = -2.7x10^{-6}
						β_{age^2CAT12} = -0.24						
						β_{age^2SPM12} = -0.17						
Sample2												
diagnosis by pipeline	2.70	.134	f^2 = 0.01	0.07	.789	f^2 = 0.00	3.90	.134	f^2 = 0.72	0.65	.134	f^2 = 0.04
			$d_{CAT12,\ AD\ -\ HS}$ = -0.66			$d_{CAT12,\ AD\ -\ HS}$ = -0.63			$d_{CAT12,\ AD\ -\ HS}$ = -0.91			$d_{CAT12,\ AD\ -\ HS}$ = -0.89
			$d_{SPM12,\ AD\ -\ HS}$ = -0.49			$d_{SPM12,\ AD\ -\ HS}$ = -0.61			$d_{SPM12,\ AD\ -\ HS}$ = -0.83			$d_{SPM12,\ AD\ -\ HS}$ = -0.77
Sample3												
diagnosis by pipeline	0.15	.696	f^2 = 0.00	0.13	.716	f^2 = 0.00	6.15	.013*	f^2 = 0.07	8.93	.003*	f^2 = 0.16
			$d_{CAT12,\ AD\ -\ HS}$ = -0.63			$d_{CAT12,\ AD\ -\ HS}$ = -0.62			$d_{CAT12,\ AD\ -\ HS}$ = -0.87			$d_{CAT12,\ AD\ -\ HS}$ = -0.92
			$d_{SPM12,\ AD\ -\ HS}$ = -0.51			$d_{SPM12,\ AD\ -\ HS}$ = -0.68			$d_{SPM12,\ AD\ -\ HS}$ = -0.83			$d_{SPM12,\ AD\ -\ HS}$ = -0.85

2.3.2.2. Effect of Alzheimer's disease diagnosis on brain volume measures using SPM12 and CAT12 (Part 2.2.)

I found a statistically significant interaction effect of diagnosis and pipeline on both hippocampi grey matter relative volumes (left: p = .013, $d_{\text{CAT12, Alzheimer's Disease (AD) – Healthy Subjects (HS)}}$ = -0.87, $d_{\text{SPM12, AD - HS}}$ = -0.83; and right: p = .003, $d_{\text{CAT12, AD - HS}}$ = -0.92, $d_{\text{SPM12, AD – HS}}$ = -0.85; **Figure 7**) when using Sample3. In particular, patients with AD showed lower volume estimates than healthy subjects in both segmentation pipelines, but with a higher difference when CAT12 was used. The interaction effect was not significant on the total grey and white matter relative volumes when using Sample3, nor on any brain volume measures when using Sample1. See **Table 8** for statistics and effect sizes.

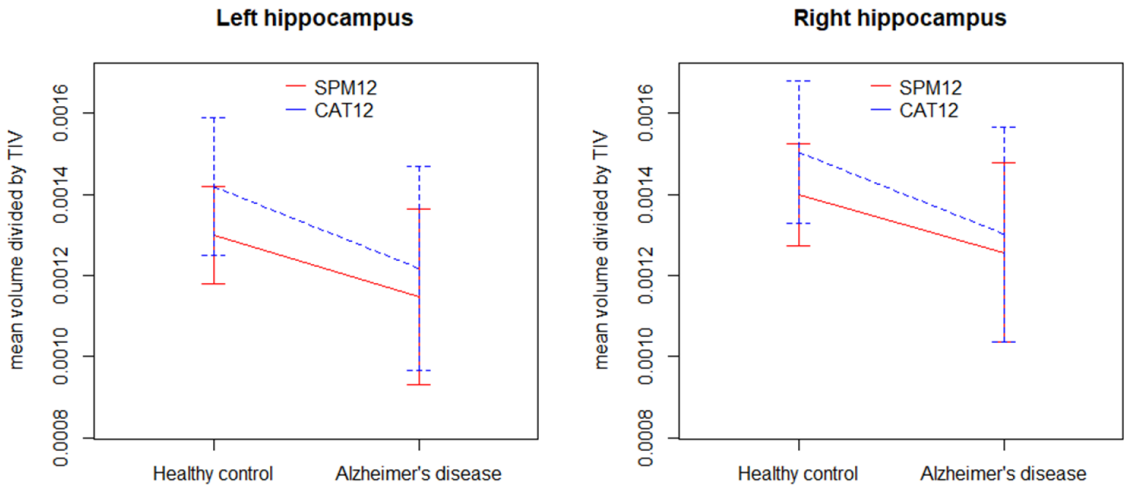


Figure 7. Mean grey matter volume of left (**left**) and right (**right**) hippocampus for healthy subjects and patients with Alzheimer's disease from Sample3 estimated using SPM12 (continuous red line) and CAT12 (dashed blue line). Error bars represent the standard deviation from the mean volume. TIV: total intracranial volume.

2.3.3. Detection of Alzheimer's disease diagnosis from brain volume measures using SPM12 and CAT12 (Part 3)

All brain volume measures were able to statistically predict AD diagnosis above chance level (AUC > 0.5; **Table 9**) when controlling for age. This result holds for both pipelines (SPM12 and CAT12) and samples (Sample2 and Sample3). The between-pipeline AUC statistical comparison showed no difference between the volume measures' predictive accuracy, of

SPM12 versus CAT12, at a statistical significance level of 5% (see **Table 9** for statistics). The ROC for each test is represented in **Figure 8**.

2.4. Discussion

Aiming to improve the comparability between studies and aid future platform choice in the research setting, I compared two brain image segmentation pipelines, SPM12 and CAT12, by 1) regressing the total grey and white matter relative volumes obtained with SPM12 with those obtained with CAT12; studying the association of total grey and white matter, and grey matter of left and right hippocampi relative volumes with 2) age in a healthy population; and 3) the diagnosis of AD; and 4) the effect of pipeline on the volume measures' estimation of an AD diagnosis (*vs.* healthy subjects).

2.4.1. Pipeline comparison in volume measurements in healthy subjects (Part 1)

Overall, the SPM12 and CAT12 grey and white matter relative volume estimates are strongly correlated in both datasets (i.e., 1.5T and 3T datasets) as expected. CAT12 total grey matter relative volume estimates are, on average, lower than the one estimated by SPM12, but the opposite is shown for grey matter of hippocampi and total white matter relative volumes. In particular, higher mean total grey matter relative volume estimates showed a higher volume difference between SPM12 and CAT12, whereas the opposite was shown for total white matter relative volume estimates. These results seem to indicate that CAT12 underestimates (relative to SPM) the total grey matter relative volume and overestimates the total white matter relative volume as the mean volume increases. This finding should be considered when comparing volumetric studies that use SPM12 with the ones that use CAT12. Furthermore, these effects are present in both 1.5T and 3T datasets, suggesting that the magnetic field strength has little impact on the choice of the brain segmentation pipeline.

Table 9. Logistic regression results between each brain volume measures and diagnosis (healthy subjects vs. patients with AD) using SPM12, CAT12, Sample1 and Sample2. Area under the receiver operating characteristic curve (AUC) is showed for each tested model together with the 95% confidence interval (CI). The AUC values were statistically compared between pipelines (SPM12 and CAT12) for each sample using the DeLong method.

Sample2

	$R^2_{Nagelkerke\ SPM12}$	$R^2_{Nagelkerke\ CAT12}$	**SPM12**	**CAT12**	**Statistical comparison**
			AUC [95% CI]	AUC [95% CI]	AUC_{SPM12} *vs.* AUC_{CAT12}
Grey matter	.15	.19	.71 [.63, .79]	.72 [.65, .80]	Z = -0.40, p = .973
White matter	.04	.05	.59 [.51, .67]	.60 [.52, .68]	Z = -0.33, p = .973
Left hippocampus	.27	.27	.76 [.69, .83]	.77 [.70, .84]	Z = -0.26, p = .973
Right hippocampus	.36	.34	.79 [.73, .86]	.79 [.73, .86]	Z = -0.03, p = .973

Sample3

	$R^2_{Nagelkerke\ SPM12}$	$R^2_{Nagelkerke\ CAT12}$	**SPM12**	**CAT12**	**Statistical comparison**
			AUC [95% CI]	AUC [95% CI]	AUC_{SPM12} *vs.* AUC_{CAT12}
Grey matter	.08	.13	.65 [.57, .73]	.69 [.62, .77]	Z = -1.53, p = .503
White matter	.12	.13	.68 [.60, .76]	.68 [.60, .75]	Z = 0.13, p = .899
Left hippocampus	.24	.25	.73 [.66, .80]	.74 [.66, .81]	Z = -0.21, p = .899
Right hippocampus	.20	.23	.72 [.64, .79]	.74 [.67, .81]	Z = -0.69, p = .899

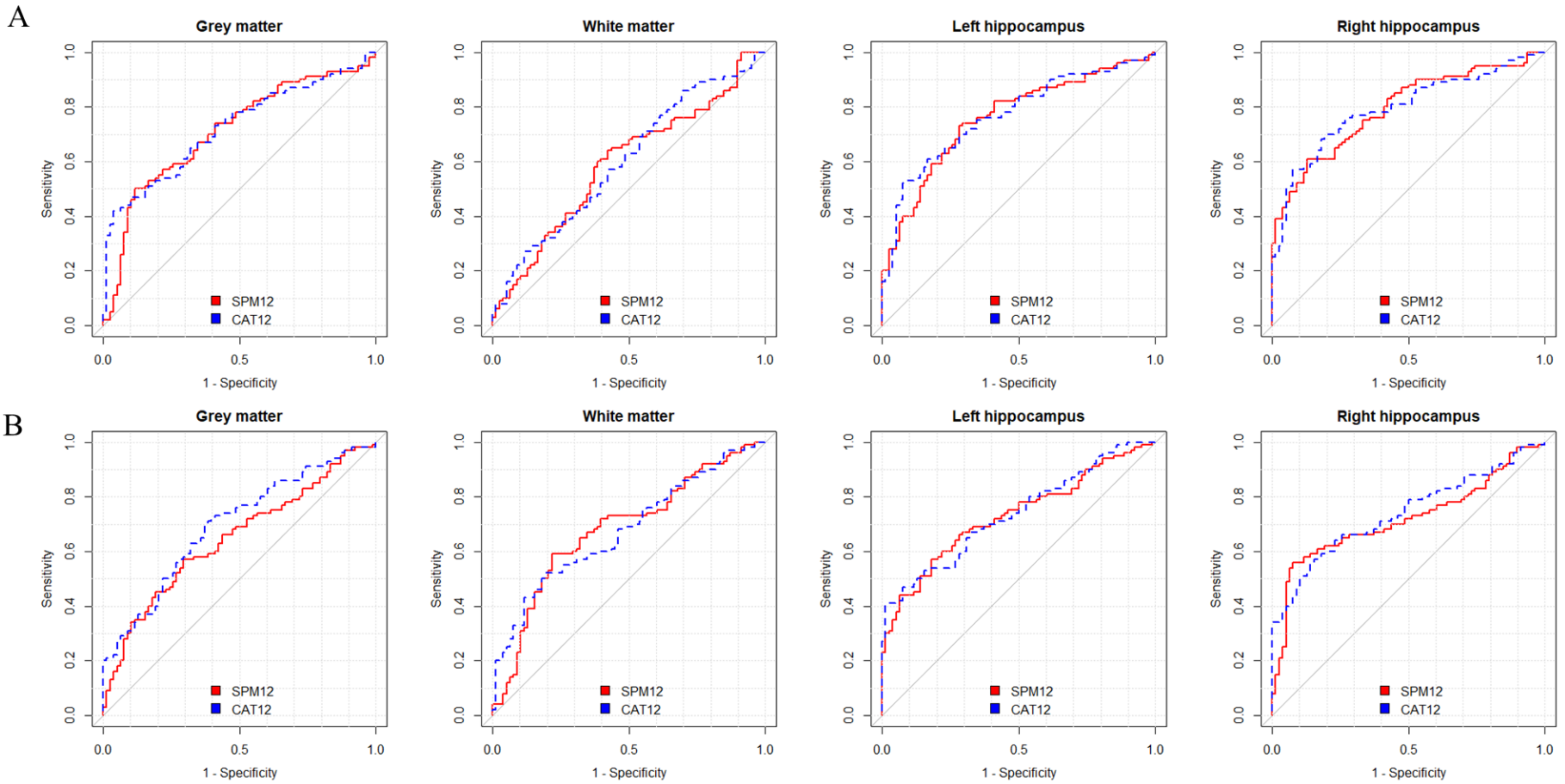


Figure 8. Receiver characteristic curves for each pipeline (SPM12 in red and CAT12 in blue) when logistically regressing each brain volume measure against diagnosis (healthy subjects versus patients with Alzheimer's Disease). Subjects used in this analysis belong to Sample2 (**A**, top row) or Sample3 (**B**, bottom row).

2.4.2. Pipeline comparison in the effect of age on brain volume measures in healthy subjects (Part 2.1.)

Also as expected, age was negatively correlated with all grey matter measures (i.e. total and hippocampi relative volumes), whereas total white matter relative volume showed an inverted U-shape relationship with age, with an inflexion point at middle age (around 45 years of age). This result is in line with previous findings using the same dataset as us herein (Fotenos et al., 2005) and as more recently reviewed (Fjell & Walhovd, 2010; Toepper, 2017), which showed that aging leads to a decrease in both grey and white matter (although only after the forth decade of life), but with a steeper grey matter atrophy. Importantly, the choice of pipeline explained between 39% (for total grey matter) and 95% (for total white matter) of the variance in all volume estimates (i.e. a 'very large' effect size). Roughly 100% of the subjects showed mean SPM12 total grey matter relative volume estimates that were higher than CAT12 ($d = 4.02$), and 88 to 95% showed mean CAT12 total white matter and grey matter of the hippocampi relative volume estimates that were higher than SPM12 ($-1.25 < d < -1.64$). Additionally, the interaction of age and pipeline explained 3 to 9% of the variance in all brain volume measures generally (i.e. a 'small' to 'medium' effect size). Total grey matter relative volume measures estimated by SPM12 showed a steeper decrease with age than the ones estimated by CAT12, whereas the opposite was found for the other three brain measures, i.e. a percentage difference of 20.0% for total grey matter, 41.0% and 43.5% for the grey matter of the left and right hippocampi, respectively, and 38.1% (β_{age}) and 34.1% (β_{age}^2) for total white matter (computed as the absolute percentage difference between the SPM12 and CAT12 beta coefficients). Although one cannot establish which segmentation pipeline translates the actual brain volume and aging association as the ground truth since the volume measures are not known, this segmentation platform comparison may be relevant for the appropriate comparison between existing and future findings using these platforms, in particular those of aging. Moreover, it is worth to note that these results hold only for 1.5T MRI scans, as the effect of segmentation pipeline choice to study the association between age and brain volume measures were not tested for using a 3T dataset.

2.4.3. Pipeline comparison in the effect of Alzheimer's disease diagnosis on brain volume measures (Part 2.2)

Additionally, I compared brain volume measures of patients with AD with those of age- and gender-matched healthy subjects and explored the interaction between diagnosis and

pipeline. Across brain volume measures, AD lowered brain volumes as expected, particularly in the hippocampi (Toepper, 2017). Interestingly, the pipeline did not significantly modulate the diagnosis effect when using 1.5T MRI scans, with this interaction explaining roughly 0% to 4% of all brain volume measures (i.e. 'small' effect size), except of the left hippocampus where 72% of its variance was explained by the pipeline choice (i.e. a 'large' effect size). On the other hand, when testing the 3T dataset, the pipeline and diagnosis interaction effect was significant on the hippocampi grey matter relative volumes, explaining 7 and 16% of its variance (a 'small' and 'medium' effect size, respectively), with SPM12 distinguishing patients with AD from healthy subjects slightly better than CAT12. There was an increase in effect size of 4.7% and 7.9% for left and right hippocampi grey matter relative volume, respectively (computed as the absolute percentage difference between the SPM12 and CAT12 Cohen's *d*), resulting in an increase of 1% and 2%, respectively, of patients showing mean relative volume estimates lower than healthy subjects when using SPM12, compared to CAT12.

2.4.4. Pipeline comparison in the detection of Alzheimer's disease diagnosis using brain volume measures (Part 3)

Finally, I compared the ability of the brain volumes obtained, with each segmentation pipeline, to distinguish patients with AD from age- and gender-matched healthy subjects. Results were similar when using 1.5T and 3T datasets. Particularly, both pipelines (SPM12 and CAT12) were able to produce volumes detecting AD above chance level, with hippocampi relative volumes showing the highest accuracies, as expected. Indeed, hippocampi volume shows a steeper grey matter atrophy rate than total grey matter already at a prodromal phase of the disease (Toepper, 2017), and it is currently the best-established imaging biomarker (research-wise) for AD (Teipel et al., 2013). Furthermore, the volumes' measure ability to detect AD diagnosis did not differ significantly between pipelines which is in line with the recent demonstration that the prediction power of MRI-based brain volume measures, in particular of hippocampi, does not depend on the measurement method, but mainly on the degree of brain tissue atrophy (Buchert et al., 2018).

2.4.5. Limitations

This study has a few limitations that need to be addressed. First, I did not compare these results with the ground truth segmentation method, that currently is defined as manual segmentation usually performed by one or more experts. This is a very time-consuming task and almost impractical when analyzing medium to big sample sizes, as this herein. Second, I could not analyze the age effect modulated by the segmentation pipeline on brain volume measures in a 3T dataset, due to the lack of availability of subjects with a matched age range with the 1.5T dataset. Furthermore, in order to statistically test the modulation effect of the magnetic field strength on the choice of pipeline in brain volume-based studies, a dataset composed by the subjects with both MRI scans ideally (i.e. 1.5T and 3T) should be used. Third, I only tested total grey and white matter relative volume measures and grey matter relative volume measures of the hippocampi. Future studies should also investigate the impact of the segmentation pipeline choice on other brain regions, by conducting whole brain voxel-wise morphometry studies.

In this study, I attempted to compare the research usefulness, albeit not clinical usefulness, of SPM12 and CAT12 neuroimaging analytical software pipelines. Taken together, these results show that 1) SPM12 and CAT12 brain volume measure differences are tissue-dependent; 2) the choice of segmentation pipeline (i.e. SPM12 or CAT12) modulated the effect of age on all brain tissue volumes and of diagnosis, albeit only on 3T MRI-based hippocampi grey matter volumes, but 3) did not impact the accuracy of the brain tissue volumes in detecting AD diagnosis. Therefore, future volume-based studies are encouraged to take into account these pipeline effects when comparing their results to other studies' findings. Furthermore, the use of CAT12 is also encouraged when conducting AD studies, as this is a more advanced brain segmentation tool and computationally less expensive than SPM12.

3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning

3.1. Introduction

In the last decade, several studies of the at-risk mental state (ARMS) population have employed machine learning using structural neuroimaging, i.e., structural magnetic resonance imaging (sMRI), to predict clinical outcome. These have been focused on two specific outcomes: functional [globally (de Wit et al., 2017; Kambeitz-Ilankovic et al., 2016), in social or role domains (Koutsouleris et al., 2018), and regarding psychotic symptoms severity (Stefania Tognin et al., 2014); **Table 10**] and transition to psychosis (Das et al., 2018; Koutsouleris et al., 2009, 2012; Koutsouleris, Riecher-Rössler, et al., 2015; Zarogianni et al., 2017, 2019), **Table 11**. For example, Kambeitz-Ilankovic and colleagues showed that a cortical surface-based pattern allows the prediction of poor *vs.* good functioning outcome at follow-up of 27 individuals at an ARMS with an accuracy of 82% (Kambeitz-Ilankovic et al., 2016). Furthermore, using a similar sample size (41 individuals at an ARMS), de Wit and colleagues demonstrated a lower accuracy of 73% in predicting the same type of outcome (i.e. poor *vs.* good functioning at follow-up) from regional brain gyrification indexes (de Wit et al., 2017). Moreover, Koutsouleris and colleagues showed, by using a substantial larger sample of 116 subjects from 7 different sites, that a grey matter volume pattern can predict poor *vs.* good social and role functioning at follow-up of individuals at ARMS with an accuracy of 77% and 57%, respectively (Koutsouleris et al., 2018). Additionally, a few studies have explored structural MRI to inform quantitative prediction of symptom or functioning progression in individuals at an ARMS with reported correlation coefficients of .34 when using cortical thickness patterns (Stefania Tognin et al., 2014) and .42 when using regional brain gyrification indexes (de Wit et al., 2017).

The feasibility of prediction transition to psychosis from an ARMS using structural MRI and ML have been consistently demonstrated. For instance, Koutsouleris and colleagues showed that by assessing patterns of whole-brain gray matter volume abnormalities, transition to psychosis from an ARMS could be predicted with an accuracy of 82% (Koutsouleris et al., 2009). These findings were later replicated a) in an independent sample (balanced accuracy = 84%) (Koutsouleris et al., 2012); b) combining both samples (balanced accuracy = 80%) (Koutsouleris, Riecher-Rössler, et al., 2015); using one of the two samples and c) graph-extracted network measures from cortical gyrification (balanced accuracy = 81%) (Das et al., 2018), or d) patterns of regional gray matter abnormalities (balanced accuracy = 74%)

(Zarogianni et al., 2019); and e) using an independent sample and assessing patterns of regional gray matter abnormalities (balanced accuracy = 77%; specificity of a replication sample of individuals at ARMS that did not developed psychosis = 68%) (Zarogianni et al., 2017). However, the reliability of the findings is unclear due to possible sampling bias, i.e. small sample size (minimum size = 33 individuals at an ARMS, maximum size = 79 individuals at an ARMS), data from one single site or a combination of data from more than one site that is not completely independent from previous reports (making them not accurately independent replications).

Herein, I sought to replicate the previous findings regarding the prediction of transition to psychosis using sMRI data of a relatively large and independent sample of 99 individuals at an ARMS from three different scan acquisition protocols. First, I aim to use this study sample as a validation sample to previous prediction models which have demonstrated to be successful (i.e. showing high accuracies in predicting transition to psychosis from ARMS). For this task I chose the study using the largest sample (66 individuals with ARMS) from Koutsouleris and colleagues (Koutsouleris, Riecher-Rössler, et al., 2015), which combines data combining data from two previously published datasets (Koutsouleris et al., 2009, 2012). Then, using an updated sMRI data processing and machine learning pipelines, several combinations of brain structural measures (i.e. whole-brain or regional gray and white matters volume, and regional cortical thickness, brain gyrification, sulci depth and complexity indexes), feature manipulation (i.e. none, feature dimensionality reduction using principal component analysis, or forward feature selection) and cross-validation strategies [leave-one scan acquisition protocol-out, leave-one subject per group-out, or 5-fold nested-cross validation (CV)]. Moreover, to the best of my knowledge I am the first one using whole brain and regional white matter volume and a combination of regional cortical thickness and brain gyrification, sulci depth and complexity indexes with ML methods to predict transition to psychosis from an ARMS at an individual level.

Table 10. Studies using structural neuroimaging and machine learning methods to predict the follow-up functioning of individuals at an ARMS (i.e. with prognosis as the outcome measure).

Study	**Sample** (ARMS criteria; site/project)	**Data** (preprocessing software; features)	**ML approach** (feature manipulation; algorithm; validation strategy; statistical significance test)	**Performance metrics**
(Stefania Tognin et al., 2014)	40 ARMS (CAARMS; OASIS)	1.5T sMRI (SPM8; voxel-based CT map)	none; relevance vector regression$^{\downarrow}$; LOO CV; 1000 permutations	r = .34 p(r) = .026* MSE = 249.63 p(MSE) = .024*
		1.5T sMRI (SPM8; voxel-based GMV map)	none; relevance vector regression$^{\downarrow}$; LOO CV; 1000 permutations	r = .14 p(r) = .627 MSE = 369.50 p(MSE) = .621
(Kambeitz-Ilankovic et al., 2016)	14 ARMS–GAF- vs. 13 ARMS–GAF+ (PACE; FETZ)	1.5T sMRI (FreeSurfer; 64 regional cortical surface areas normalized by the total cortical surface)	none; linear L1-regularized multivariate logistic regression; LOO nested-CV; 5000 permutations	BAC = 82% SE = 79%, SP = 85% p = .050*
(de Wit et al., 2017)	17 ARMS–GAF+ vs. 24 ARMS–GAF- (SIPS, FIGS and BSABS-P; BCRM)	1.5T sMRI (FreeSurfer; 30 regional subcortical GMV with correction for the effects of age and gender by regressing them out)	none; linear SVM; LOO CV; 1000 permutations	ACC = 67% SE = 59%, SP = 75% p = .047*
		1.5T sMRI (FreeSurfer; 70 regional surface areas with correction for the effects of age and gender by regressing them out)	none; linear SVM; LOO CV; 1000 permutations	ACC = 44% SE = 47%, SP = 42% p > .050
		1.5T sMRI (FreeSurfer; 68 regional CT with correction for the effects of age and gender by regressing them out)	none; linear SVM; LOO CV; 1000 permutations	ACC = 49% SE = 47%, SP = 50% p > .050
		1.5T sMRI	none; linear SVM;	ACC = 69% SE = 71%, SP = 67%

		(FreeSurfer; 75 regional cortical GMV with correction for the effects of age and gender by regressing them out)	LOO CV; 1000 permutations	p = .015*
		1.5T sMRI (FreeSurfer; 68 regional gyrification indexes with correction for the effects of age and gender by regressing them out)	none; linear SVM; LOO CV; 1000 permutations	ACC = 73% SE = 69%, SP = 78% p = .016*
	41 ARMS (SIPS, FIGS and BSABS-P; BCRM)	1.5T sMRI (FreeSurfer; 30 regional subcortical GMV with correction for the effects of age and gender by regressing them out)	none; linear SVR$^{\downarrow\downarrow}$; LOO CV; 1000 permutations	r = .38 p = .048*
		1.5T sMRI (FreeSurfer; 68 regional gyrification indexes with correction for the effects of age and gender by regressing them out)	none; linear SVR$^{\downarrow\downarrow}$; LOO CV; 1000 permutations	r = .42 p = .008*
(Koutsouleris et al., 2018)	66 ARMS–SocialF- vs. 50 ARMS–SocialF+ (SPI-A and/or SIPS; PRONIA consortium – 7 sites)	1.5T (1 site) and 3T (6 sites) sMRI (SPM12/CAT12; voxel-based GMV maps)	feature dimensionality reduction with PCA and feature selection using a greedy sequential backward elimination algorithm; linear SVM LSO nested-CV 1000 permutations	BAC = 77% SE = 70%, SP = 84% p < .001***
	69 ARMS–RoleF- vs. 47 ARMS–RoleF+ (SPI-A and/or SIPS; PRONIA consortium – 7 sites)	1.5T (1 site) and 3T (6 sites) sMRI (SPM12/CAT12; voxel-based GMV maps)	feature dimensionality reduction with PCA and feature selection using a greedy sequential backward elimination algorithm; linear SVM LSO nested-CV 1000 permutations	BAC = 57% SE = 67%, SP = 47% p < .001***

Abbreviations: ACC: accuracy; ARMS: at-risk mental state; ARMS-GAF-: individuals at ARMS showing an impaired global functioning at follow-up; ARMS-GAF-: individuals at ARMS showing a good global functioning at follow-up; ARMS-RoleF-: individuals at ARMS showing an impaired role functioning at follow-up; ARMS-RoleF+: individuals at ARMS showing a good role functioning at follow-up; ARMS-SocialF-: individuals at ARMS showing an impaired social functioning at follow-up; ARMS-SocialF+: individuals at ARMS showing a good social functioning at follow-up; BAC: balanced accuracy; BCRM: Department of Psychiatry at the University Medical Center Utrecht, Brain Center Rudolf Magnus in the Netherlands; BSABS-P: Bonn Scale for the Assessment of Basic Symptoms - Prediction List; CV: cross-validation; FETZ: Early Detection and Intervention Centre for Mental Crises, Ludwig-Maximilians-University; FIGS: Family Interview for Genetic Studies; GMV: gray matter volume; LOO: leave-one-out; OASIS: Outreach and Support in South London service in London, United Kingdom; PACE: Personal Assessment and Crisis Evaluation; RVR: relevance vector regression; SE: sensitivity; SIPS: Structured Interview for Prodromal Syndromes; sMRI: structural magnetic resonance imaging; SP: specificity; SPI-A: Schizophrenia Proneness

Instrument; SVM: support vector machines; SVR: * p < .050; ** p < .010; *** p < .001; ↓regressed against the positive and negative syndrome scale total scores; ↓↓regressed against the modified global assessment functioning scores.

Table 11. Studies using structural neuroimaging and machine learning methods to predict the transition to psychosis from an ARMS (i.e. with prognosis as the outcome measure).

Study	**Sample** (ARMS criteria; site/project)	**Data** (preprocessing software; features)	**ML approach** (feature manipulation; algorithm; validation strategy; statistical significance test)	**Performance metrics**
(Koutsouleris et al., 2009)	15 ARMS-T vs. 18 ARMS-NT (PACE; FETZ)	1.5T sMRI (SPM5/VBM5; voxel-based GMV maps scaled to the TIV with correction for age and sex with partial correlations. Feature dimensionality reduction with PCA)	none; non-linear SVM; 5-fold CV; 5000 permutations	ACC = 82% SE = 83%, SP = 80% p < .001***
(Koutsouleris et al., 2012)	16 ARMS-T vs. 21 ARMS-NT (BSIP and SANS; FePsy)	1.5T sMRI (SPM5/VBM5; voxel-based GMV maps scaled to the TIV)	correction for age and sex with partial correlations and feature dimensionality reduction with PCA; non-linear SVM; 10-fold nested-CV; none	BAC = 84% SE = 81%, SP = 88%
(Koutsouleris, Riecher-Rössler, et al., 2015)	33 ARMS-T vs. 33 ARMS-NT (PACE or BSIP and SANS; FETZ and FePsy)	1.5T sMRI (SPM8/VBM8; voxel-based GMV maps)	correction for study center effects using partial correlations and feature dimensionality reduction with PCA; linear SVM; LPO nested-CV; none	BAC = 80% SE = 76%, SP = 85%
(Zarogianni et al., 2017)	17 ARMS-T vs. 17 ARMS-NT (familial high-risk; EHRS)	1.0T sMRI (SPM5/VBM5; 116 regional GMD)	recursive feature elimination; linear SVM; LOO nested-CV; 1000 permutations	BAC = 77% SE = 76%, SP = 77% p = .003**
	40 ARMS-NT (validation sample)	1.0T sMRI (SPM5/VBM5; 116 regional GMD)		SP = 68%

	(familial high-risk; EHRS)			
(Das et al., 2018)	16 ARMS-T vs. 63 ARMS-NT (BSIP, BPRS, and SANS; FETZ)	3T sMRI (FreeSurfer; 6 graph-extracted network measures from cortical gyrification)	unspecified correction for age, sex, and TIV effects; randomized trees with class-weighing; 5-fold CV with 30 repetitions; 5000 permutations	BAC = 81% SE = 66%, SP = 97% p <.001***
(Zarogianni et al., 2019)	16 ARMS-T vs. 19 ARMS-NT (BSIP, BPRS, and SANS; FePsy)	1.5T sMRI (SPM5/VBM5; 116 regional GMD)	recursive feature elimination; linear SVM; LOO nested-CV; 1000 permutations	BAC = 74% SE = 63%, SP = 84% p = .002**

Abbreviations: ACC: accuracy; ARMS: at-risk mental state; ARMS-T: individuals at ARMS that later transitioned to psychosis; ARMS-NT: individuals at ARMS that did not transitioned to psychosis; BAC: balanced accuracy; BCRM: Department of Psychiatry at the University Medical Center Utrecht, Brain Center Rudolf Magnus in the Netherlands; BPRS: Brief Psychiatric Rating Scale; BSIP: Basel Screening Instrument for Psychosis; CV: cross-validation; EHRS: Edinburgh High Risk Study; FePsy: early detection of psychoses; FETZ: Early Detection and Intervention Centre for Mental Crises, Ludwig-Maximilians-University; GMD: gray matter density; GMV: gray matter volume; LOO: leave-one-out; LPO: leave-pair-out; PACE: Personal Assessment and Crisis Evaluation; SANS: Scale for the Assessment of Negative Symptoms; SE: sensitivity; SIPS: Structured Interview for Prodromal Syndromes; sMRI: structural magnetic resonance imaging; SP: specificity; SVM: support vector machines; * $p < .050$; ** $p < .010$; *** $p < .001$.

3.2. Materials and methods

3.2.1. Sample description

The total sample consisted of 246 individuals at an ARMS, recruited at first presentation from consecutive referrals to the Outreach and Support in South London (OASIS) high-risk service, South London and Maudsley NHS Foundation Trust (Broome et al., 2005). OASIS is a clinical service located in Lambeth, South London, that offers treatment to individuals between 14 and 35 years of age who meet the ARMS criteria for psychosis. The presence of the ARMS was assessed using the comprehensive assessment of the ARMS (CAARMS), a detailed clinical assessment (Phillips et al., 2000). In particular, an individual can meet the criteria for the ARMS if 1) had a recent decline in function coupled with either schizotypal personality disorder or a first degree relative with psychosis, examined using the family interview for genetic studies (FIGS) (Nurnberger Jr et al., 1994); or 2) experienced attenuated positive psychotic symptoms; or 3) had a Brief Limited Intermittent Psychosis (BLIP), a brief psychotic episode of less than 1 week's duration that resolved without antipsychotic medication. Exclusion criteria applied to all participants was a history of significant head injury and current (in the last 12 months) substance dependency according to DSM-IV diagnostic criteria.

When the subjects were first diagnosed as being at an ARMS (i.e., baseline) a set of data were acquired: a) a sMRI scan; b) genome-wide genotypes; and c) environmental risk factors assessments. Subjects were labeled as transitioned to psychosis (ARMS-T) if they later presented a first episode of psychosis (FEP) or as not-transitioned to psychosis (ARMS-NT) if they did not present a FEP within at least a period of 2 years. Furthermore, transition to psychosis during the follow-up period was established according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) (American Psychiatric Association, 2000) criteria based on clinical consensus between at least two experienced psychiatrists. Additional socio-demographic and clinical measures were also assessed at baseline, including age, sex, handedness, self-reported ethnicity, full scale intelligence quotient measured by National Adult Reading Test (Nelson, 1982), years of education, and global assessment of function using Global Assessment of Functioning tool (GAF; at baseline and at follow-up) (American Psychiatric Association, 2000), and CAARMS (at baseline and follow-up) (Phillips et al., 2000). Socio-demographic and clinical variables were analyzed using Univariate Analysis of Variance (ANOVA) for continuous data and a chi square test or Fisher's exact test (if there were less than 5 subjects in one group) for

ordinal data (**Table 12**). These statistical analyses were performed using the Statistical Package for the Social Sciences 26 (SPSS 26 for Windows, Chicago, IL, USA).

From the pool of 246 individuals at an ARMS, 99 had a baseline sMRI scan acquired up to six months after the follow-up of transition dates. Over the 2-year follow-up period, 23 individuals at an ARMS developed psychosis (ARMS-T) and the remaining 76 did not (ARMS-NT). Two ARMS-T were scanned 68 and 136 days after the transition to psychosis date. Most of the ARMS group (84/99; 85%) were naïve to antipsychotics at the time of scanning; the remaining 15 (15%) had been exposed to antipsychotics.

3.2.2. Structural magnetic resonance imaging

sMRI scans were acquired with two different scanners using three enhanced fast gradient echo 3-Dimensional (efgre3D) protocols: a) acquisition protocol 1 – 1.5T Signa scanner (General Electric Medical Systems, USA; voxel size = 0.9 x 0.9 x 1.5 mm^3; matrix of acquisition = 256 x 256 x 124; field-of-view = 220 mm; gap = 0 mm; repetition/echo/inversion times = 15.9 s/5.2 s/300 s; flip angle = 20°; 19 scans: 3 ARMS-T, 16 ARMS-NT); b) acquisition protocol 2 – 1.5T Signa scanner (General Electric Medical Systems, USA; voxel size = 0.9 x 0.9 x 1.5 mm^3; matrix of acquisition = 256 x 256 x 124; field-of-view = 220 mm; gap = 0 mm; repetition/echo/inversion times = 21.3 s/5.1 s/0 s; flip angle = 20°; 33 scans: 14 ARMS-T, 19 ARMS-NT); and c) acquisition protocol 3 – 3T Signa scanner (General Electric Medical Systems, USA; voxel size = 1.1 x 1.1 x 1.1 mm^3; matrix of acquisition = 256 x 256 x 146~196; field-of-view = 280 mm; gap = 1.1 mm; repetition/echo/inversion times = 7.0~9.6 s/2.8~2.9 s/450 s; flip angle = 20°; 33 scans: 14 ARMS-T, 19 ARMS-NT).

Table 12. Socio-demographic and clinical information of the ARMS sample with structural MRI data.

	Protocol 1		**Protocol 2**		**Protocol 3**		**Group comparison**
	ARMS-T (n = 14)	**ARMS-NT (n = 19)**	**ARMS-T (n = 3)**	**ARMS-NT (n = 16)**	**ARMS-T (n = 6)**	**ARMS-NT (n = 41)**	
Age at baseline (years)	23.2 ± 3.4 [15.6 26.9]	24.5 ± 4.8 [19.2 34.5]	26.2 ± 7.0 [20.1 33.8]	24.5 ± 5.2 [17.8 35.3]	23.4 ± 4.5 [17.5 29.2]	21.8 ± 4.3 [17.1 33.1]	Protocol: p = .271 Transition: p = .592 Protocol x Transition: p = .447
Age at follow-up or transition (years)	25.6 ± 4.2 [17.3 33.4]	32.7 ± 5.2 [22.6 43.9]	29.2 ± 5.4 [20.2 38.8]	28.8 ± 5.6 [22.9 43.1]	25.2 ± 4.8 [18.3 31.0]	25.6 ± 4.8 [20.3 41.2]	Protocol: p = .027* Transition: p = .099 Protocol x Transition: p = .025*
Age at scan (years)	23.0 ± 3.6 [17.5 27.8]	23.9 ± 4.8 [18.5 34.8]	27.0 ± 8.2 [20.2 36.1]	25.1 ± 5.4 [18.6 37.4]	24.1 ± 4.8 [18.3 30.8]	22.4 ± 4.6 [17.7 38.3]	Protocol: p = .261 Transition: p = .499 Protocol x Transition: p = .541
Interval between baseline and scan age (years)	-0.2 ± 1.4 [-2.3 1.9]	-0.5 ± 1.1 [-2.3 2.1]	0.9 ± 1.3 [0.1 2.4]	0.5 ± 0.5 [0.1 2.1]	0.6 ± 0.5 [0.2 1.6]	0.6 ± 1.0 [0.0 5.1]	Protocol: p < .001*** Transition: p = .419 Protocol x Transition: p = .795
Sex (male/female)	11/3	9/10	2/1	14/2	3/3	19/22	Protocol x Transition: Protocol 1: p = .070 Protocol 2: p = .422 Protocol 3: p = 1

Handedness [a] (right/left/ambidextrous)	12/0/1	16/0/0	3/0/0	13/1/0	4/0/0	36/4/0	Protocol x Transition: Protocol 1: p = .448 Protocol 2: p = 1 Protocol 3: p = 1
Self-reported ethnicity (white/black/asian/other)	7/5/1/1	11/5/1/2	2/1/0/0	13/1/1/1	4/1/1/0	19/19/1/2	Protocol x Transition: Protocol 1: p = .933 Protocol 2: p = .530 Protocol 3: p =. 212
Years of education	13.4 ± 2.1 [10 18]	13.1 ± 1.9 [10 17]	11.7 ± 2.3 [9 13]	14.1 ± 2.6 [11 20]	15.2 ± 2.5 [11 18]	13.0 ± 2.2 [9 20]	Protocol: p = .298 Transition: p = .966 Protocol x Transition: p = .024*
IQ (z-standardized) [b]	-1.1 ± 1.1 [-2.1 1.0]	0.0 ± 1.1 [-2.1 1.8]	0.1 ± 0.1 [0.0 0.2]	0.5 ± 0.9 [-1.3 1.6]	-0.1 ± 1.3 [-2.1 1.6]	0.1 ± 1.1 [-2.1 3.5]	Protocol: p = .427 Transition: p = .252 Protocol x Transition: p = .923
GAF at baseline	52.9 ±16.0 [35 90]	57.8 ± 11.4 [35 75]	58.7 ± 3.2 [55 61]	58.6 ± 9.9 [41 75]	50.3 ± 11.4 [35 65]	53.6 ±14.8 [0 75]	Protocol: p = .402 Transition: p = .475 Protocol x Transition: p = .877
GAF at follow-up [c]	49.3 ± 18.6 [10 69]	58.5 ± 17.9 [20 94]	27.3 ± 6.8 [22 35]	62.3 ± 13.5 [46 93]	52.5 ± 20.0 [30 78]	66.2 ± 13.6 [33 87]	Protocol: p = .064 Transition: p < .001*** Protocol x Transition: p = .095
CAARMS at baseline [d]	33.2 ± 13.0	28.4 ± 15.3	29.3 ± 21.9	23.2 ± 14.3	39.7 ± 24.1	28.5 ± 16.7	Protocol: p = .505

	[9 56]	[8 58]	[12 54]	[0 51]	[0 69]	[0 81]	Transition: p = .153 Protocol x Transition: p = .824
CAARMS at follow-up [f]	19.6 ± 23.0 [0 63]	11.6 ± 10.9 [0 31]	42.0 ± 43.3 [6 90]	14.7 ± 18.4 [0 54]	42.7 ± 42.1 [0 102]	15.5 ± 17.2 [0 60]	Protocol: p = .082 Transition: p = .001*** Protocol x Transition: p = .262

Data format: mean ± standard deviation [min max]. Information not available for [a]1 ARMS-T and 3 ARMS-NT (Protocol 1), 2 ARMS-NT (Protocol 2), 2 ARMS-T and 1 ARMS-NT (Protocol 3); [b]1 ARMS-T and 1 ARMS-NT (Protocol 2), 1 ARMS-NT (Protocol 3); [c]2 ARMS and 5ARMS-NT (Protocol 1), 4 ARMS-NT (Protocol 2), 8 ARMS-NT (Protocol 3); [d]2 ARMS-T and 7 ARMS-NT (Protocol 1), 3 ARMS-NT (Protocol 2), 2 ARMS-NT (Protocol 3); [f]3 ARMS-T and 6 ARMS-NT (Protocol 1), 3 ARMS-NT (Protocol 2), 8 ARMS-NT (Protocol 3). ARMS: at-risk mental state; ARMS-T: individuals at ARMS that later transitioned to psychosis; ARMS-NT: individuals at ARMS that did not transitioned to psychosis. *p<.05; **p<.01; ***p<.001

3.2.3. Image processing

T1w images were processed with CAT12 (v1092, http://www.neuro.uni-jena.de/cat/), a SPM12 add-on (v6909, http://www.fil.ion.ucl.ac.uk/spm/) using default settings and MATLAB (9.3) – i.e. using the same pipeline as described in **Comparing SPM12 with CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study** – chapter **2**). In summary, the preprocessing included bias field inhomogeneity correction; image segmentation into grey matter, white matter, and cerebrospinal fluid; spatial normalization to a template derived from 555 healthy subjects of the IXI-database (http://www.brain-development.org) using the DARTEL algorithm (Ashburner, 2007); and a spatial smoothing of Full Width at Half Maximum (FWHM) of 8 mm. Finally, grey and white matter volumes for 64 regions of interest (ROI; description of each ROI is in the **Table 13**) were computed by dividing the volume of grey and white matter of each ROI (i.e. the sum of all voxels classified as grey or white matter inside that ROI, respectively) by the total intracranial volume (TIV; the sum of all voxels classified as grey or white matter or as cerebrospinal fluid). ROIs were defined by the Hammers atlas (Hammers et al., 2003).

Table 13. Regions of interest (ROI) for which the grey and white matter volume were extracted. Volumes were extracted for the left and right side of each ROI and using the Hammers atlas.

Hippocampus	Inferolateral remainder of parietal lobe
Amygdala	Caudate nucleus
Anterior temporal lobe, medial part	Nucleus accumbens
Anterior temporal lobe, lateral part	Putamen
Parahippocampal and ambient gyri	Thalamus
Superior temporal gyrus	Pallidum
Middle and inferior temporal gyri	Corpus callosum
Fusiform gyrus	Precentral gyrus
Cerebellum	Gyrus rectus
Brainstem	Orbitofrontal gyri
Insula	Inferior frontal gyrus
Lateral remainder of occipital lobe	Superior frontal gyrus
Gyrus cinguli, anterior part	Postcentral gyrus
Gyrus cinguli, posterior part	Superior parietal gyrus
Middle frontal gyrus	Lingual gyrus
Posterior temporal lobe	Cuneus

Additionally, the segmented images (i.e. grey and white matter images) were used to estimate cortical thickness and central surface using the projection-based thickness method

(Dahnke, Yotter, et al., 2013). In summary, this method uses grey matter voxel-wise segmentation image to estimate the distance between the inner boundary (i.e. white matter – grey matter) and the outer boundary (i.e. grey matter – cerebrospinal fluid). Then projects the local maxima (which is equal to the cortical thickness) to other gray matter voxels by using a neighbor relationship described by the white matter distance (i.e. from the inner boundary). This central surface is generated at the 50% distance between gray matter inner and outer boundaries. The projection-based thickness method allows the handling of partial volume information, sulcal blurring, and sulcal asymmetries without explicit sulcus reconstruction. For a more detailed description of this method please refer to (Dahnke, Yotter, et al., 2013). Moreover, the central surface was corrected for topology defects using spherical harmonics (Rachel Aine Yotter et al., 2011), spherically mapped into a common coordinate system to allow inter-subject analysis (Rachel A. Yotter, Thompson, et al., 2011), and spherically registered using an adaptation of the DARTEL algorithm (Ashburner, 2007) to spherical maps. Finally, regional-based cortical thickness and surface measures (i.e. folding measures) – gyrification index, i.e. the absolute mean curvature (Luders et al., 2006), the depth of sulci and the measurement of local surface complexity, i.e. the fractal dimension (Rachel A. Yotter, Nenadic, et al., 2011), were extracted for 68 ROIs (description of each ROI is in the **Table 14**) defined by the Desikan-Killiany atlas (Desikan et al., 2006).

Table 14. Regions of interest (ROI) for which cortical thickness, gyrification index, depth of sulci, and local surface complexity were extracted. These measures were extracted for the left and right side of each ROI and using the Desikan-Killiany atlas.

Banks of the superior temporal sulcus	Medial orbital frontal cortex	Rostral anterior cingulate cortex
Caudal anterior-cingulate cortex	Middle temporal gyrus	Rostral middle frontal gyrus
Caudal middle frontal gyrus	Parahippocampal gyrus	Superior frontal gyrus
Cuneus cortex	Paracentral lobule	Superior parietal cortex
Entorhinal cortex	Pars opercularis	Superior temporal gyrus
Fusiform gyrus	Pars orbitalis	Supramarginal gyrus
Inferior parietal cortex	Pars triangularis	Frontal pole
Inferior temporal gyrus	Pericalcarine cortex	Temporal pole
Isthmus–cingulate cortex	Postcentral gyrus	Transverse temporal cortex
Lateral occipital cortex	Posterior-cingulate cortex	Insula
Lateral orbital frontal cortex	Precentral gyrus	
Lingual gyrus	Precuneus cortex	

3.2.4. Image quality control

The quality of each processed image (which is decomposed in several volume- and surface-based brain measures as described above) was empirically assessed using the quality assurance framework of CAT12. This framework describes the properties of the image before CAT12 processing and estimates four qualitative measures using the tissue segmentation results: a) noise contrast ratio (NCR) – it yields information regarding the amount of noise in the image by measuring the local standard deviation in the optimized white matter segment scaled by the minimum tissue contrast; b) inhomogeneity contrast ratio (ICR) – it yields information regarding the bias field inhomogeneity measured as the global standard deviation within the optimized white matter segment scaled by the minimum tissue contrast; c) root-mean-squared resolution (RES) – it described the resolution of the image by measuring the root-mean-square of the voxel size; and d) weighted average image quality rating (IQR) – averaging the above three measures. The quality ascertainment framework maps the rating scores to image quality grades: from A (excellent) to F (unacceptable/failed). We set the subject's image inclusion threshold at D (sufficient), i.e. only subjects which images had a IQR of A to D were included in the final sample, as it has been shown that typical scientific data (clinical) data get good to satisfactory ratings (Dahnke, Ziegler, et al., 2013). NCR, ICR, RES and IQR were analyzed using Univariate ANOVA to test for the effects of 'Protocol', 'Transition' and 'Protocol x Transition' using SPSS.

3.2.5. Machine learning approach

3.2.5.1. Previous model-based approach

I intended to use this study's ARMS sample as an external and independent validation sample to the transition to psychosis from ARMS prediction model published by Koutsouleris and colleagues (Koutsouleris, Meisenzahl, et al., 2015). For this purpose, a collaboration was established in the summer of 2018 between the group of Dr. Diana Prata (Biomedical Neurosciences Lab) from the Institute of Biomedical Engineering and Biophysics, Faculty of Sciences, University of Lisbon, Portugal, and the group of Dr. Nikolaos Koutsouleris (Neurodiagnostische Verfahren Lab - Neurodiagnostic Applications Lab) from the Ludwig-Maximillian's University of Munich, Germany. To accomplish this task a member of our group, Dr. Vasco Diogo (Dr. Diana Prata's PhD student) was hosted by Dr. Koutsouleris's group under a short-term mission (with a duration of 1 month, in February of 2020). Three attempts to accomplish this task were made. First, we intended to

retrieve the prediction model used to produce the results reported in (Koutsouleris, Meisenzahl, et al., 2015). However, our group was told that this model was, indeed, built with an early and bugged version of the ML software NeuroMiner [see below a summarized description of the ML strategy used in (Koutsouleris, Meisenzahl, et al., 2015) to build the prediction model]. Therefore, a direct validation of this already published transition to psychosis model was not possible. Second, we tried to retrieve the data (i.e. FETZ and FePsy) that was used to train the previous prediction model [see below a summarized description of the data used in (Koutsouleris, Meisenzahl, et al., 2015) to build the prediction model] and to re-train the model using the same ML pipeline, but using the newer and improved version of NeuroMiner. Unfortunately, not all subjects from the sample that was previously used were available. Therefore, neither a replication of the already published transition to psychosis model, nor a validation of this re-trained model were possible. Third, as a last attempt of replicate the previous findings, we tried to train a new prediction model using a) the data that was previously published (i.e. FETZ and FePsy) and still available; b) the same preprocessing pipeline that we used with our ARMS sample (an updated preprocessing pipeline compared to the previous publication; see below a summarized description of the image processing pipeline used in (Koutsouleris, Meisenzahl, et al., 2015) to build the prediction model and **3.2.3. Image processing** for the updated one); and c) the same ML pipeline previously used. Regrettably, this was not possible to be accomplished within the duration of the short-term mission of Dr. Vasco Diogo due to logistics constraints. In particular, there was no physical computational resources to run the analysis at the time and there is no authorization to physically store the FETZ or FePsy data at another location or to access it outside the physical installations of Dr. Koutsouleris's group (i.e., access it remotely). Efforts are still being made to build a cross-sample validated transition to psychosis prediction model under this collaboration with Dr. Koutsouleris.

In summary, a sample composed by 66 individuals at an ARMS (33 ARMS-T – 16 and 17 from FePsy and FETZ datasets, respectively; 33 age-, gender-, and years of education-matched ARMS-NT – 16 and 17 from FePsy and FETZ datasets, respectively) was used to build the previously published transition to psychosis prediction model (Koutsouleris, Riecher-Rössler, et al., 2015). sMRI data from these subjects were processed using VBM8 toolbox (a previous version of CAT12; http://dbm.neuro.uni-jena.de/wordpress/vbm/), a SPM8 add-on (http://www.fil.ion.ucl.ac.uk/spm/). Voxel-based gray matter (VBGM) maps were extracted using a similar approach as described in **3.2.3 Image processing**, except the

spatial smoothing (that was performed using a FWHM of 4 mm). For the ML analysis NeuroMiner toolbox (http://proniapredictors.eu/neurominer/index.html) was used and it consisted in the following steps. First, each VBGM were corrected for center effects (i.e., FETZ of FePsy) using partial correlations and scaled voxel-wise to the range [0 1]. Second, a two-step feature dimensionality reduction was performed by a) extracting sparse, nonredundant voxel sets according to their weighted contribution to the geometric distance between the two groups (i.e. ARMS-T and ARMS-NT) (Yijun Sun et al., 2010); and b) extracting uncorrelated principal components using robust principal component analysis (PCA) (Hubert et al., 2005, 2009) (see **3.2.5.2.3. Feature manipulation** section for a detailed description of robust PCA). Third, these principal components were entered as features in a linear SVM algorithm (Bernhard Scholkopf et al., 2000) (for a detailed description of the SVM algorithm see **3.2.5.2.4. Support vector machines classification section**). Forth, the model was trained and tested using a leave-one-pair-out nested CV approach. In the inner CV cycle the whole sample, except one subject per group (i.e. ARMS-T and ARMS-NT) was used to tune the model hyperparameters – the optimal number of principal components and the SVM algorithm's ν parameter – and to generate an ensemble classifier from all the SVM base learners trained in each fold of the inner CV cycle. The membership of the two subjects left out in the outer CV cycle was decided by majority voting of each ensemble classifiers from the inner CV cycle. Lastly, the performance of the prediction model was assessed by measuring the balanced accuracy, the sensitivity, and the specificity (see **3.2.5.2.6. Performance measures** for a detailed description of performance measures) in classifying all the left-out subjects from the outer CV cycle.

3.2.5.2. Novel model-based approach

Several ML strategies to generate prediction models for transition to psychosis from sMRI data using our ARMS sample were followed. These include a) sample balancing and bootstrapping; testing several b) feature types; c) feature manipulation approaches; and d) cross-validation approaches. Furthermore, the analyses were conducted using the neuroimaging ML tool NeuroMiner v1.0 ELESSAR (http://proniapredictors.eu/neurominer/index.html)

3.2.5.2.1. Sample balancing and bootstrapping

The final sample used in the ML analyses was composed by the 23 ARMS-T subjects and 23 ARMS-NT subjects randomly selected to match the ARMS-T for age, sex, and scan acquisition protocol. The matching criteria for age and sex was based on the non-rejection of the null hypothesis that the ARMS-T and ARMS-NT groups had the same median age (tested with a two-sided Mann-Whitney U-test) and sex (tested with a chi-square test). The matching for the scan acquisition protocol was done in a one-to-one manner, i.e., the number of ARMS-NT subjects within each protocol is the same as the number of ARMS-T. The subsampling was repeated 5 times, i.e., 5 bootstrapped samples were created, and the subsequent ML analyses were conducted for each of the bootstrapped sample.

3.2.5.2.2. Feature types

Individual ML models (i.e. classifiers) were trained and validated for each of the following brain measures: a) VBGM maps (297 811 initial features); b) voxel-based white matter (VBWM) maps (204 706 initial features); c) regional-based grey (ROIGM) and d) white (ROIWM) matter volumes (each with 64 initial features) scaled to the TIV; and d) surface based regional cortical thickness, and gyrification, sulci, and complexity indexes (ROISurface; 272 initial features). All the features were scaled between 0 and 1 before entering the SVM classification algorithm.

3.2.5.2.3. Feature manipulation

Feature dimensionality reduction was performed for VBGM and VBWM features using robust PCA (Hubert et al., 2005, 2009). This algorithm projects the correlated voxels within the voxel-based maps to a given number of uncorrelated principal components, which represent compact sets of neuroanatomical features with reduced content of noise. Furthermore, these principal components are linear combinations of the correlated voxels computed to maximize the variance of the data. Robust PCA is specially designed to deal with non-symmetrical data, i.e., data under the influence of outliers, and is more suitable when the data is characterized by a much higher number of features than observations, such as herein. Moreover, the main advantages of PCA are the reduction of a) the computational complexity of classification caused by the high dimensionality of structural MRI data; and b) the generalization error of classification by optimizing the number of principal

components for data projections, thus maximizing the degree of anatomical information while minimizing the impact of noise. Here the robust PCA was applied during the inner CV cycle (see the **3.2.5.2.5. Cross-validation** section). The number of principal components that were retained explained up to 80% of the variance in the data and were limited by the inner CV cycle's sample size, n, i.e., a maximum of only $n/2$ components could indeed be extracted. **Table 15** shows the maximum number of principal components that can be extracted for each inner CV cycle in each CV scheme that was used (see also the **Cross-validation** section).

Table 15. Number of principal components extracted from the data in each inner CV cycle in each CV scheme that was used (i.e., leave-one scan protocol-out CV, leave-one per group-out CV, 5-fold CV) and for each feature type (i.e., voxel-based grey (VBGM) or white (VBWM) matter volume maps) for which principal component analysis was used to reduce feature space dimensions. Both the maximum number of components that is possible to extract from data and the average number of components explaining up to 80% of the variance in the data per inner CV across bootstrapped samples that were indeed extracted are shown.

Inner CV cycle	**Principal components to extract** Maximum Mean (standard deviation)
Leave-one scan protocol-out CV scheme	
Inner CV cycle 1: Protocol 1 (n=28)	14 VBGM: 6.8 (0.4) VBWM: 6.8 (0.4)
Inner CV cycle 2: Protocol 2 (n=6)	3 VBGM: 1 (0) VBWM: 1 (0)
Inner CV cycle 3: Protocol 3 (n=12)	6 VBGM: 3 (0) VBWM: 2.4 (0.5)
Leave-one per group-out CV scheme	
Each inner CV cycle (n=42)	21 VBGM: 9.9 (0.0) VBWM: 9.6 (0.0)
5-fold CV scheme	
Each inner CV cycle (n=30)	15 VBGM: 3.6 (0.2) VBWM: 3.4 (0.4)

Feature selection was performed on regional brain features (i.e., ROIGM, ROIWM, and ROISurface) using a greedy forward search feature selection algorithm. This is a stepwise algorithm that starts with an empty set of features and then, tests the predictive value of every single feature, selecting the ones improving the overall accuracy across the inner CV cycle

folds (see the **3.2.5.2.5. Cross-validation** section). The final set of features is, then, composed by the 10% most predictive variables. Additionally, no feature selection, i.e., using the total number regional brain features, was also tested.

3.2.5.2.4. Support vector machine classification

Binary classification of transition to psychosis from an ARMS (i.e. ARMS-T *vs.* ARMS-NT) was performed using linear SVM (Burges, 1998; Cortes & Vapnik, 1995). SVM is a supervised classification algorithm that by learning from an initial training dataset how to best distinguish between two (or more) classes is able to classify new cases into those groups. The training dataset is composed by well-defined observations in the form $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i$ represents feature vectors (e.g., VBGM) and y_i represent class labels (e.g., $y_i = +1$ for ARMS-T and $y_i = -1$ for ARMS-NT) and is used to define a hyperplane that optimally separates the two classes. This hyperplane is, in turn, determined by the support vectors, which maximize the distance between the nearest data instances of opposite classes (**Figure 9**). Features will, therefore, be weighted according to their importance in the definition of these support vectors. When data is not optimally separable with a linear relationship, a kernel function can be used to transform the data into a higher dimensional space where classes become linearly separable. In this study we exclusively used a linear kernel SVM to reduce the risk of overfitting the data (given our final sample size being relatively small). Furthermore, the linear SVM classifier has a penalty parameter C that controls the trade-off between having zero training error and allowing misclassification. Larger C values will result in a preference of smaller-margin hyperplane if it yields a better training accuracy. On the other hand, small values of C will give preference to a larger-margin hyperplane, even if it will result in a lower training accuracy. Herein, a parameter search was carried out to identify the optimal C value (i.e., $2^l, l = [-5{:}\,1{:}\,4]$) in the inner CV cycle (see the **Cross-validation** section).

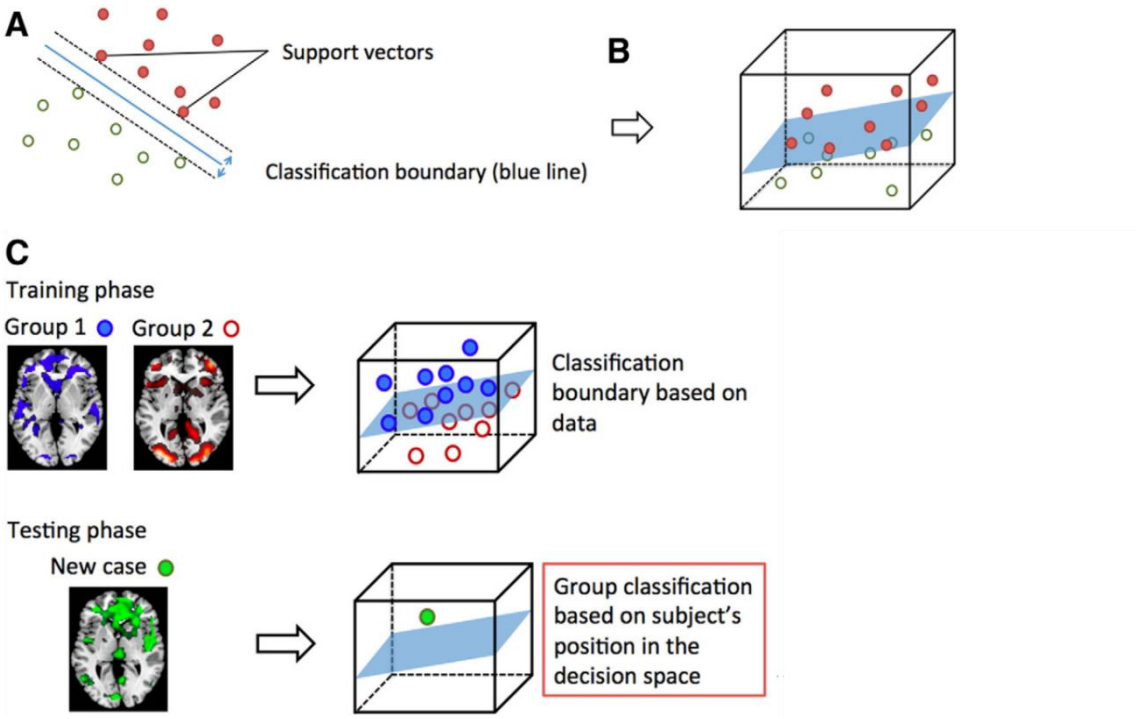


Figure 9. The support vector machine (SVM) classifier. **(A)** Illustration of a classification problem between two classes [e.g. individuals at an at-risk mental state that developed psychosis (ARMS-T) and individuals at an ARMS that did not (ARMS-NT)] for a two-dimensional case. Each brain image (e.g. voxel-based gray matter map) corresponds to a point in the input (i.e. feature) space and each voxel in the image represents one dimension of this space. The red circles represent the images of ARMS-T and the green circumferences images of ARMS-NT. The classification boundary (blue line) is created based on the maximum margin (dashed black lines) space between the data distributions of the two classes. Only data points near the margin affect the classification boundary and these are the support vectors. **(B)** If the feature space is composed by more than two dimensions (i.e., *k*-dimensional space), than the classification boundary will be a hyperplane that best separates the two classes in the *k*-dimensional space (represented by the blue plane). **(C)** The optimal classification boundary is discovered in the training phase based on information from the two classes. In the testing phase, new cases are positioned in the decision space and classified as belonging to either class. Adapted from (Gifford et al., 2015).

3.2.5.2.5. Cross-validation

Each classifier was trained in a nested-CV scheme (**Figure 10**; see also the section **1.2. Machine Learning** – chapter **1**). The sample was split into *k* non-overlapping folds and each fold was iteratively used to estimate the performance of the classifier on validation data, while the remaining k-1 folds were used to train the decision rule of the classifier. Specifically, the purpose of the inner CV cycle is to use the training and testing subsamples to optimize a) feature manipulation, i.e., to find the optimal number of principal components

to extract from the original feature set (i.e., feature dimension reduction with PCA when using VBGM or VBWM), or to select the optimal feature set (i.e., greedy forward feature selection when using ROIGM, ROIWM or ROISurface); and b) the linear SVM hyperparameter C. The outer CV cycle is used to estimate the generalizability of the trained prediction model by measuring the model performance on the validation subsample. Herein, I tested three different nested-CV schemes: a) leave-one scan acquisition protocol-out (LSO), b) leave-one per group from the same scan acquisition protocol-out (LPO), and b) classical 5-fold CV. Furthermore, the optimal hyperparameters/feature set were chosen as the ones yielding the highest balanced accuracy (see below) across the inner CV cycle.

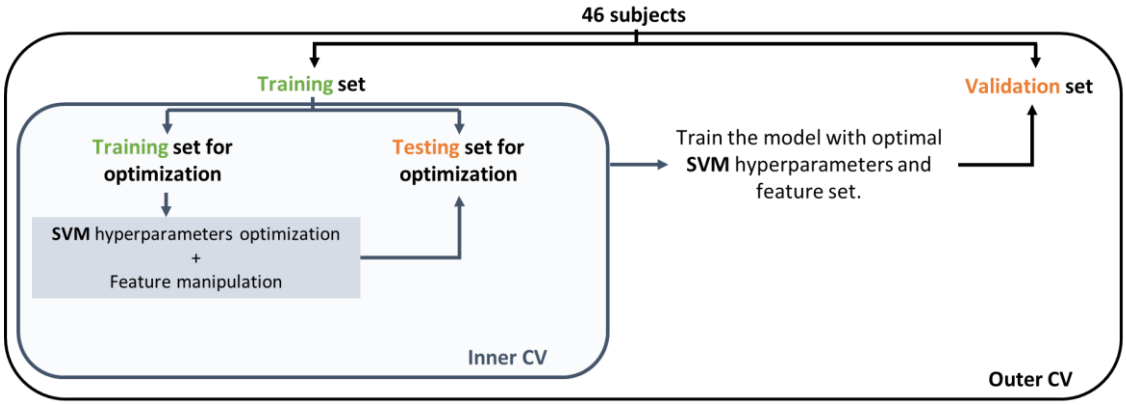


Figure 10. Nested-cross-validation scheme.

3.2.5.2.6. Performance measures

The classifier's performance was evaluated using measures derived from the confusion matrix (**Figure 11**): sensitivity (SE), specificity (SP), balanced accuracy (BAC), positive likelihood ratio (PLR), negative likelihood ratio (NLR), and diagnostic odds ratio (DOR). SE (also known as the true positive rate) reflects the proportion of ARMS-T individuals (positive class) that have been correctly identified (**Eq. 1**), whereas SP (also known as the true negative rate) measures the proportion of ARMS-NT (negative class) that have been correctly identified (**Eq. 2**). BAC represents the average of individuals that are correctly identified as being ARMS-T or ARMS-NT (**Eq. 3**). PLR translates the probability of an ARMS-T being identified as an ARMS-T divided by the probability of an ARMS-NT being identified as an ARMS-T (**Eq. 4**), whereas NLR translates the probability of an ARMS-T being identified as an ARMS-NT divided by the probability of an AMRS-NT being

identified as an ARMS-NT (**Eq. 5**). Finally, DOR is a measure of how efficient the diagnostic test (i.e. the classifier) is. It is measured by dividing the PLR by the NLR. All the performance measures were measured using the validation subsamples in the outer CV cycle and averaged across the 5 bootstrapped samples as a measure of stability of the performance using different samples (albeit the bootstrapped samples were all drawn from the same population).

		Actual class	
		ARMS-T (positive)	ARMS-NT (negative)
Predicted class	ARMS-T (positive)	True positives	False positives
	ARMS-NT (negative)	False negatives	True negatives

Figure 11. Confusion matrix illustration.

$$SE = \frac{true\ positives}{positives} \quad \textbf{(Eq. 1)}$$

$$SP = \frac{true\ negatives}{negatives} \quad \textbf{(Eq. 2)}$$

$$BAC = \frac{SE+SP}{2} \quad \textbf{(Eq. 3)}$$

$$PLR = \frac{SE}{1-SP} \quad \textbf{(Eq. 4)}$$

$$NLR = \frac{1-SE}{SP} \quad \textbf{(Eq. 5)}$$

The statistical significance for the BAC of each classifier was assessed using permutation testing. Here, the class labels were permutated 1000 times (i.e., training subjects were randomly attributed with ARMS-T or ARMS-NT label) and then, the nested-CV procedure was repeated. The *p*-value for the classification was derived by calculating the number of times in which the BAC for the permuted labels were higher than that obtained for the real labels and dividing this number by 1000 (**Eq. 6**).

$$p \text{ value} = \sum_{i=1}^{1000} \frac{\left(BAC_{permuted_i} > BAC_{observed}\right)}{n} \quad \textbf{(Eq. 6)}$$

BAC was considered statistically significant at a significance level of 5% (i.e., $p < .05$). Furthermore, the prediction ability of each tested combination of feature type, feature manipulation, and CV scheme was evaluated by testing the statistical significance of the median BAC across bootstrapped samples using a one-tailed Wilcoxon signed rank test (i.e., to test if the BAC is higher than chance level – 50%, with a statistical significance level of 5%). P-values were not adjusted for multiple comparisons due to non-independence of the samples used in each statistical test.

3.2.5.2.7. Comparison between testing and validation balanced accuracies

Additionally, the difference between the testing (i.e., from the inner CV cycle) and validation (i.e., from the outer CV cycle) BAC was assessed with a paired two-sided Wilcoxon signed rank test and through Bland-Altman plots for each tested combination of feature type, feature manipulation, and CV strategy and across bootstrapped samples. The difference between the testing and validation BAC was considered statistically significant at a significance level of 5% (i.e., $p < .05$). Moreover, the effect size of the difference was extracted by computing the Cohen's d for paired samples (**Eq. 7**), where $\bar{x}_{validation}$ and $\bar{x}_{testing}$ represent the mean validation and testing BAC, respectively; $s_{validation}$and $s_{testing}$ represent the standard deviation of the validation and testing BAC, respectively; and r the Pearson correlation coefficient between the validation and testing BAC.

$$\text{Cohen's } d = \frac{\bar{x}_{validation} - \bar{x}_{testing}}{\sqrt{s^2_{validation} + s^2_{testing} - 2r s_{validation} s_{testing}}} \quad \textbf{Eq. 7}$$

3.2.5.2.8. Association between the classification error and demographic, clinical and imaging variables

The main effect of several demographic, clinical and imaging variables (age at scan, sex, IQ, GAF and CAARMS at baseline, interval between baseline and scan acquisition age, scan acquisition protocol and IQR – independent variables) on the classification error (dependent

variable) of each tested classifier was tested using a repeated measures design with generalized estimation equations (GEE). In detail, a GEE model was fit for each effect of interest including the independent variable of interest as a between subjects variable and each tested combination of feature type (i.e., ROIGM, ROIWM, ROISurface, VBGM, or VBWM), feature manipulation (i.e., feature dimensionality reduction – principal component analysis or feature selection – no feature selection or forward feature selection), and CV scheme (i.e., LSO, LPO, or 5 fold CV), and bootstrapped samples (i.e., each of the five samples) as within-subjects variables. GEE was chosen given a) that not all ARMS-NT subjects will be represented in all bootstrapped samples (as opposite to the ARMS-T group, which is the same across bootstrapped samples); and b) its superior management of missing data in repeated measures designs, relative to ANOVA (Hubbard et al., 2010). Furthermore, GEE was run with an "Unstructured" covariance matrix between each level of the within-subject factors and a binomial distribution with a logit link was assumed for the classification error. The statistical significance of each effect of interest was tested with a Wald chi squared test and the correspondent *p*-values were corrected for multiple testing (i.e., for 8 statistical tests – every independent variable of interest). The effect was considered statistically significant at a level of significance of 5% (i.e., FDR corrected $p < .05$). Moreover, effect sizes were computed for each effect as the odds ratio (OR), which was computed from the beta (β) parameter estimate from the GEE model (**Eq. 8**) for the independent continuous variable (i.e., age at scan, intelligence quotient, GAF and CAARMS at baseline, interval between baseline and scan ages and weighted average image quality rating) or each level of the independent categorical variables (i.e., sex and scan acquisition protocol).

$$OR = e^{\beta} \qquad \textbf{Eq. 8}$$

3.2.5.2.9. Clinical interpretability of features and applicability of the best classification model

The clinical interpretability of the features selected by the best classification model (i.e., the one showing the best BAC) was assessed in two steps. First, the importance of the selected features was computed as follows: 1) extracted the median of the feature weights (defined by the SVM algorithm) across the best models during the inner CV cycle (i.e., across the models with the best testing BAC during the inner CV cycle) for each outer CV fold; 2)

extracted the standard error of the feature weights across the best models during the inner CV cycle for each outer CV fold; 3) divided this inner CV median by its standard error (inner CV ratio); and 4) summed the inner CV ratio and divide it by the number of outer CV folds. Second, a correlation analysis was conducted between the feature values and the clinical assessment scores (i.e., GAF and CAARMS at baseline) using Pearson's correlation. The correlation was considered statistically significant if its *p*-value was below 5% (after correction for multiple testing, i.e., for every selected feature). Additionally, the clinical applicability of the structural neuroimaging features as a clinical biomarker was assessed for the best classification model using a workflow that have been previously developed (Prata et al., 2014). In detail, the clinical applicability was evaluated with a two-dimensional scale assessing the quality of evidence and effect size of the biomarker. Each of these dimensions were scored from 1 to 4 according to the criteria delineated in th**e Table 16**. The structural neuroimaging features were considered a clinically applicable biomarker, i.e., particularly worthy of clinical consideration, if it reached a sum score of 6 (out of 8).

Table 16. Criteria for quality of evidence and effect size to grade the structural neuroimaging features used in the best classification model as a clinical biomarker. Adapted from (Prata et al., 2014).

Quality of evidence	
Grade	An observation of a positive result (p-value < .05, corrected for multiple comparisons) in:
-	An uncontrolled study.
1	A study controlled for relevant extraneous variables (confounding, nuisance or effect modifiers), i.e., matched, restricted or adjusted for, e.g., age, sex.
2	A study as above (grade 1), but with an explicit a priori intent to discover a precisely defined biomarker, i.e., with a given measure/modality, cut off and direction of effect of both biomarker and response (i.e., transition to psychosis from an ARMS).
3	A study as above (grade 2), but designed with adequate power informed by previous positive studies of the same biomarker, i.e., replication in a larger cohort.
4	At least 2 studies as above (grade 3).
Effect size	
Grade	An observation of a positive result (p-value < .05, corrected for multiple comparisons) with a:
-	Estimate from studies with quality of evidence ≤1.
1	Marginal effect (DOR < 1.3).
2	Small effect size (DOR 1.3–1.5).
3	Medium effect size (DOR 1.5–2.0).
4	Large effect size (DOR > 2.0).

3.3. Results

3.3.1. Image quality control

Image quality measures, i.e., NCR, ICR, RES, and IQR, are in **Table 17**. A main effect of scan acquisition protocol was found on every measure ($p < .001$). The main effects of transition and interaction effect of protocol by transition were not statistically significant on any of the quality measures. Furthermore, 48 (48.5%) images achieved a good weighted quality (23.3% rated with a B and 25.3% with a B-), 49 (49.5%) achieved a satisfactory weighted quality (24.2% rated with a C+, 16.1% with a C, and 6.1% with a C-), and 2 (2%) achieved a sufficient weighted quality (1% rated with a D+ and 1% with a D). Images with good overall image quality ratings (i.e., rated with B/B-) were acquired mainly with scan acquisition protocol 3, whereas images from protocols 1 and 2 were mainly rated as having a satisfactory overall quality (i.e., rated with C+/C/C-) (**Figure 12**). All the subjects' images surpassed the overall weighted quality threshold (i.e., sufficient/D) and, therefore, were included in further analyses.

Table 17. Image quality assessment performed in the structural MRI data of the ARMS sample.

	Protocol 1		Protocol 2		Protocol 3		Group comparison
	ARMS-T (n = 14)	ARMS-NT (n = 19)	ARMS-T (n = 3)	ARMS-NT (n = 16)	ARMS-T (n = 6)	ARMS-NT (n = 41)	
NCR	3.2 ± 0.4 [2.6 4.0]	3.2 ± 0.4 [2.8 4.3]	2.9 ± 0.1 [2.8 3.0]	3.0 ± 0.2 [2.5 3.3]	2.2 ± 0.3 [2.0 2.8]	2.2 ± 0.2 [1.9 2.9]	Protocol: p < .001*** Transition: p = .848 Protocol x Transition: p = .825
ICR	1.6 ± 0.1 [1.4 1.7]	1.6 ± 0.1 [1.4 1.8]	1.7 ± 0.2 [1.6 1.9]	1.7 ± 0.1 [1.5 1.9]	2.7 ± 0.4 [2.2 3.5]	2.4 ± 0.3 [1.9 3.0]	Protocol: p < .001*** Transition: p = .234 Protocol x Transition: p = .113
RES	2.2	2.2	2.2	2.2	2.2	2.2	Protocol: p = < .001*** Transition: p = .146 Protocol x Transition: p = .074
IQR	2.9 ± 0.3 [2.4 3.7]	2.9 ± 0.3 [2.6 3.9]	2.7 ± 0.1 [2.6 2.8]	2.8 ± 0.2 [2.4 3.1]	2.2 ± 0.2 [2.1 2.6]	2.2 ± 0.1 [2.1 2.7]	Protocol: p < .001*** Transition: p = .882 Protocol x Transition: p = .789

Data format: mean ± standard deviation [min max]. **Abbreviations:** ARMS: at-risk mental state; ARMS-T: individuals at an ARMS who later transitioned to psychosis; ARMS-NT: individuals at an ARMS who did not transitioned to psychosis; ICR: inhomogeneity contrast ratio; IQR: weighted average image quality rating; NCR: noise contrast ratio; RES: root-mean-squared resolution. *p<.05; **p<.01; ***p<.001

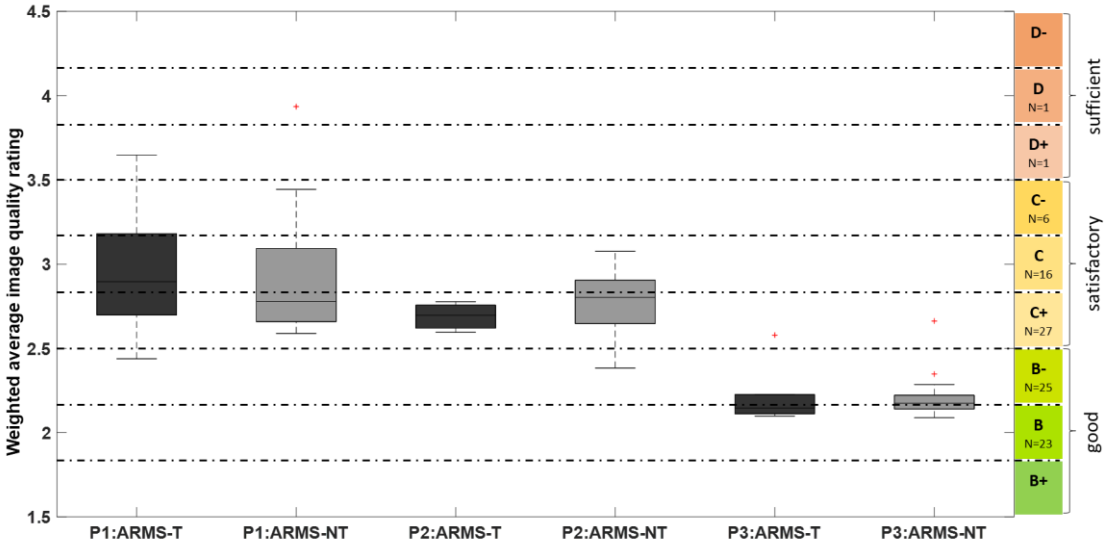


Figure 12. Weighted average image quality rating computed by the quality ascertainment framework of CAT12 using the noise (i.e., noise contrast ratio) and bias (i.e., inhomogeneity contrast ratio) information of each scan protocol [i.e., scan acquisition protocol 1 (P1), 2 (P2), or 3 (P3)] and for each group [i.e., individuals at an at-risk mental state who transitioned to psychosis (ARMS-T) or who did not (ARMS-NT)]. The quality ascertainment framework maps the rating scores to image quality grades (A-F) shown on the right side of the figure.

3.3.2. Structural neuroimaging classification analysis

Overall, the BAC of the classification models trained and validated using each combination of feature type (i.e., ROIGM, ROIWM, ROISurface, VBGM, or VBWM), feature manipulation (i.e., feature dimensionality reduction through PCA; no feature selection; or forward feature selection), CV scheme (i.e., LSO CV; LPO CV; or 5-fold CV), and bootstrapped sample (i.e., one of the five samples) ranged from 37% to 67% (**Figures 13** and **14**; **A1.Tables 1-3** in the **Appendix 1**). The highest classification performance (BAC = 67%) was observed in two classification models, both trained with the ROISurface features using a forward feature selection and a) a LSO CV scheme (SE = 65%, SP = 70%, DOR = 4.3, p = .016; best model 1) or b) a 5 fold CV scheme (SE = 70%, SP = 65%, DOR = 4.3, p = .004; best model 2). Furthermore, only 6 classification models (in a total of 120 trained models) showed a BAC higher than chance level at an individual statistically significant level. However, this statistical significance did not survive the FDR correction for multiple comparisons.

Moreover, the median BAC across bootstrapped samples of each combination of feature type, feature manipulation, and CV scheme did not surpass chance level ($p > .05$; **Tables 18** and **19**; **Figures 13** and **14**).

Table 18. Performance measures of each classification model based on regional features across bootstrapped samples. Measures for each tested combination of regional feature type [i.e., regional-based grey (ROIGM) and white (ROIWM) matter volume; and surface-based regional cortical thickness, gyrification, sulci and complexity indexes (ROISurface)], feature manipulation [i.e., no feature manipulation (No-FS); and forward feature manipulation (FFS)], and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; and 5-fold CV] are presented. Statistical significance of the median BAC across bootstrapped samples was tested using a one-tailed Wilcoxon signed rank test.

	ROIGM		ROIWM		ROISurface	
	No-FS	FFS	No-FS	FFS	No-FS	FFS
LSO CV scheme						
SE (%)	55.7 ± 6.4 [47.8 65.2]	59.1 ± 11.7 [47.8 78.3]	57.4 ± 19.1 [30.4 82.6]	62.6 ± 13.3 [52.2 82.6]	41.7 ± 14.6 [26.1 60.9]	39.1 ± 20.6 [17.4 65.2]
SP (%)	55.7 ± 10.4 [43.5 69.6]	40.9 ± 10.5 [26.1 52.2]	46.1 ± 14 [21.7 56.5]	27.8 ± 22.3 [0.0 56.5]	61.7 ± 17.8 [34.8 82.6]	63.5 ± 12.1 [43.5 73.9]
BAC (%)	55.7 ± 5.5 [47.8 63.0]	50.0 ± 3.8 [43.5 52.2]	51.7 ± 5.6 [43.5 58.7]	45.2 ± 6.6 [37.0 54.3]	51.7 ± 7.4 [43.5 63.0]	51.3 ± 9.8 [43.5 67.4]
PLR	1.3 ± 0.3 [0.9 1.9]	1.0 ± 0.1 [0.8 1.1]	1.1 ± 0.2 [0.7 1.4]	0.9 ± 0.2 [0.7 1.2]	1.2 ± 0.4 [0.8 1.8]	1.1 ± 0.6 [0.6 2.1]
NLR	0.8 ± 0.2 [0.6 1.1]	1.0 ± 0.2 [0.8 1.4]	0.9 ± 0.2 [0.7 1.2]	1.5 ± 0.7 [0.8 2.5]	1.0 ± 0.3 [0.6 1.4]	1.0 ± 0.3 [0.5 1.2]
DOR	1.7 ± 0.8 [0.8 3.0]	1.0 ± 0.3 [0.6 1.3]	1.3 ± 0.5 [0.6 2.0]	0.6 ± 0.5 [0.0 1.4]	1.4 ± 0.9 [0.6 2.9]	1.5 ± 1.6 [0.5 4.3]
p-value	.188	.313	.313	.969	.688	.688
LPO CV scheme						
SE (%)	47.8 ± 6.1 [43.5 56.5]	67.0 ± 7.9 [56.5 73.9]	49.6 ± 8.5 [39.1 60.9]	39.1 ± 9.2 [26.1 47.8]	53.9 ± 6.6 [43.5 60.9]	52.2 ± 6.9 [43.5 60.9]
SP (%)	54.8 ± 6.6 [47.8 60.9]	44.3 ± 10.8 [34.8 60.9]	52.2 ± 5.3 [43.5 56.5]	50.4 ± 12.5 [34.8 69.6]	54.8 ± 5.0 [47.8 60.9]	54.8 ± 12.9 [39.1 69.6]
BAC (%)	51.3 ± 4.8 [45.7 58.7]	55.7 ± 7.1 [45.7 65.2]	50.9 ± 5.2 [43.5 56.5]	44.8 ± 6.6 [37.0 54.3]	54.3 ± 4.3 [50.0 60.9]	53.5 ± 7.5 [43.5 60.9]
PLR	1.1 ± 0.2 [0.8 1.4]	1.3 ± 0.3 [0.9 1.8]	1.0 ± 0.2 [0.8 1.3]	0.8 ± 0.3 [0.5 1.3]	1.2 ± 0.2 [1.0 1.6]	1.2 ± 0.3 [0.8 1.6]
NLR	1.0 ± 0.2 [0.7 1.2]	0.8 ± 0.3 [0.5 1.3]	1.0 ± 0.2 [0.8 1.3]	1.3 ± 0.3 [0.9 1.5]	0.8 ± 0.1 [0.6 1.0]	0.9 ± 0.3 [0.6 1.3]
DOR	1.2 ± 0.5 [0.7 2]	1.9 ± 1.1 [0.7 3.6]	1.1 ± 0.4 [0.6 1.7]	0.7 ± 0.4 [0.3 1.5]	1.5 ± 0.6 [1.0 2.4]	1.5 ± 0.8 [0.6 2.4]
p-value	.500	.188	.500	.969	.063	.500
5-fold CV scheme						
SE (%)	42.6 ± 3.6 [39.1 47.8]	40.9 ± 5.8 [34.8 47.8]	59.1 ± 6.6 [52.2 69.6]	45.2 ± 8.5 [34.8 56.5]	53.0 ± 10.4 [39.1 65.2]	57.4 ± 8.4 [47.8 69.6]
SP (%)	54.8 ± 12.5 [34.8 65.2]	40.9 ± 7.3 [30.4 47.8]	40.9 ± 8.5 [30.4 52.2]	53 ± 11.3 [34.8 65.2]	54.8 ± 6.6 [47.8 60.9]	53 ± 11.7 [39.1 65.2]
BAC (%)	48.7 ± 6.6 [39.1 56.5]	40.9 ± 2.4 [37.0 43.5]	50.0 ± 5.1 [45.7 56.5]	49.1 ± 2.5 [45.7 52.2]	53.9 ± 5.2 [47.8 60.9]	55.2 ± 9.3 [47.8 67.4]
PLR	1.0 ± 0.3 [0.7 1.4]	0.7 ± 0.1 [0.6 0.8]	1.0 ± 0.2 [0.9 1.2]	1.0 ± 0.1 [0.9 1.1]	1.2 ± 0.2 [0.9 1.6]	1.3 ± 0.5 [0.9 2.0]
NLR	1.1 ± 0.3 [0.8 1.6]	1.5 ± 0.2 [1.3 1.9]	1.0 ± 0.3 [0.7 1.3]	1.1 ± 0.1 [0.9 1.3]	0.9 ± 0.2 [0.6 1.1]	0.9 ± 0.3 [0.5 1.1]

DOR	1.0 ± 0.5 [0.4 1.7]	0.5 ± 0.1 [0.3 0.6]	1.1 ± 0.5 [0.7 1.8]	0.9 ± 0.2 [0.7 1.2]	1.5 ± 0.6 [0.8 2.4]	2.0 ± 1.6 [0.8 4.3]
p-value	.688	>.999	.813	.875	.313	.688

Data format: mean ± standard deviation [min max]. **Abbreviations:** BAC: balanced accuracy; DOR: diagnostic odds ratio; NLR: negative likelihood ratio; PLR: positive likelihood ratio; SE: sensitivity; SP: specificity.

Table 19. Performance measures of each classification model based on voxel-wise features across bootstrapped samples. Measures for each tested combination of voxel-wise feature type [i.e., voxel-based grey (VBGM) and white (VBWM) matter volume maps, feature dimensionality reduction through principal component analysis and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; and 5-fold CV] are presented. Statistical significance of the median BAC across bootstrapped samples was tested using a one-tailed Wilcoxon signed rank test.

	LSO CV scheme		LPO CV scheme		5-fold CV scheme	
	VBGM	VBWM	VBGM	VBWM	VBGM	VBWM
SE (%)	20.9 ± 34.8 [0 82.6]	46.1 ± 38.2 [4.3 78.3]	47.0 ± 10.4 [34.8 60.9]	50.4 ± 11.3 [34.8 60.9]	30.4 ± 10.2 [21.7 43.5]	41.7 ± 8.5 [34.8 56.5]
SP (%)	72.2 ± 35.7 [8.7 91.3]	53 ± 35.4 [21.7 95.7]	55.7 ± 8.4 [43.5 65.2]	53.0 ± 7.1 [47.8 65.2]	51.3 ± 7.8 [43.5 60.9]	52.2 ± 6.1 [43.5 60.9]
BAC (%)	46.5 ± 3.6 [43.5 52.2]	49.6 ± 2.4 [45.7 52.2]	51.3 ± 7.5 [45.7 63.0]	51.7 ± 2.8 [47.8 54.3]	40.9 ± 2.4 [37.0 43.5]	47.0 ± 6.8 [41.3 58.7]
PLR	0.6 ± 0.6 [0.0 1.5]	0.9 ± 0.3 [0.3 1.1]	1.1 ± 0.4 [0.8 1.8]	1.1 ± 0.1 [0.9 1.2]	0.6 ± 0.1 [0.5 0.8]	0.9 ± 0.3 [0.7 1.4]
NLR	1.3 ± 0.4 [1.0 2.0]	1.0 ± 0.1 [0.9 1.1]	1.0 ± 0.3 [0.6 1.2]	0.9 ± 0.1 [0.8 1.1]	1.4 ± 0.1 [1.3 1.5]	1.1 ± 0.3 [0.7 1.4]
DOR	0.7 ± 0.9 [0.0 2.2]	0.8 ± 0.4 [0.1 1.1]	1.3 ± 1.0 [0.7 3.1]	1.1 ± 0.2 [0.8 1.4]	0.4 ± 0.1 [0.2 0.6]	0.9 ± 0.7 [0.5 2.1]
p-value	.969	.875	.813	.313	>.999	.969

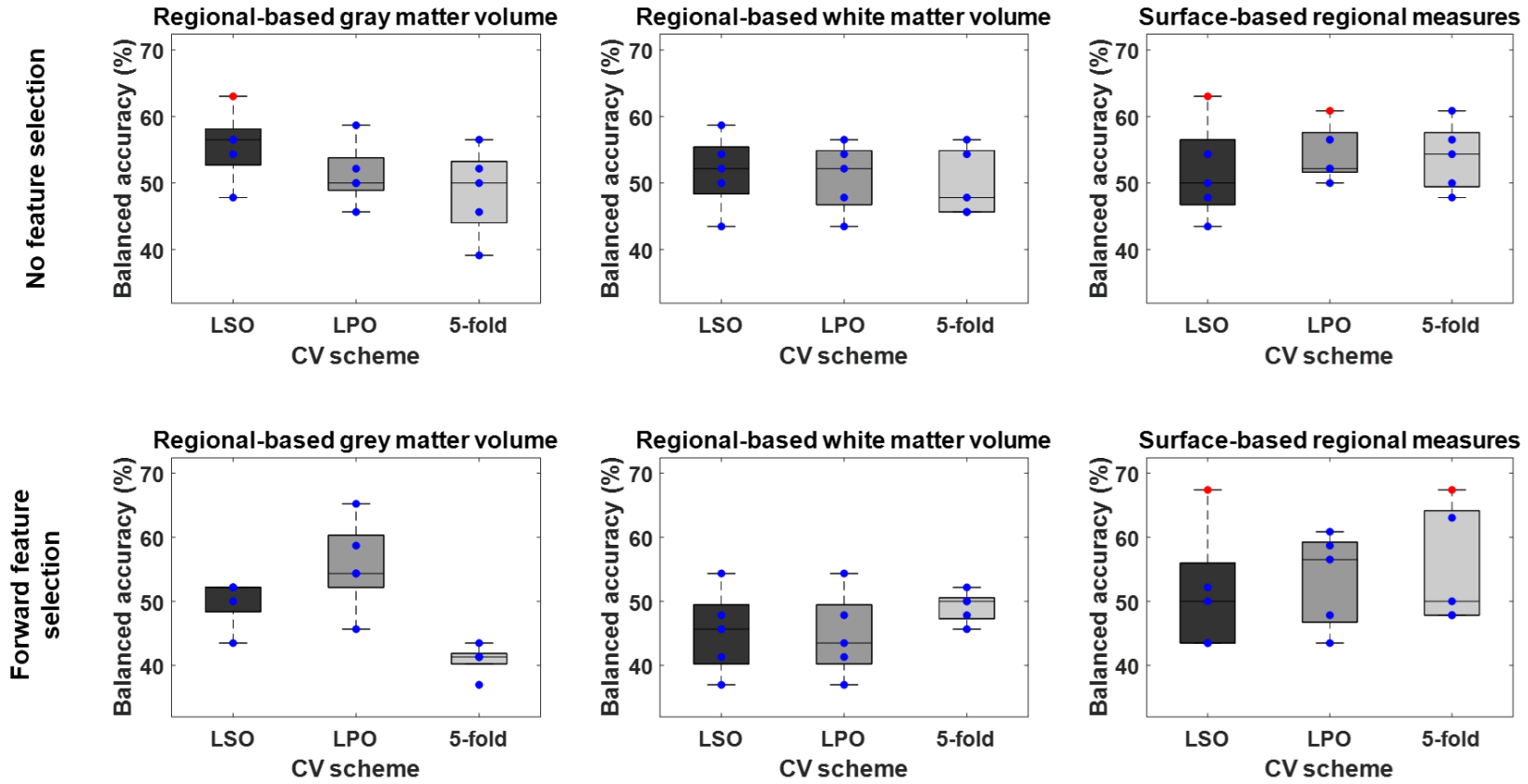


Figure 13. Balanced accuracy across bootstrapped samples for each tested combination of regional feature type [i.e. regional-based grey and white matter volume; and surface-based regional cortical thickness, gyrification, sulci and complexity indexes (surface-based regional measures)], feature selection [i.e., no feature selection; and forward feature selection (FFS)], and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; and 5-fold CV]. Dots represent the balanced accuracy value in each of the five bootstrapped samples and are red colored if the balanced accuracy is statistically significant (i.e., $p < .05$) or blue colored if it is not (i.e., $p > .05$). The statistical significance of the balanced accuracy in each bootstrapped sample was evaluated through permutation testing.

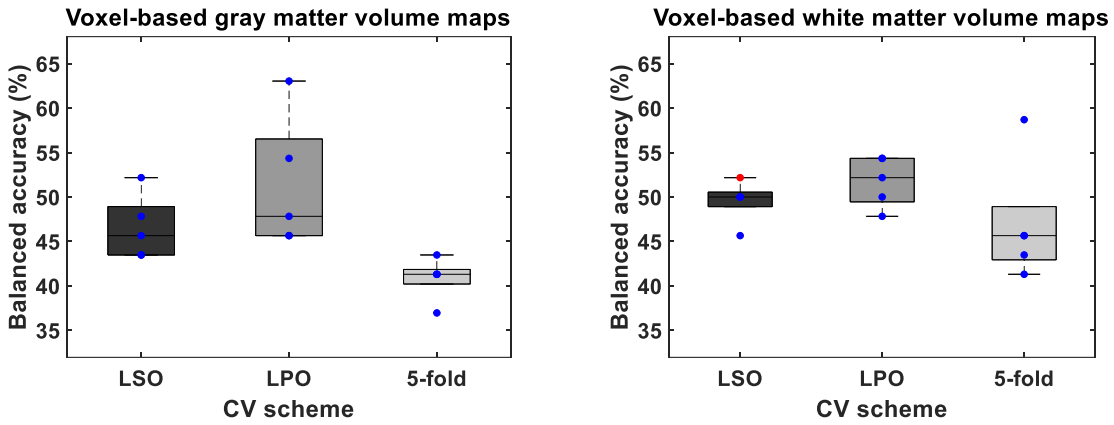


Figure 14. Balanced accuracy across bootstrapped samples for each tested combination of voxel-wise feature type [i.e., voxel-based grey (VBGM) and white (VBWM) matter volume maps], feature dimensionality reduction through principal component analysis and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; and 5-fold CV. Dots represent the balanced accuracy value in each of the five bootstrapped samples and are red colored if the balanced accuracy is statistically significant (i.e. $p < .05$) or blue colored if it is not (i.e. $p > .05$). The statistical significance of the balanced accuracy in each bootstrapped sample was evaluated through permutation testing.

3.3.3. Comparison between testing and validation balanced accuracies

The difference between the testing and the validation BAC was statistically significant for 15 classifiers (out of 24 possible combinations of feature type and manipulation and CV schemes; $p < .05$; **A1.Tables 4** and **5**). In cases where the difference was significant, the validation BAC was on average lower than the testing BAC (**Figures 15-18** and **A1.Figures 1-4**; see also **A1.Tables 4** and **5** for effect sizes). Moreover, very large effect sizes (i.e., $d > 1.2$) were observed for 4 classifiers trained with the following combinations: ROIGM, forward feature selection and a) LSO; b) LPO; or c) 5-fold CV; and d) ROISurface, forward feature selection and LPO CV.

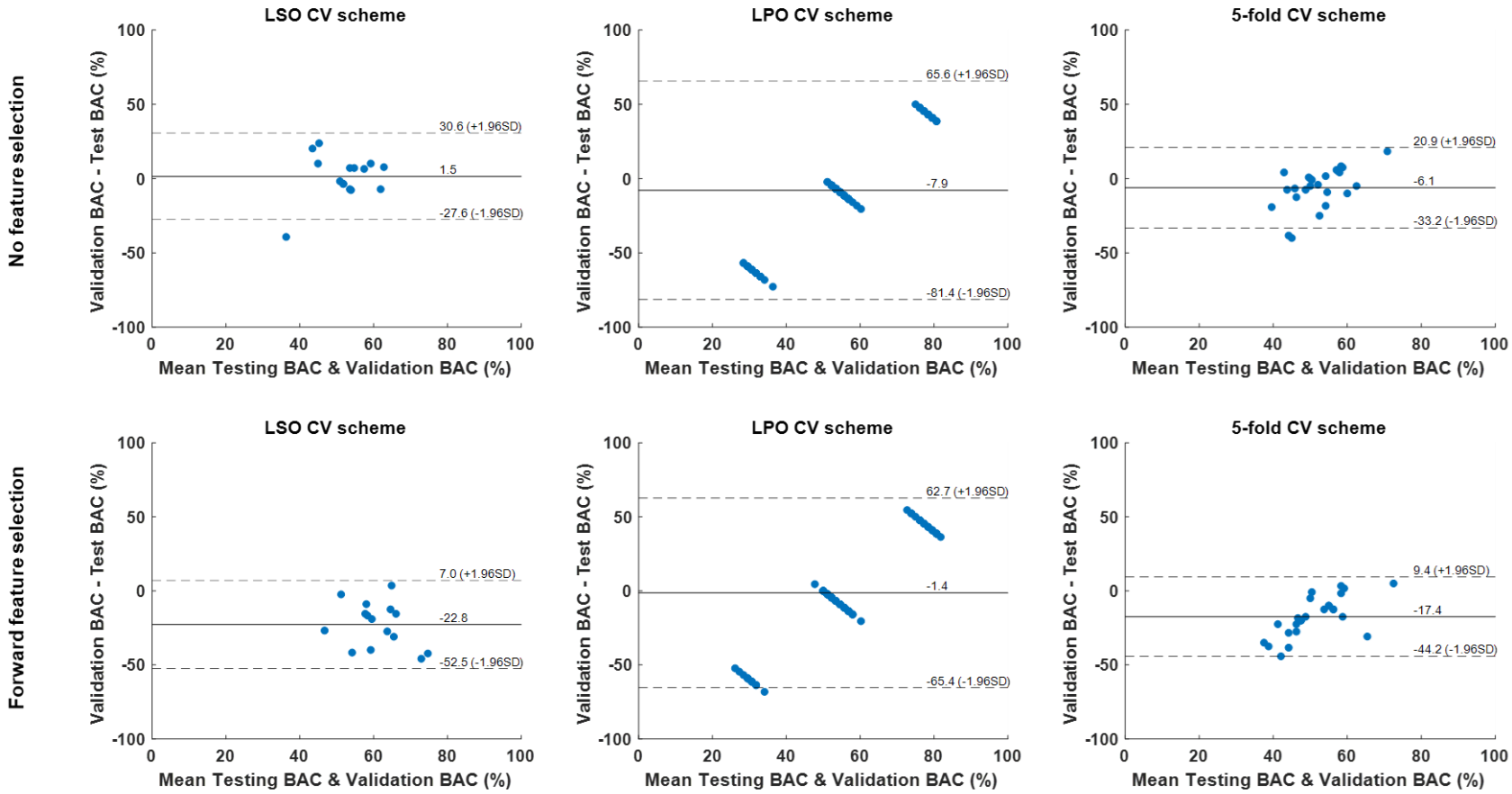


Figure 15. Bland-Altman plot with limits (dashed lines) of agreement for mean (continuous line) testing and validation balanced accuracies (BAC) of classification models trained with regional-based grey matter volumes in combination with feature selection [i.e., no feature selection (top); or forward feature selection (bottom)] and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; and 5-fold CV].

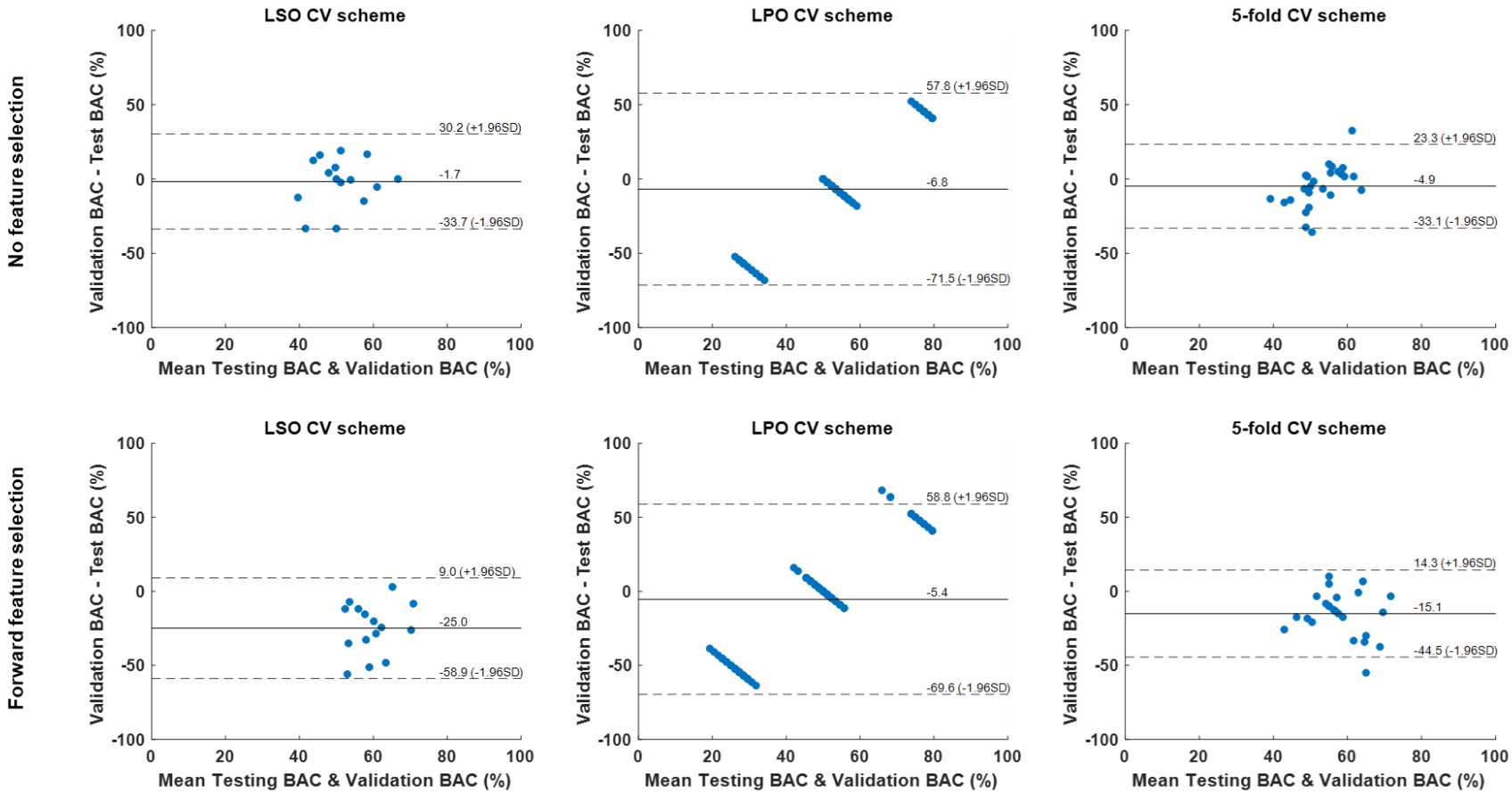


Figure 16. Bland-Altman plot with limits (dashed lines) of agreement for mean (continuous line) testing and validation balanced accuracies (BAC) of classification models trained with regional-based white matter volumes in combination with feature selection [i.e., no feature selection (top); or forward feature selection (bottom)] and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; and 5-fold CV].

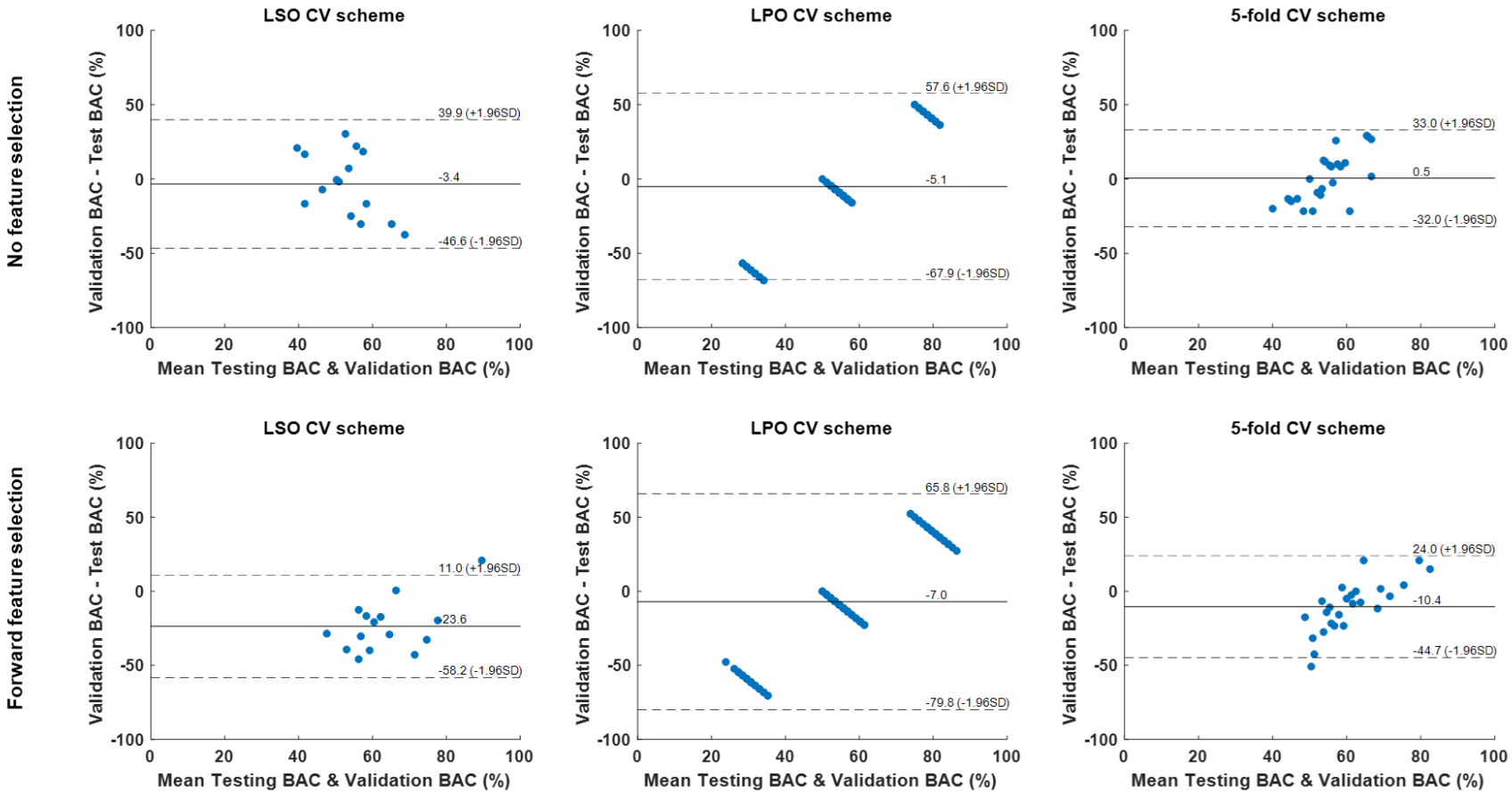


Figure 17. Bland-Altman plot with limits (dashed lines) of agreement for mean (continuous line) testing and validation balanced accuracies (BAC) of classification models trained with surface-based regional cortical thickness, gyrification, sulci and complexity indexes in combination with feature selection [i.e., no feature selection (top); or forward feature selection (bottom)] and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; and 5-fold CV].

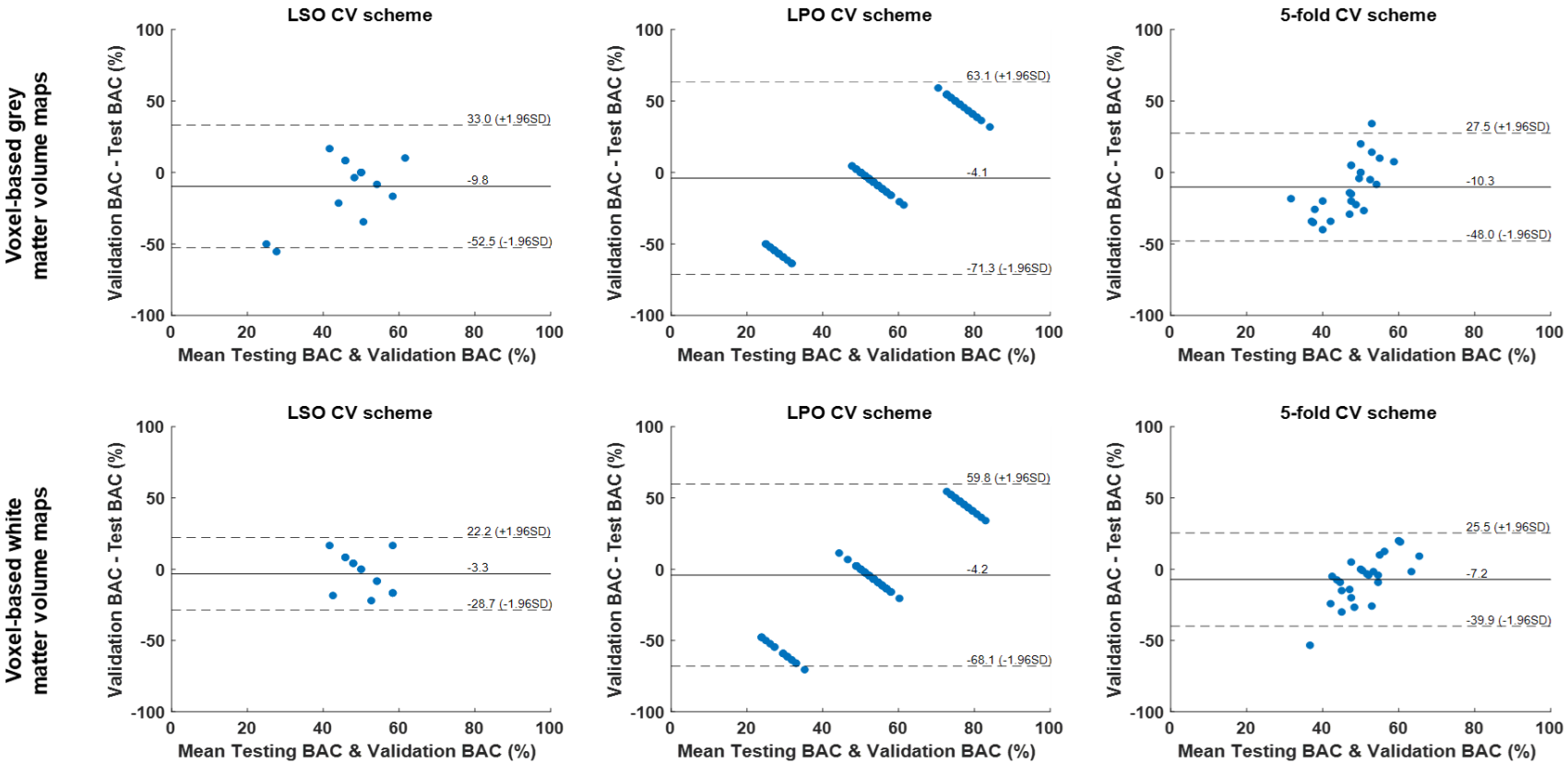


Figure 18. Bland-Altman plot with limits (dashed lines) of agreement for mean (continuous line) testing and validation balanced accuracies (BAC) of classification models trained with voxel-based grey (top) or white (bottom) matter volume maps in combination feature dimensionality reduction through principal component analysis with cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; and 5-fold CV].

3.3.4. Association between the classification error and demographic, clinical and imaging variables

The main effect of scan acquisition protocol (FDR corrected $p = .016$; $OR_{protocol\ 1} = 0.62$ and $OR_{protocol\ 2} = 0.69$, when compared to protocol 3) and IQR (FDR corrected $p = .019$; OR = 0.72) on the classification error were statistically significant (**Table 20**). Subjects are less likely to be wrongly classified if they are a) scanned using protocol 1 or 2 (instead of protocol 3) and b) if their structural MRI's weighted average quality rating increases. The main effect of age at scan, sex, IQ, GAF and CAARMS at baseline and interval between baseline and scan age on the classification error were not significant.

Table 20. Main effect of demographic, clinical and imaging variables on the classification error of each tested classifier. The effect was considered statistically significant at a significance level of 5% (i.e., FDR corrected *p* value < .05). Effects sizes were computed for each effect as the odds ratio (OR) for the continuous variables (i.e., age at scan, intelligence quotient, GAF and CAARMS at baseline, interval between baseline and scan ages and weighted average image quality rating) and each level of the categorical variables (i.e., sex, handedness, self-reported ethnicity, and scan acquisition protocol).

	Wald χ^2(df,n)	***p***	**FDR *p***	**Effect size** OR; 95% CI
Age at scan	$\chi^2(1,5520) = 1.41$	.235	.471	0.99; [1.97 1.01]
Sex	$\chi^2(1,5520) = 0.32$	.569	.608	Male: 1.00 Female: 1.05 [0.88 1.26]
Intelligence quotient	$\chi^2(1,5520) = 0.26$	.608	.608	0.98 [0.90 1.07]
GAF at baseline	$\chi^2(1,5520) = 0.46$	.499	.608	1.00 [0.99 1.01]
CAARMS at baseline	$\chi^2(1,5478) = 2.21$	.138	.367	1.00 [1.00 1.01]
Interval between baseline and scan age	$\chi^2(1,5520) = 0.97$	.326	.521	0.96 [0.89 1.04]
Scan acquisition protocol	$\chi^2(2,5520) = 12.46$	.002**	.016*	Protocol 1 = 0.62 [0.48 0.81] Protocol 2 = 0.69 [0.51 0.93] Protocol 3 = 1.00
Weighted average image quality rating	$\chi^2(1,5520) = 8.02$	.005**	.019*	0.72 [0.58 0.91]

*p<.05

3.3.5. Clinical interpretability of features and applicability of the best classification model

The features with the highest importance (i.e., with magnitude 1.96 standard deviations above or below the mean importance across all features) assigned by the best classification

models were the gyrification index of the left medial orbital frontal cortex and the complexity index of the left banks of the superior temporal sulcus for the best model 1 (see also **A1.Tables 6-9**); and the cortical thickness of the left precentral gyrus and the gyrification indexes of the left lateral orbital frontal cortex, the right precentral gyrus, and the right pars triangularis for the best model 2 (see also **A1.Tables 10-14**).

The correlation analysis between the value of the features selected by the best classification models and the clinical assessment scores (i.e., GAF and CAARMS at baseline) was not statistically significant for any of the features (i.e., FDR corrected *p*-value > .05; **A1.Tables 6-14**).

The two best classification models used as features (and potential clinical biomarker) surface-based regional cortical thickness, and gyrification, sulci depth and complexity indexes. As there is no previous consistent evidence of the positive effect of each of these features on the transition to psychosis from an ARMS, nor of the use of these same features in transition to psychosis from an ARMS classification models, the present study may be considered as exploratory. Furthermore, herein I controlled for relevant extraneous variables (i.e., matching groups – ARMS-T and ARMS-NT – for age, sex, and scan protocol). Therefore, this biomarker scores only 1 (out of 4) in the quality of evidence dimension of the clinical applicability assessment scale (**Table 16**). Moreover, although the two best classification models showed a large effect size (DOR = 4.6 for both models), the biomarker is not assigned a score for the effect size dimension, because it was assigned with a quality of evidence score of only 1 (**Table 16**). Therefore, this biomarker should not be considered as a clinically applicable one.

3.4. Discussion

This study applied ML to structural neuroimaging data using a relatively larger sample and an updated ML approach to detect transition to psychosis from an ARMS aiming to replicate previous positive findings (Das et al., 2018; Koutsouleris et al., 2009, 2012; Koutsouleris, Riecher-Rössler, et al., 2015; Zarogianni et al., 2017, 2019). Furthermore, I explored, for the first time, the use of whole brain white matter volume and regional white matter volume, cortical thickness and surface-based brain gyrification, sulci depth and complexity indexes with ML to predict transition to psychosis.

3.4.1. Prediction of transition to psychosis using structural neuroimaging

Overall, and unexpectedly, I could not replicate previous findings. After balancing the samples for binary classification of transition to psychosis accounting for age, sex and the three different scan acquisition protocols to avoid overoptimistic results, the performance of all tested combinations were not significantly better than chance level. Compared to the previous studies reporting high balanced accuracies in predicting transition to psychosis from sMRI maps (Das et al., 2018; Koutsouleris et al., 2009, 2012; Koutsouleris, Riecher-Rössler, et al., 2015; Zarogianni et al., 2017, 2019), this study has some advantages. First, this study's sample is drawn from a more naturalistic ARMS population as it includes subjects whose sMRI images acquired three different scan acquisition protocols. Indeed, in clinical practice the characteristic of the neuroimaging data of an ARMS clinical help will depend on the imaging center from which the data is acquired. Training a classification model with data from different centers potentially increases the generalizability of that model. Only one of the previous transition to psychosis prediction studies used a two-site group balanced sample (Koutsouleris, Riecher-Rössler, et al., 2015), by combining the samples reported in two previous studies by the same authors (Koutsouleris et al., 2009, 2012). The main differences between this report and mine, herein, are the following. a) Their sample was larger than my balanced bootstrapped samples (i.e., 36% larger than mine, measured as the absolute value of the change in sample size, divided by the average of the size of the two samples). However, I tested my ML models on five balanced subsamples (i.e., through bootstrapping), allowing me to obtain a measure of generalizability of these models' performance. Moreover, they do not present a measure of the statistical significance of the model's balanced accuracy, which I do herein. b) They have controlled the effect of site on the classification using partial correlations during the training phase of the CV cycle, whereas I controlled it by keeping the same proportion of subjects at an ARMS that transitioned to psychosis and those who did not in each scan protocol during the training phase of the CV cycle (i.e., when using the LPO CV scheme as the previous study did). Additionally, I also guaranteed that the pair of subjects left out for testing/validation were from the same site. This potentially increases the generalizability of the classification model by training it with a more heterogeneous sample (and, as explained above, more naturalistic) and diminishing the effect of site on the testing/validation classification accuracy (which is not taken into account in the previous report (Koutsouleris, Riecher-Rössler, et al., 2015).

Second, I trained my classification models with samples balanced for group (subjects at an ARMS who later transitioned to psychosis and who did not), age at scan and sex. Balancing for group is important to avoid biasing the classification model to the most represented group and it was not taken into account by three previous reports (Koutsouleris et al., 2009, 2012; Zarogianni et al., 2019). Moreover, the effects of age (Fjell & Walhovd, 2010) and sex (Ruigrok et al., 2014) on brain structure have been consistently reported and, therefore, should be taken into account in the context of ML with sMRI data. All previous report used a sample matched for age and sex (as me herein) (Koutsouleris et al., 2009, 2012; Koutsouleris, Riecher-Rössler, et al., 2015; Zarogianni et al., 2017, 2019), except for one (Das et al., 2018). Das and colleagues reported a statistically significant and better than chance level balanced accuracy in predicting transition to psychosis using a sample unbalanced for both group and sex. Although they used a ML algorithm with class (i.e., group) weighing – which in summary increases the influence of the minority class when training the classifier by assigning higher weights to rare cases, the authors performed an unspecified correction for sex effect (as well as for age and TIV effects) to the data during the training CV cycle. This approach may not be the most appropriate given the known effect of sex on brain structure and the empirically tested association between sex and group (i.e., transition to psychosis from an ARMS *vs.* no transition), which categorizes sex as a potential confounder in this analysis. Furthermore, in three of the previous reports, the effects of age and sex were corrected before entering the ML analysis (Koutsouleris et al., 2009), and during the training CV cycle (Das et al., 2018; Koutsouleris et al., 2012) using partial correlations (Koutsouleris et al., 2009, 2012) or an unspecified method (Das et al., 2018) – which I did not perform. Correction for age effects in ML analysis has been previously shown to increase classification accuracy in Alzheimer's disease, when it is estimated from healthy subjects (Falahati et al., 2016). Correction for effects of no interest in ML analysis should be done with extreme caution as it can easily remove relevant subject-specific information (Wachinger et al., 2021). This is especially important when the correction is being performed in a non-healthy (i.e., non-standard) population, because the effect of external variables such as age and sex might be modulated by the presence of the disease (e.g., being at ARMS or having schizophrenia).

Third, this study's sample is composed by subjects whose clinical diagnosis of an ARMS was based on having a schizotypal personality disorder or on the subject's familial-high risk coupled with functioning decline and on the CAARMS (Yung et al., 2005), which mainly

evaluates positive symptoms. These were not the same criteria as those used in the previous studies predicting transition to psychosis from an ARMS. Indeed, all used samples of subjects clinically assessed with tools that evaluate not only positive symptoms, but also basic and negative symptoms (Das et al., 2018; Koutsouleris et al., 2009, 2012; Koutsouleris, Riecher-Rössler, et al., 2015; Zarogianni et al., 2019), except one (Zarogianni et al., 2017), which included only familial-high risk subjects in its sample. This potentially increases the inclusion of subjects in the early phase of the psychosis prodrome (characterized by the presence of basic and negative symptoms), whereas my sample includes mainly subjects in the late prodromal phase of psychosis (characterized mainly by the presence of positive symptoms) (Paolo Fusar-Poli, Borgwardt, et al., 2013). Therefore, my results suggest that previously reported accuracies in predicting transition to psychosis may be population-specific and poorly generalizable to differently and clinically characterized populations (as mine herein). Unfortunately, I could not directly test this by using my sample as an external validation sample to the previously reported classification models due to logistic constrains.

3.4.2. Comparison between testing and validation balanced accuracies

I further investigated the possible reasons for the poor performance of the tested classification models (i.e., with balanced accuracies that were not significantly better than chance level). I found a significant decrease in the validation balanced accuracies, compared to the testing ones in 63% of the tested classification models (i.e., in 15 out of 24). This difference was more pronounced (i.e., with a large effect size; $d > 1.2$) when testing regional gray matter volume, cortical thickness and surface-based brain gyrification, sulci depth and complexity indexes features in combination with the forward feature selection method and one of the possible CV schemes (i.e., 27% of the models which difference between the testing and validation balanced accuracies was significant – 4 out of 15). Specifically, in these cases, at least 89% of validation balanced accuracies are lower than the mean of the testing balanced accuracies. This is suggestive of some degree of overfitting during the training/testing phase of the classification models and might explain the poor generalization when they are tested on unseen data (i.e., during the validation phase). Moreover, although the observed testing performance of 79% of the tested models (i.e., 19 out of 24) is significantly higher than chance level (data not shown), it rarely reaches balanced accuracies higher than 80% (data is shown only graphically). Thus, the low performance of these models may be explained by other factors.

3.4.3. Association between the classification error and demographic, clinical and imaging variables

I also investigated the association between the model's classification error and several extraneous variables, namely demographic, clinical and imaging variables. Surprisingly, I found a significant association between scan acquisition protocol and sMRI's weighted average quality rating and the classification error, albeit they were balanced between the groups (i.e., ARMS-T and ARMS-NT). Indeed, these results show that subjects are more likely to be wrongly classified if they are scanned with the acquisition protocol from the 3T scanner and if their MRI images have a lower weighted average quality rating, i.e., better quality. As a matter of fact, the images acquired with the 3T protocol are the ones showing the highest overall quality from the whole sample, therefore, it is not surprising that both variables' (i.e., the protocol and average quality) effects are significant. Although I applied a subsampling strategy that keeps the groups balanced within each protocol, maximizing in this way the sample size, this seems to be insufficient to dilute the effect of protocol on the classification accuracy.

3.4.4. Clinical interpretability of features and applicability of the best classification model

As a proof of concept, I additionally analyzed the two models that individually yield the best significant performance (i.e., the highest balanced accuracy with a $p < .05$) in terms of its clinical interpretability and potential applicability. These models were trained with surface-based regional cortical thickness and brain folding measures (gyrification, sulci depth and complexity indexes). Among the features contributing the most for the prediction of the transition to psychosis are a) the gyrification indexes of the left lateral and medial orbital frontal cortex and the right precentral gyrus, the complexity index of the left banks of the superior temporal sulcus – contributing with a negative weight; and b) the cortical thickness of the left precentral gyrus and the right pars triangularis – contributing with a positive weight (note that negative final score implies a classification of a subject as ARMS-T). Indeed, 67% of these features (i.e., 4 out of 6) can be interpreted as to have a risk effect in psychosis, i.e., given the brain of a subject at an ARMS, the more folded these brain regions are, the more likely this subject is to be classified as ARMS-T. On the contrary, the thicker

the left precentral gyrus and the more folded the pars triangularis, the more likely this subject is to be classified as ARMS-NT. Interestingly, and in line with what I found, the transition to psychosis has been shown to be associated to a higher gyrification index in the prefrontal cortex, which includes the orbital frontal cortex (Harris et al., 2007). Moreover, lower gyrification index in the pars triangularis (a region belonging to the Broca's area, which is very important in language processing) in patients with chronic schizophrenia compared to healthy controls has been previously reported (Cachia et al., 2008). Although no significant differences in cortical thickness between ARMS who later transition to psychosis and those who do not have been consistently reported (Sprooten et al., 2013; S Tognin et al., 2014), overall cortical thinning has been observed in FEP (Sprooten et al., 2013) and schizophrenia (Van Haren et al., 2011) patients, when compared to healthy controls. Furthermore, the correlation between the regional cortical thickness and the brain folding measures selected by the best classification models and the clinical assessments at baseline (i.e., GAF and CAARMS) was not significant in any of the measures. This suggestively supports the use of structural neuroimaging as a complement to the clinical assessments as they seem to convey non-overlapping information. However, before these regional cortical thickness and brain folding measures can be considered a potential clinically applicable biomarker, at least a replication of this study must be conducted. Only then its potential clinical usefulness can, indeed, be discussed. To the best of my knowledge, only one study used graph-extracted network measures from cortical gyrification to predict transition to psychosis using randomized trees (Das et al., 2018), which is not the same method I used herein.

3.4.5. Limitations

This study has a few limitations that need to be addressed. First, this study is not an exact replication to the previous findings, e.g., I used updated versions of the imaging processing tools (i.e., SPM12/CAT12) and employed a strategy to balance group (i.e., ARMS-T and ARMS-NT) for important extraneous, such as age, sex and scan acquisition protocol. Second, although my bootstrapped samples' size, i.e., the ones used to train the classification models, were at the same order of magnitude as previous studies using balanced samples, it is small given it gathers data from three different imaging acquisition protocols. Indeed, the strategy I followed to dilute the effect of protocol on the classification accuracy does not seem to be effective, possibly due to lack of power, i.e., small sample and/or uneven representation of each protocol in the sample. Moreover, to rule out the hypothesis that my

negative results are originated by the lack of power, i.e., due to the small sample size, the same methods as described herein should be replicated in significantly larger samples, which will be provided by multicenter projects, such as NAPLS 2 (Addington et al., 2012), PSYSCAN (Stefania Tognin et al., 2020), and PRONIA (http://pronia.eu) over the next years. Third, to estimate the generalizability of all the classification models I performed bootstrapping, i.e., draw 5 subsamples from the original sample using a semi-random approach, and then tested the significance of the resulting balanced accuracy across the models trained with each subsample individually. An alternative approach would be to increase the number of bootstrapped samples drawn from the original sample (maximizing, in this way, the use of the sample that is available) and combining the individual classifiers output, i.e., trained with each bootstrapped sample, into an ensemble decision, e.g., by majority voting, as done in recent schizophrenia structural neuroimaging ML studies (Das et al., 2018; Koutsouleris, Meisenzahl, et al., 2015). However, I could not apply this approach to our study due to logistic and computational resources constrains – the ML software NeuroMiner could not perform bootstrapping with customize conditions as I did herein, i.e., balancing groups in each scan acquisition protocol.

In this study, I attempted to replicate previous findings showing the usefulness of structural MRI in predicting transition to psychosis from an ARMS using ML, i.e., with ML classifiers showing accuracies better than chance level, with a completely independent sample. Overall, after balancing the samples for binary classification of transition to psychosis accounting for age, sex and the three different scan acquisition protocols to avoid overoptimistic results, the performance of all tested combinations were poor. These results suggest that structural MRI data, alone, cannot predict transition to psychosis at better than chance-level, suggesting that the prediction value of structural MRI data from a prodromal stage of psychosis should be reconsidered.

4. Creating a brain-based gene expression quantitative trait loci score (eGenScore) tool

4.1. Introduction

The emergence of the genome-wide studies (GWAS) has allowed the identification of thousands of variants (mainly single nucleotide polymorphisms [SNPs]) that influence the susceptibility of complex diseases, such as schizophrenia (Trifu et al., 2020). However, there is still a gap between the variants and their functional role of in the diseases' etiology, in particular in regards to SNPs (Lappalainen et al., 2013). Indeed, nearly 90% of these genetic variations occur in non-coding DNA sequences, and only about 4-5% of plausibly causal variants in GWAS-associated regions are coding variants, which suggests that the main mechanism by which variation in these regions acts is not by altering protein structure. In comparison, about 50% of plausibly causal variants are expression quantitative trait loci (eQTL), suggesting moderation of gene expression is an important mechanism of action (Hindorff et al., 2009; Watanabe et al., 2019). As such, it is crucial to consider and efficiently utilize variants correlated with gene expression, i.e. eQTL, to better understand the mechanisms behind the role of specific genes (especially if implicated by the hypothesis-free GWA approach) in intermediate or complex phenotypes (Aguet et al., 2017).

The degree of expression of genes is typically inferred from the transcriptome, i.e. the messenger RNA (mRNA) levels of all genes expresses in the tissue. Measuring the expression of a given gene is invasive for many tissues, including the human brain, requiring postmortem sampling. Recently, efforts have been put forward to compile large-scale concomitant transcriptomic and genomic datasets, i.e. eQTL datasets, such as the Brain eQTL Almanac (BrainEAC) (Ramasamy et al., 2014) and the CommonMind Consortium (CMC) (Hoffman et al., 2019), with both being brain specific. These datasets provide a source for data on eQTLs in numerous tissue samples and for the analysis of tissue-specific gene expression (Aguet et al., 2017, 2020; Ardlie et al., 2015; Barbeira et al., 2021; Ramasamy et al., 2014). Using these emerging eQTL datasets, gene expression can be used as an intermediate molecular phenotype to potentially address the functional gap in GWAS findings and get a much needed step closer to understanding the underlying mechanisms and molecular pathways of complex disorders.

Herein I present a brain-based gene expression quantitative trait loci score tool, eGenScore, to predict gene expression levels based on genome-wide genotypes. This tool uses a polygenic method, i.e. a weighted sum of SNP's alleles, to compose an eQTL score, i.e. with

SNPs that have been shown to have a statistical association with gene expression levels. The eQTL score acts as a proxy for gene expression level in a specific tissue. We trained eQTL models, i.e. the formula to compute the eQTL scores, with concomitant transcriptomic (i.e. gene expression levels measured through mRNA data) and genomic (i.e. SNP's genotypes) data specifically from the frontal cortex (i.e. brain tissue) from the BrainEAC database. Furthermore, I analyzed the performance of the eQTL models using an internal cross-validation approach and an external validation approach by applying the models to the CMC database.

4.2. Materials and methods

An overview of the datasets and methods used in this study is represented in **Figure 19**. All quality control procedures, described below, were performed by the database providers.

4.2.1. Genomic and transcriptomic datasets

4.2.1.1. BrainEAC

The BrainEAC dataset was used to train and internally validate eQTL models using eGenScore. The dataset belongs to the UK Brain Expression Consortium (UKBEC) (Ramasamy et al., 2014), was downloaded from the first version of the BrainEAC website (http://www.braineac.org/), and is composed by genome-wide genotypes and gene expression levels from postmortem samples from 10 brain tissues, i.e. the frontal, the temporal, the occipital, and the cerebellar cortices, the hippocampus, the putamen, the substantia nigra, the thalamus, and the intralobular white matter, of 134 individuals. In this study we used only gene expression data from the frontal cortex and only 127 individuals had data from this brain tissue. A full description of the acquisition and processing of transcriptomic and genomic data can be found in the original study (Ramasamy et al., 2014).

The individuals included in the database were restricted to the ones having European ancestry for two main reasons. First, populations from geographically different origins (e.g. Europeans, Asians, Africans) have been shown to have different allele frequencies throughout the genome (Hellwege et al., 2017). Therefore, any analysis based on alleles' frequencies (e.g. genome-wide association analysis, polygenic risk score) should always take into account the genetic structure of the population from which the sample was drawn from, either by restricting it to a specific one, which is the most common approach, or by

controlling for it as a covariate of no interest, which is less common and more likely to control its confounding effect. Second, most of the statistical-based genetics studies use populations with an European ancestry due to its availability being in a much higher proportion than the other populations. Regarding the genomic data, samples were composed by genome-wide genotypes of SNPs and the final number of SNPs included in this database passed the following quality control. First, samples and SNPs had a call rate higher than 95% to exclude individuals or SNPs with high rates of genotypes missingness. Second, SNPs deviating from the Hardy–Weinberg equilibrium (HWE) with a $p < 10^{-4}$ were excluded. HWE states that in an indefinitely large population, the frequencies of alleles and genotypes in the genome are constants over generations. In statistical-based genetic studies, when the expected and the observed genotype frequencies differ statistically, it is assumed it is a result of genotyping errors, and, therefore, these genotypes are excluded from the sample. Third, as the number of SNPs that are actually genotyped is limited by the number of genetic markers of the chip that is used, these SNPs genotypes were used to statistically infer the genotypes of the vicinity SNPs throughout the genome in an approach called imputation (Li et al., 2009). This method is based on the genetic structure of a reference, or standardized, population that should correspond to the one from which the sample being studied was drawn from, i.e. blocks of SNPs that are inherited with a high non-random correlation, or in high linkage disequilibrium. In the BrainEAC database, the genotyped sample was imputed using the European reference panel of the 1000 Genome Project (see also below) (Auton et al., 2015). Finally, the imputed SNPs were further filtered out if their post-imputation quality was below .50 or if the frequency of their minor allele (MAF), i.e. the one with the lowest frequency, was below 5%, resulting in approximately 5.8 million SNPs. The main reason for this last criterion is that most of the studies do not have enough power to detect associations with SNPs with low MAF.

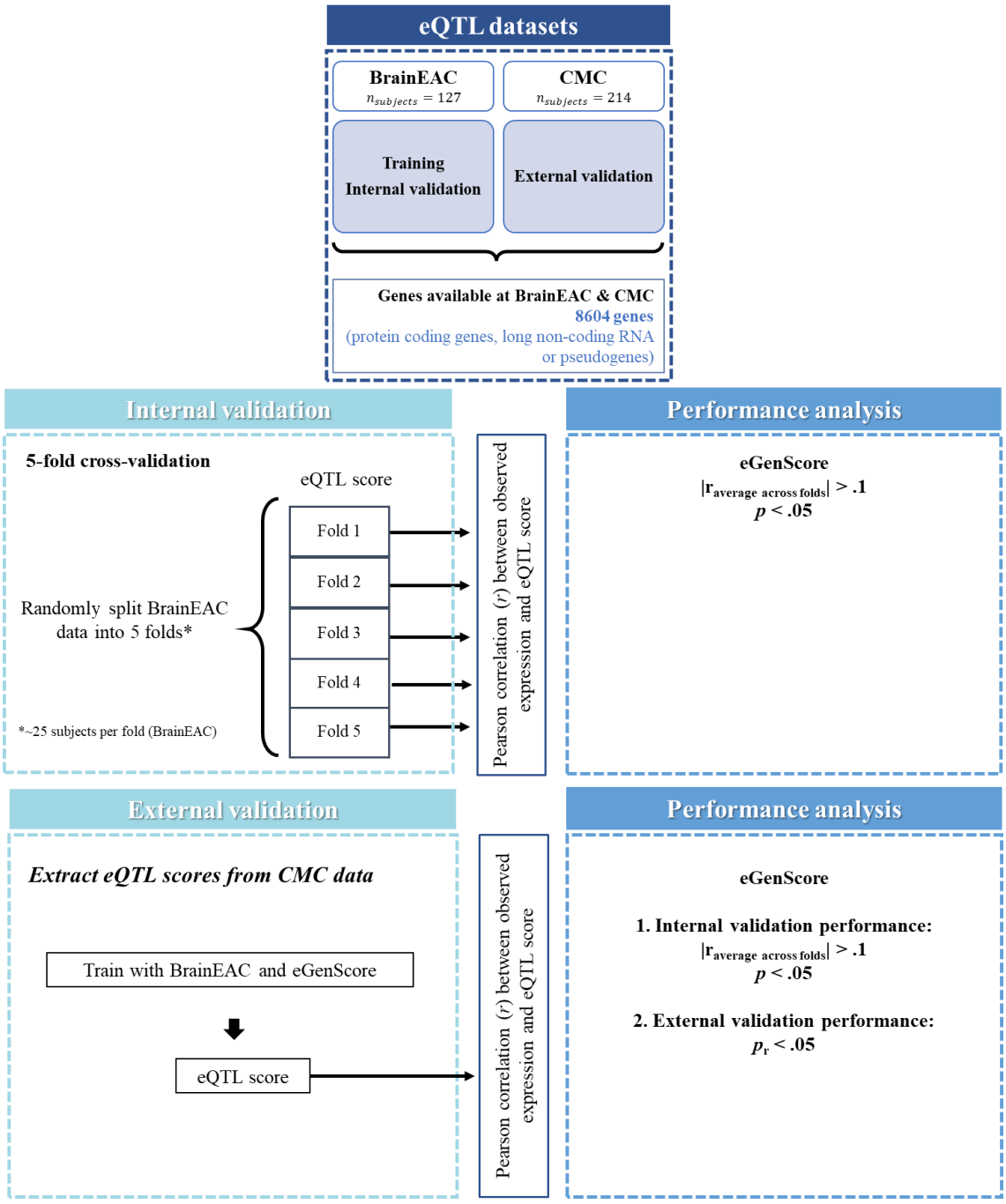


Figure 19. Representation of the steps taken for the selection of genes (A) for which an eQTL model was trained and validated, both internally (B) and externally (C).

Regarding the transcriptomic data, gene expression levels for around 25 thousand genes were available. In specific, for each gene, the exon-specific RNA expression data is average using winsorized mean, normalized using robust multi-array average, log2 transformed across all tissues and corrected for batch effects, sex and brain bank. Furthermore, genomic and transcriptomic data were mapped onto the human genome assembly GRCh37/hg19 and

annotated according to NCBI Reference Sequence build 36, and only SNPs and exon-specific transcripts from chromosomes 1 to 22 were included in this study (i.e. sex chromosomes were excluded).

4.2.1.2. CMC

The CMC dataset (Release 1) was used to externally validate the eQTL models trained with eGenScore and BrainEAC dataset. The dataset belongs to the CommonMind Consortium (Hoffman et al., 2019) and comprises genome-wide genotypes and gene expression levels in the frontal cortex (i.e. dorsolateral prefrontal cortex) of 214 individuals. All samples and approximately 40 million SNPs passed quality control (exclusion of individuals with neuropsychiatric diseases – bipolar disorder, schizophrenia or affective disorder – and with non-European ancestry, samples with call rate < 90%, p-value for deviation from HWE < $5x10^{-5}$, genotyping call rate < 98%). From these 214 individuals, gene expression levels for 15 478 genes in counts per million (normalized by scaling each sample's read count by the total counts by gene, $\log_2$ transformed and corrected for covariates using surrogate variables analysis) were provided by the CMC dataset. Furthermore, genomic and transcriptomic data were mapped onto the human genome assembly GRCh37/hg19, annotated to Gencode (v26, https://www.gencodegenes.org/human/release_26.html), and only SNPs and exon-specific transcripts from chromosomes 1 to 22 were included in this study (i.e. sex chromosomes were excluded).

4.2.1.3. 1000 genomes

The 1000 genomes dataset (phase 3, October 2015, EUR panel) were used to compute the linkage disequilibrium and to adjust the weight of each SNP in the gene expression models trained with eGenScore using the BrainEAC dataset. This data is part of the 1000 Genomes project (Auton et al., 2015) and was chosen as representative of the worldwide population (i.e. of a standardized population) in terms of genetic variance. The dataset is composed by genome-wide genotypes of individuals with European ancestry only and comprises approximately 78 million SNPs mapped onto the human genome assembly GRCh37/hg19 of 503 individuals. Furthermore, only SNPs from chromosomes 1 to 22 were included in this study (i.e. sex chromosomes were excluded).

4.2.2. Genes overlap between datasets

In this study I analyzed only genes a) that were labeled as protein coding, long non-coding RNA or pseudogenes in Gencode (v26), and b) if expression levels were available simultaneously in BrainEAC and CMC datasets. The correspondence between gene transcripts IDs from BrainEAC and the CMC gene ensemble IDs was done using BioMart tool (Durinck et al., 2005) using the following criteria: a) the transcript ID and the gene ensemble ID should be from the same strand (i.e. positive or negative); and b) when more than one transcript ID in the BrainEAC database corresponded to the same gene ensemble ID in CMC database, the transcript ID with the largest overlap (in nucleotide base pairs) with the gene ensemble ID is chosen. eQTL models were trained and validated (i.e. internally and externally) for 8604 genes which had expression levels simultaneously in BrainEAC and CMC databases.

4.2.3. eQTL model training

An eQTL model for each gene was trained using eGenScore framework and BrainEAC datasets. The first step was to select SNPs located 1 million base pairs upstream and downstream of the gene location in the genome. Second, the association between the SNPs genotypes and the gene expression level was tested using linear regression and an additive allele coding (i.e. 0, 1 or 2 tested alleles) for each SNP individually (**Eq. 9**) as implemented in the Matrix eQTL engine (Shabalin, 2012).

$$expression_{gene_j} = \beta_i \times SNP_i \qquad \textbf{(Eq. 9)}$$

SNPs nominally associated (i.e. $p < .05$) with gene expression were filtered using linkage disequilibrium (LD) from the 1000 genomes dataset. In detail, the SNPs were first ordered by statistical significance (i.e. from the lowest to the largest *p*-value). Second, for every possible unique pair of SNPs the LD was measured using the 1000 genomes dataset. Third, for each pair of SNPs in high LD (i.e. $r^2 > .3$), the SNP with the lowest significance (i.e. the highest *p*-value) was excluded. Fourth, the third step was iterated across all pairs of SNPs in high LD. Fifth, the weight of the contribution to the gene expression level (i.e. the beta

coefficients from the linear regression described above) of each SNP to be included in the gene expression model was extracted.

The eQTL score, which represents the predicted gene expression, was computed as the weighted sum of each SNP's tested alleles adjusted to the expected proportion of those alleles in a standardized population (i.e. the 1000 genomes dataset) (**Eq. 10** and **11**). For each SNP (SNP_i), this adjustment centers the expected contribution of the SNP_i at zero. In this way, replacing a missing genotype for a given SNP with zero sets the contribution of this SNP to its expected contribution in the population (Vassos et al., 2020).

$$eQTL\ score_j = \sum_i(\beta_i \times SNP_i - ajustment\ factor_i) \quad \textbf{(Eq. 10)}$$

$$adjustment\ factor_i = \beta_i \times Proportion_{1\ ref\ allele;\ i} + 2 \times \beta_i \times Proportion_{2\ ref\ allele;i} \quad \textbf{(Eq. 11)}$$

4.2.4. Internal validation

The eQTL models were internally validated using a 5-fold cross-validation approach. In each iteration, the following measures were computed using the hold-out fold: a) the Pearson correlation coefficient (*r*) between the observed gene expression and the eQTL score, and b) the *p*-value corresponding to the null hypothesis of no correlation between the observed gene expression and the eQTL score. Then, as an overall performance measure of the eQTL score model, the Pearson correlation coefficient was averaged across the 5 folds (r_{avg}) and squared (r_{avg}^2). Furthermore, the r_{avg}^2 is herein interpreted has the variance in the observed gene expression levels that can be explained by the eQTL score (i.e. the predicted gene expression levels). The global *p*-value was computed using the Fisher's method (Fisher, 1992). Furthermore, the models were considered significant if the averaged correlation between the observed gene expression and the eQTL score was statistically significant (i.e. Fisher's *p*-value < .05) and of at least small size (i.e. $|r_{avg}| > .1$).

4.2.5. External validation

The eQTL models that were shown to be significant at the internal validation were externally validated using the CMC dataset. The external validation performance was assessed by computing the squared Pearson correlation coefficient (r^2) between the observed gene expression and the eQTL score in the CMC dataset and considered statistically significant if the *p*-value corresponding to the null hypothesis of no correlation between the observed gene expression and the eQTL score was below .05.

4.2.6. Statistical analysis

The squared average Pearson correlation coefficient (i.e. r_{avg}^2) and the squared Pearson correlation coefficient (i.e. r^2), of the internal and external validations, respectively, were compared for the genes whose models were significant using a two-sided paired *t*-test. Cohen's *d* was computed as the effect size of the difference and its interpretation was performed using Kristoffer Magnusson web tool (Interpreting Cohen's d effect size, https://rpsychologist.com/d3/cohend/).

4.3. Results

4.3.1. Internal validation performance of the eQTL score models

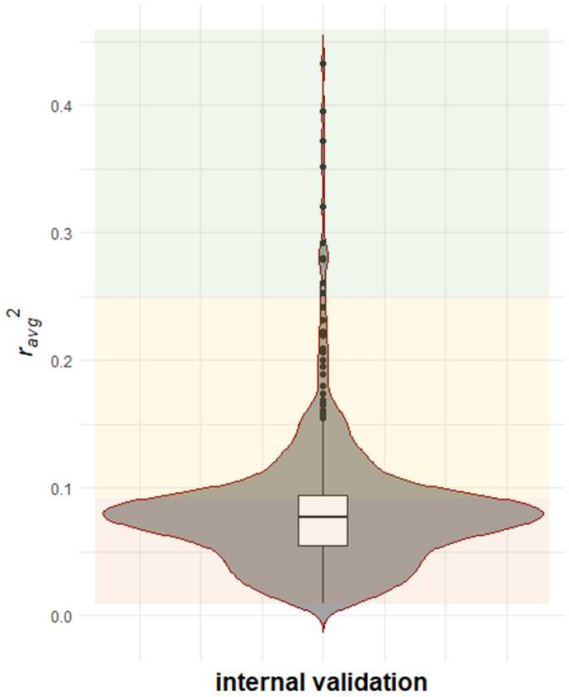


Figure 20. Performance of the eQTL models during the internal cross-validation, measured as the squared average correlation between the observed gene expression and the eQTL score across the hold-out folds (r_{avg}^2).

The proportion of eQTL models shown to be significant (i.e. $|r_{avg}| > .1$ and Fisher's *p*-value < .05) was 6.9% (594 out of 8604 genes). Moreover, 1.9% of the significant models showed a high performance (i.e. $r_{avg}^2 > .25$; 11 out of 594 genes), 29.1% showed a medium performance (i.e. $.09 < r_{avg}^2 < .25$; 173 out of 594 genes), and 69.0% showed a low performance (i.e. $.01 < r_{avg}^2 > .09$; 410 out of 594 genes) (**Figure 20**).

4.3.2. External validation performance of the eQTL model

The proportion of eQTL models shown to have a significant correlation between the observed gene expression and the eQTL score in the external database CMC (i.e. $p < .05$) was 23.2% (138 out of 594 genes with significant expression models at the internal validation). Moreover, 17.4% of these models showed a high performance (i.e. $r^2 > .25$; 24 out of 138 genes), 26.8% showed a medium performance (i.e. $.09 < r^2 < .25$; 37 out of 138 genes), and 55.8% showed a low performance (i.e. $.01 < r^2 > .09$; 77 out of 138 genes). Furthermore, the performance of the models during the internal and the external validation (i.e. r_{avg}^2 *vs.* r^2) were not significantly different ($t(137) = -1.06$, $p = .292$, Cohen's $d = 0.09$) (**Figure 21**).

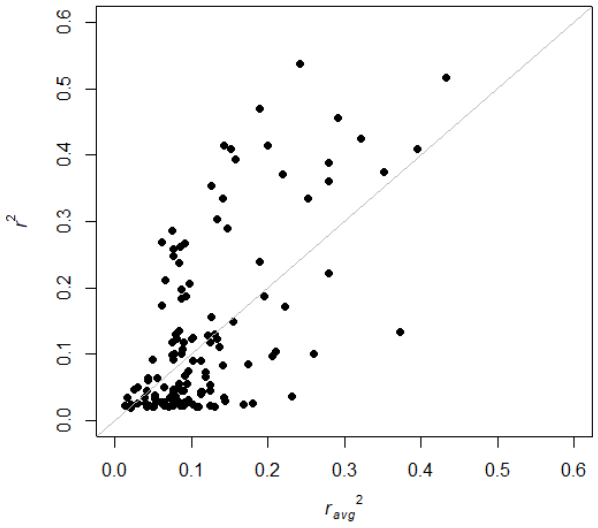


Figure 21. Performance of the eQTL models during the internal (r_{avg}^2) and the external validation (r^2).

4.4. Discussion

I presented eGenScore, a brain-based gene expression quantitative trait loci score tool that uses a polygenic method to compute eQTL scores, a proxy for gene expression levels. I analyzed the performance of the eQTL models using an internal cross-validation approach and an external validation approach by applying the models to the CMC database. Overall, these results showed that eGenScore was able to predict gene expression with a significant internal cross-validation performance (i.e. with an absolute Pearson correlation between the observed gene expression and the eQTL score > 0.1 and a Fisher's *p*-value < 0.05) for almost 7% of the total genes for which a model was trained. One reason for the lack of significance might be the low statistical power to test the correlation between the observed gene expression and the eQTL score, i.e. for each hold out fold in the internal cross-validation there was only around 25 observations to test the correlation. Moreover, the external validation results are in line with this argument, i.e. around 23% of the significant (i.e. at the internal validation) eQTL models where able to predict gene expression in an external dataset with 214 observations (i.e. with a statistically significant correlation between the observed and predicted gene expression).

Furthermore, only a small proportion (1.9%) of significant eQTL models (i.e. at the internal validation) were able to explain more than 25% of the variance of the observed expression. Indeed, the eGenScore is only able to predict the genetically regulated component of the expression of a given gene. However, the gene expression is also regulation by epigenetic mechanisms, such as DNA methylation and imprinting (Hoopes, 2008). Therefore, it is reasonable to posit that the expression of the genes for which the eQTL model performed poorly might had been mainly regulated by other epigenetic mechanisms. Moreover, this might also explain why only a low proportion (i.e. 17.4%) of gene expression models were able to explain more than 25% of the variance in the observed gene expression of the external dataset (i.e. CMC). Additionally, the brain samples of the datasets used to train and internally validate the eQTL models (i.e. BrainEAC) and to externally validate them (i.e. CMC) from which the gene expression levels were measured were not extracted exactly from the same brain region. It is reasonable to expect that the expression level of a given gene might slightly vary depending on which location in the brain the same is taken from. Therefore, the performance of the eQTL models on the external dataset might be influenced by this factor.

4.4.1. Limitations

This study has a few limitations that need to be addressed. First, as the gene expression data of the BrainEAC and CMC databases were annotated to different human genome libraries (i.e. in BrainEAC to NCBI Reference Sequence build 36 and in CMC to Gencode v26) which hinders the exact correspondence between the transcriptomic data of the two databases and, therefore, narrows the number of possible genes that we can analyze. Second, the performance of the eQTL models measured trough the internal cross-validation was indeed limited by the small sample size of the test subsamples (i.e. the hold out folds), which restricts the statistical power in detecting significant correlations between the observed gene expression and the eQTL score. Third, the eQTL models trained by eGenScore using BrainEAC as the training database are only valid on transcriptomic and genomic samples with an European ancestry. The validity of these models in samples with other ancestries (e.g. African) remains unknown and would be interesting to explore in future studies. Fourth, another gene-based method, PrediXcan, has been recently developed in parallel to the eGenScore and have been used to predict gene expression levels from SNPs (Gamazon et al., 2015; Huckins et al., 2019). PrediXcan uses a regularized linear regression method to estimate the gene expression levels, whereas eGenScore uses a polygenic approach. Furthermore, the most recent models trained by PrediXcan used as the training dataset the Genotype-Tissue Expression (GTEx version 8 – 2019) (Lonsdale et al., 2013), which, compared to BrainEAC, is a more comprehensive transcriptomic and genomic database with a slightly larger sample size to what regards brain gene expression, and, very importantly, up-to-date data reads (i.e. the availability of whole-genome sequencing data and gene expression data annotated to the most recent human genome assemblies – Gencode 26). The comparison of the gene expression models' performance across methods (i.e. eGenScore *vs.* PrediXcan) as well as across training datasets (BrainEAC *vs.* GTEx) would be important explore and has been explored after the formulation of this thesis (Vânia Tavares et al., 2021).

In this study, I presented eGenScore, a polygenic method to predict gene expression levels in the brain from genotypes. This method opens an opportunity to study neuropsychiatric diseases using a mechanistic or functional approach by only requiring availability of genotyping information, which is easier to get and, in particular for brain samples, non-invasive. Future studies are still needed to prove its applicability to other populations, e.g.

with African ancestry, and a comprehensive comparison with parallelly developed similar approaches to estimate gene expression from genotyped data is also required and has been explored after the formulation of this thesis.

5. Individual prediction of transition to psychosis using genetics and machine learning

5.1. Introduction

Genetics have been successfully used in psychosis classification studies (i.e. schizophrenia/first episode psychosis (FEP) patients/At-risk mental state (ARMS) individuals *vs.* healthy controls or FEP patients *vs.* ARMS individuals; see **1.4.2. Individual prediction of psychosis using genetics and machine learning** in **1. General introduction**. However, to the best of my knowledge, genetics have never been explored as a predictor of the transition to psychosis from an ARMS. Moreover, several types of genetic markers have been revealed as promising predictors of psychosis, namely single nucleotide polymorphisms (SNPs), a composite score of SNPs and gene expression profile (for a review see **1.4.1 Potential genetic biomarkers** in **1. General introduction**).

Herein, I explored, for the first time, the potential of genetics to predict transition to psychosis from an ARMS using a sample of 75 individuals and two model approaches: a previous and a novel model-based approach. In the previous model-based approach, I used a schizophrenia polygenic risk score (PRS) that have been previously shown to distinguish FEP patients from healthy controls (Vassos et al., 2016) as a transition to psychosis predictor. In the novel model-based approach, I tested two different predictors: (1) a set of SNPs that have been shown to be associated with schizophrenia in the most recent genome-wide meta-analysis study (Pardiñas et al., 2018) and (2) a brain-specific expression Quantitative Trait Loci (eQTL) score of a set of genes that have been identified as schizophrenia-associated in the same genome-wide meta-analysis study (Pardiñas et al., 2018). For this last set of predictors, I used the eGenScore tool (see **4. Creating a brain-based gene expression quantitative trait loci score (eGenScore) tool**) to extract the eQTL score for each gene and several brain tissues (i.e. frontal, temporal, and occipital cortices; putamen; substantia nigra; hippocampus; cerebellum; and white matter). Moreover, the eQTL score is herein used as a proxy for the gene expression profile in the several brain tissues.

5.2. Materials and methods

5.2.1. Sample description

The sample used in this study was described in detail in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning section**.

From the pool of 246 individuals at an ARMS, 135 had genome-wide genotyped data. Over the 2-year follow-up period, 41 individuals at an ARMS developed psychosis (ARMS-T) and the remaining 94 did not (ARMS-NT).

5.2.2. Genotyping and imputation

Genotyping procedures have been previously described (Bramon et al., 2014; Vassos et al., 2017). In detail, DNA was extracted from blood or cheek swabs. The samples were genotyped either at the South London and Maudsley (SLaM) NHS Foundation Trust/King's College London Biomedical Research Centre Genomics Laboratory on the Illumina HumanCore Exome BeadChip ("SLaM sample", 935 subjects from which 134 were ARMS – 40 ARMS-T and 94 ARMS-NT) or at the Wellcome Trust Sanger Institute (WTSI, Cambridge, UK) on the Genome-wide Human SNP Array 6.0 ("WTSI sample", 786 subjects from which 1 was ARMS-T). Quality control (QC) included exclusion of SNPs with minor allele frequency (MAF) <1% or 2%, SNPs with genotypic failure >1% or >5%, and individuals with genotypic failure >1% or 2%, and SNPs with Hardy Weinberg equilibrium $p<10^{-5}$ in healthy subjects only or $p<10^{-6}$ in SLaM or WTSI sample, respectively. Imputation was performed with IMPUTE2 (Howie et al., 2009) based on the 1000 Genomes phase 3 reference panel (Auton et al., 2015). The imputed markers underwent a second stage of QC to exclude SNPs that were missing in >5% or 1% of individuals in SLaM or WTSI sample, respectively, or had imputation information score (INFO) <0.8.

5.2.3. Genotyped samples merging and population stratification analysis

SLaM and WTSI samples were merged keeping only overlapped imputed SNPs. To account for genotyping and imputation QC differences in the two samples and following standard GWAs QC guidelines (Marees et al., 2018), an extra QC was run excluding SNPs missing in >2% of individuals and with a MAF <5%. After the QC, a population stratification analysis was conducted to select only subjects with an European ancestry (the number of subjects per self-reported ethnicity is in the **A2.Table 1** in the **Appendix 2**). In detail, a principal component analysis was applied to a) the merged total sample (i.e. 935 subjects from the SLaM sample and 786 subjects from the WTSI sample); and b) the 1000 Genomes phase 3 dataset (i.e. 2504 subjects). Then, the first two principal components from each dataset were extracted and plotted against each other (**Figure 22**). The 1000 Genomes

dataset is used in this analysis as a reference for the genetic structure across populations. It is comprised by five super populations: African, American, East Asian, European, and South Asian ancestry. As the genetic ancestry of the subjects in the reference dataset is known, the ancestry of the subjects from a non-reference dataset can be inferred by visually comparing the two first principal components plotted against each other with those from the reference dataset. For this study only subjects with an European ancestry were selected for further analysis according to the following criteria: a) subjects had to have reported as being 'white' (i.e. self-reported ethnicity is a proxy of the genetic ancestry); and b) subjects had to show a genetic structure explained by the first two components roughly in a similar manner as in the reference dataset, i.e. the scatter plot had to be visually similar. Seventy-five subjects met these criteria (21 ARMS-T and 54 ARMS-NT) and two subjects were excluded as they have reported as being white, but showed a genetic structure similar to other populations, i.e. Asian and African (**Figure 22**). Samples merging, QC and population stratification analysis were conducted using PLINK 1.9 (https://www.cog-genomics.org/plink2) (Chang et al., 2015).

Socio-demographic (age at baseline and follow-up or transition, sex, years of education, and intelligence quotient) and clinical variables (global assessment of functioning and comprehensive assessment at risk mental state scores at baseline and follow-up or transition) were analyzed using a two-tailed independent t-test for continuous data and a chi square test for ordinal data (**Table 21**). These statistical analyses were performed using the Statistical Package for the Social Sciences 26 (SPSS 26 for Windows, Chicago, IL, USA).

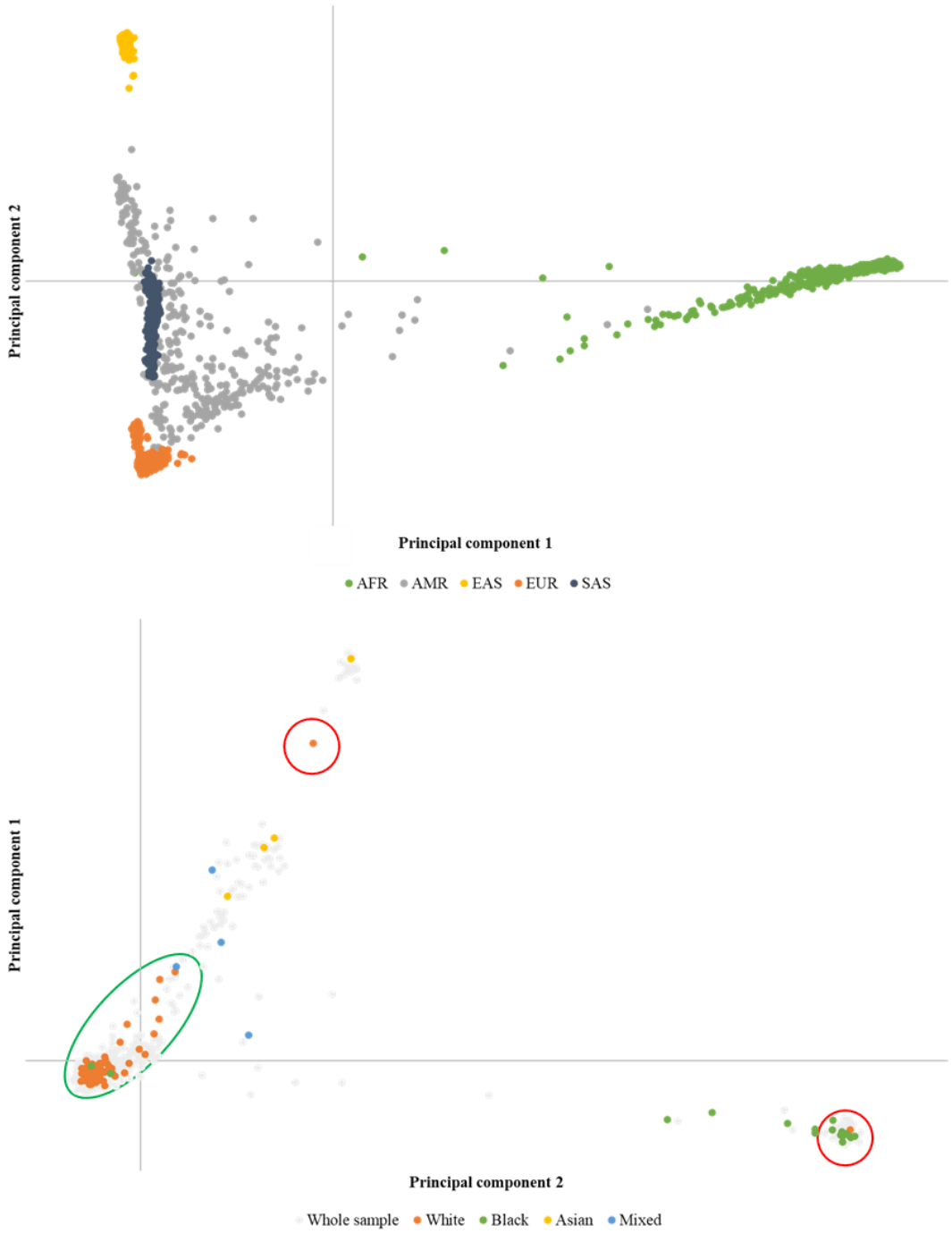


Figure 22. Population stratification analysis from the reference dataset (i.e. 1000 Genomes; top) and the study sample (i.e. ARMS sample, bottom). ARMS subjects selected for further analysis met the following two criteria: a) self-reported as being 'white' (orange dots in the bottom plot); and b) show a genetic structure similar to the that of the reference dataset's subjects with an European ancestry (orange dots in the top plot). Therefore, the ARMS subjects included (i.e. 75 subjects) in the final sample are highlighted by the green ancestry (i.e. met the two criteria) and the ones excluded (i.e. 2 subjects) are highlighted by the red circles (i.e. they have reported as being white, but showed a genetic structure similar to other populations – Asian and African. AFR: African, AS: Asian, AMR: American, EAS: East Asian, EUR: European, SAS: South Asian ancestries.

Table 21. Socio-demographic and clinical information of the ARMS sample with genetic data and an European ancestry.

	ARMS-T (n = 21)	**ARMS-NT** (n = 54)	**Group comparison**
Age at baseline (years)	23.8 ± 5.3 [15.6 33.8]	22.5 ± 4.0 [14.6 34.5]	p = .284
Age at follow-up or transition (years)	25.3 ± 5.9 [17.3 38.8]	27.9 ± 5.1 [18.5 43.9]	p = .069
Sex (male/female)	14/7	30/24	p = .380
Years of education	13.0 ± 2.2 [10.0 18.0]	12.0 ± 4.4 [0 18.0]	p = .292
IQ (z-standardized) [a]	0.1 ± 1.0 [-2.1 2.2]	0.2 ± 1.0 [-2.1 1.8]	p = .678
GAF at baseline	54.0 ± 15.7 [0 80]	53.6 ± 16.0 [0 78]	p = .923
GAF at follow-up [b]	47.8 ± 24.3 [0 79]	59.2 ± 21.0 [0 94]	p = .050
CAARMS at baseline [c]	37.6 ± 17.5 [6 69]	29.9 ± 16.2 [0 81]	p = .097
CAARMS at follow-up [d]	24.4 ± 27.9 [0 90]	12.4 ± 14.0 [0 60]	p = .019*

Data format: mean ± standard deviation [min max]. Information not available for [a]2 ARMS-T and 9 ARMS-NT; [b]4 ARMS-NT; [c]1 ARMS and 9 ARMS-NT; [d]1 ARMS-T and 3 ARMS-NT. ARMS: at-risk mental state; ARMS-T: individuals at ARMS that later transitioned to psychosis; ARMS-NT: individuals at ARMS that did not transitioned to psychosis. *p<.05.

5.2.4. Machine learning approach

5.2.4.1.Sample balancing and bootstrapping

The final sample used in the genetic machine learning (ML) analysis was composed by 19 (for the previous model-based approach) or 21 (for the novel model-based approach) ARMS-T subjects and 19 or 21 ARMS-NT subjects, respectively, randomly selected to match the ARMS-T for age, and sex. The matching criteria for age and sex was based on the non-rejection of the null hypothesis that the ARMS-T and ARMS-NT groups had the same median age (tested with a two-sided Mann-Whitney U-test) and sex (tested with a chi-square test). The subsampling was repeated 100 times, i.e., 100 bootstrapped samples were created, and the subsequent ML analysis was conducted for each of the bootstrapped sample.

5.2.4.2. Previous model-based approach

Vassos and colleagues have previously shown that PRS for schizophrenia could discriminate FEP patients from healthy controls with a medium effect size (R^2 = 9.4%, p < .001) (Vassos

et al., 2017) using a sample of 445 FEP patients and 265 controls. Herein, I tested PRS for schizophrenia as a predictor of the transition to psychosis from ARMS. This analysis was conducted using R software 4.0.5 (R Core Team, 2018).

5.2.4.2.1. Polygenic risk score extraction

The PRS was computed for each subject ($PRS_i, i = 1, \dots 70$) in this study's sample (i.e. ARMS sample) as the sum of the alleles of the SNPs that have been previously associated with the diagnosis of schizophrenia (versus healthy controls; $SNP_j, j = 1, \dots, m$), weighted by their effect size, $\ln\left(OR_{SNP_j}, j = 1, \dots, m\right)$, in that association (**Eq. 12**). The association between each SNP and the schizophrenia diagnosis was conducted by the Psychiatric Genomics Consortium GWAS meta-analysis study (Ripke et al., 2014) in a sample of 36 989 patients with schizophrenia and 113 075 healthy controls. The number of SNPs to be include in the PRS was determined by Vassos and colleagues the ones below an association with schizophrenia *p*-value below 0.1. The reason this *p*-value threshold was chosen is that it explained the highest proportion of variance between the PRS and the diagnosis of FEP (*vs.* healthy controls) (Vassos et al., 2017). The actual computation of the PRS for the ARMS sample was done by Dr. Evangelos Vassos at the Institute of Psychiatry, Psychology and Neuroscience, King's College London, United Kingdom, and was only available for 70 ARMS subjects (19 ARMS-T and 51 ARMS-NT).

$$PRS_i = \sum_j^m \ln\left(OR_{SNP_j}\right) SNP_j \qquad \textbf{(Eq. 12)}$$

5.2.4.2.2. Logistic regression for classification

Binary classification of transition to psychosis from an ARMS (i.e. ARMS-T *versus* ARMS-NT) was performed using logistic regression. Logistic regression is a regression for binary outcomes (as herein). In the case where we have only one predictor, X, as I have herein, the probability of observing the outcome is Y is given by **Eq. 13**.

$$P(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} + \varepsilon \qquad \textbf{(Eq. 13)}$$

For binary classification, a threshold of 0.5 is commonly applied to the probability of observing the outcome. For example, a subject at an ARMS with a $P(Transition\ to\ psychosis) > 0.5$ would be predicted to transition to psychosis, while a subject at an ARMS with a $P(Transition\ to\ psychosis) < 0.5$ would be predicted to not develop a psychotic disorder, i.e. would not transition to psychosis.

5.2.4.2.3. Cross-validation

The logistic regression was trained and tested in a simple leave-one per group-out (LPO) CV scheme (**Figure 10-inner CV cycle**; see also the section **1.2. Machine Learning** in **1. General introduction** and the section **3.2.5.2.5. Cross-validation** in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**).

5.2.4.3. Novel model-based approach

Two novel prediction models for transition to psychosis from genetic data using this study's ARMS sample were trained and tested. One using as predictors SNPs' alleles and other using eQTL scores of genes that have been previously associated with psychosis (i.e. patients with schizophrenia versus healthy controls). The selection of both SNPs and genes was based on a recent GWAS meta-analysis (Pardiñas et al., 2018) which used a sample of 40 675 patients with schizophrenia and 64 643 healthy controls. The analyses were conducted using R software 4.0.5 (R Core Team, 2018).

5.2.4.3.1. Feature types

5.2.4.3.1.1. Psychosis-associated SNPs

The SNP sets to be included in the ML analyses were defined as follows. First, SNPs were ranked by their psychosis association significance (i.e. *p*-value) according to the (Pardiñas et al., 2018) study. Then, four subsets were defined by selecting all the SNPs meeting a statistical significance level threshold (several thresholds were tested: $p < 10^{-2}$, $p < 10^{-3}$, $p < 10^{-4}$, or $p < 10^{-5}$). The number of SNPs from the original study and the ones available in our sample are in the **Table 22**.

Table 22. Number of psychosis associated SNPs with a statistical significance level below a given threshold (i.e. $p < 10^{-2}$, $p < 10^{-3}$, $p < 10^{-4}$, or $p < 10^{-5}$) in the original study (Pardiñas et al., 2018). The number of SNPs from each subset (i.e. defined by each threshold) that are available in the ARMS sample is also represented.

#SNPs	$p < 10^{-2}$	$p < 10^{-3}$	$p < 10^{-4}$	$p < 10^{-5}$
(Pardiñas et al., 2018)	307 632	98 130	41 442	21 498
ARMS sample	79 247	31 565	14 791	8 102

5.2.4.3.1.2. eQTL scores of psychosis-associated genes

The eQTL score of a list of psychosis-associated genes identified by the same study from which the list of SNPs was extracted ((Pardiñas et al., 2018); see also **Psychosis-associated SNPs** section) was computed. In specific, from a total of 571 genes that have been identified in the previous GWAS meta-analysis (Pardiñas et al., 2018), 426 genes were available in the BrainEAC database (see also the section **4.2.1.1. BrainEAC** in **4. Creating a brain-based gene expression quantitative trait loci score (eGenScore) tool**). Then, an eQTL score for each gene and each brain tissue available in BrainEAC (i.e. frontal, temporal, and occipital cortices; putamen; substantia nigra; hippocampus; cerebellum; and white matter) was computed using the eGenScore tool (see also **4. Creating a brain-based gene expression quantitative trait loci score (eGenScore) tool**). Furthermore, only eQTL scores for which there was a model that have showed (1) a statistically significant internal validation performance (i.e. $p < .05$) and (2) an absolute Pearson correlation coefficient between observed gene expression and eQTL score of 0.1 at the internal validation were considered for the ML analysis (for details regarding eQTL score model significance see **4. Creating a brain-based gene expression quantitative trait loci score (eGenScore) tool**). The number of genes per brain tissue for which an eQTL score was computed is in **Table 23**.

Table 23. Number of genes per brain tissue for which an eQTL score was computed.

Brain tissue	**#genes**
Frontal cortex	21

Temporal cortex	17
Occipital cortex	16
Putamen	16
Substantia nigra	13
Hippocampus	11
Cerebellum	28
White matter	19

5.2.4.3.2. Elastic net for classification

Binary classification of transition to psychosis from an ARMS (i.e. ARMS-T *versus* ARMS-NT) was performed using logistic regularized regression with elastic net (Zou & Hastie, 2005). This regularization method applies L1 ($\|\beta\|_1$; from least absolute shrinkage and selection operator or LASSO regression) and L2 ($\|\beta\|^2$; from the ridge regression) penalties when estimating the weights (β) of each predictor (X) in the regression (**Eq. 14**).

$$\hat{\beta} = \arg min_{\beta} \|y - X\beta\|^2 + \lambda(l_1\|\beta\|_1 + l_2\|\beta\|^2) \quad \textbf{(Eq. 14)}$$

L1 penalty can shrink coefficients to zero, which can help to reduce overfitting and perform feature selection. However, if a group of predictors are collinear, L1 penalty tends to select only one of the predictors, while ignoring the others (i.e. shrinking their coefficients to zero). Moreover, it can select at most the same number of predictors as the number of observations used to fit the regression model. These limitations are overcome by adding the L2 penalty, which also shrinks the coefficients, but in the case of collinearity it equalizes the coefficients of that group of predictors (i.e. instead of selecting only one). This helps to reduce the model complexity and multi-collinearity. The final elastic net penalty to be applied to the regression coefficients is defined by l_1 and l_2, where $l_2 = 1 - l_1$ and λ. If $l_1 = 1$, then only L1 penalty would be applied, whereas if $l_1 = 0$, then only L2 penalty would be applied. Moreover, λ is a numerical value that defined the amount of shrinkage to be applied to the regression coefficients. Herein, a search was carried out to identify the optimal l_1 and λ values (i.e., $l_1 = 0{:}\,0.1{:}\,1;\ \lambda = 0.01{:}\,0.01{:}\,1$) in the inner CV cycle (see the **Cross-validation** section). Furthermore, the implementation of the elastic net was done using the 'glmnet' v4.0 R package.

5.2.4.3.3. Cross-validation

The two classifiers were trained using the nested-CV scheme described in the section **3.2.5.2.5. Cross-validation** in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**. In summary, a 5-fold CV was applied to the inner cycle to find the optimal number of initial SNPs to be included in the model (i.e. defined by the statistical significance threshold $p = 10^{-2}$, $p = 10^{-3}$, $p = 10^{-4}$, or $p = 10^{-5}$, when using the psychosis-associated SNPs' alleles as predictors) and the elastic net hyperparameters l_1 and λ. the optimal hyperparameters/feature set were chosen as the ones yielding the highest balanced accuracy (see below) across the inner CV cycle. The outer CV cycle is used to estimate the generalizability of the trained prediction model and was done with a LPO scheme.

5.2.4.4. Performance measures

The classifier's performance was evaluated using sensitivity (SE), specificity (SP), balanced accuracy (BAC), positive likelihood ratio (PLR), negative likelihood ratio (NLR), and diagnostic odds ratio (DOR) (see also the section **3.2.5.2.6. Performance measures** in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**). All the performance measures were measured using the validation subsamples in the CV and averaged across the 100 bootstrapped samples as a measure of stability of the performance using different samples (albeit the bootstrapped samples were all drawn from the same population). BAC was considered statistically significant across bootstrapped samples at a significance level of 5% (i.e., $p < .05$; see also the section **3.2.5.2.6. Performance measures** in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**). Additionally, the statistical significance for the BAC of each classifier (i.e. trained with each bootstrapped sample) was assessed using an one-sided binomial test as implemented in the caret R package and considered statistically significant with a $p < .05$.

5.2.4.5. Comparison between testing and validation balanced accuracies

The differences between the testing (i.e., from the inner CV cycle) and validation (i.e., from the outer CV cycle) BAC was assessed with a paired two-sided Wilcoxon signed rank test

and through Bland-Altman plots across bootstrapped samples for the ML models that have not reach a BAC statistically better than chance level (i.e. $p > .05$; see also the section **3.2.5.2.7. Comparison between testing and validation balanced accuracies** in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**). Moreover, this analysis was conducted considering only the novel model-based approach and the classifiers which BAC averaged across bootstrapped samples was not statistically significant ($p > .05$).

5.2.4.6. Clinical interpretability of features and applicability of the best classification model

The clinical interpretability of the features selected by the best classification model (i.e., the one showing the best BAC across bootstrapped samples and considering only the novel model-based approach) was assessed as described in the section **3.2.5.2.9. Clinical interpretability of features and applicability of the best classification model** in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**. The importance of the selected features was, however, computed using a different method. First, the beta coefficients of each selected genetic feature (each SNP or each eQTL score) were extracted for each CV outer cycle. Then, the median of the beta coefficients per risk factor was extracted and scaled by the number of times the risk factor was selected during the CV outer cycle. Moreover, the association between each selected genetic feature and the GAF and CAARMS score at baseline was assessed using Pearson correlation for eQTL scores (i.e. continuous feature) or univariate analysis of variance for SNP (i.e. categorial feature with three levels as each SNP can have one of three possible genotypes – allele 1/allele 1, allele 1/allele 2, or allele2/allele 2).

5.3. Results

5.3.1. Genetic classification analysis

Overall, the BAC of the classification models trained using the PRS, the list of psychosis-associated SNPs or the list of psychosis-associated genes for which a brain eQTL score was extracted and a bootstrapped sample (i.e., one of the 100 samples) ranged from 18% to 88% (**Figure 23**; **Tables 24**). Regarding the novel model-based approach (i.e. classification model trained with the SNPs or the eQTL scores), the model with the best classification

performance was trained with the SNPs as features and showed a BAC of 88% (SE = 86%, SP = 90%, DOR = 57.0, p = < .001). Furthermore, for the SNPs-based classifier, 23 classification models (in a total of 100 trained models) showed a BAC higher than chance level at an individual statistically significant level, 11 of which survived FDR correction for multiple comparisons. For the eQTL-based classifier, only 4 (out of 100) classification models showed a BAC higher than chance level, which did not survive FDR correction for multiple comparisons. Moreover, the median BAC across bootstrapped samples of the classification model trained with the psychosis-associated SNPs surpassed chance level with a statistical significance of 5%. However, the median BAC of the models trained with the PRS and the eQTL scores were not statistically better than chance level.

Table 24. Performance measures of the genetic classification models (i.e. trained with a schizophrenia polygenic risk score (PRS), a list of psychosis-associated single nucleotide polymorphisms (SNPs), or expression quantitative trait loci (eQTL) scores of a list of psychosis-associated genes expressed in the brain) across bootstrapped samples. Statistical significance of the median BAC across bootstrapped samples was tested using a one-tailed Wilcoxon signed rank test. *p < .05

	PRS	SNP	eQTL score
SE (%)	45.7 ± 16.1 [10.5 63.2]	52.1 ± 14.1 [19.0 85.7]	64.1 ± 16.9 [64.1 100]
SP (%)	42.2 ± 15.6 [10.5 63.2]	54.5 ± 15.8 [23.8 90.5]	29.3 ± 23.0 [0.0 85.7]
BAC (%)	43.9 ± 14.8 [18.4 63.2]	53.3 ± 13.4 [28.6 88.1]	46.7 ± 7.5 [31.0 66.7]
PLR	0.9 ± 0.5 [0.2 1.7]	1.5 ± 1.3 [0.4 9.0]	1.0 ± 0.4 [0.6 3.3]
NLR	1.8 ± 1.4 [0.6 7]	1.0 ± 0.5 [0.2 2.8]	1.7 ± 1.1 [0.5 6.0]
DOR	1.1 ± 1.0 [0.0 2.9]	3.3 ± 7.3 [0.2 57.0]	0.9 ± 1.0 [0.0 5.5]
p-value	>.999	.027*	>.999

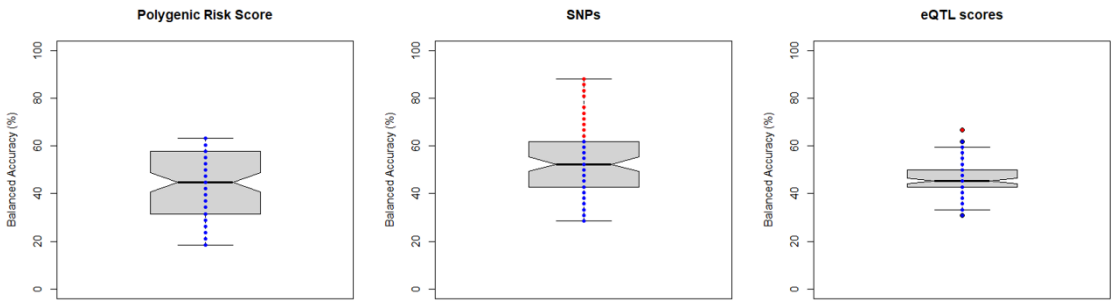


Figure 23. Balanced accuracy across bootstrapped samples for each classifier trained with the polygenic risk score, the list of psychosis-associated single nucleotide polymorphism (SNPs) or with the list of psychosis-associated genes for which an expression quantitative trait loci (eQTL) score was extracted. Dots represent the balanced accuracy value in each of the 100 bootstrapped samples and are red colored if the balanced accuracy is statistically significant (i.e., $p < .05$) or blue colored if it is not (i.e., $p > .05$). The statistical significance of the balanced accuracy in each bootstrapped sample was evaluated through a binomial test.

5.3.2. Comparison between testing and validation balanced accuracies

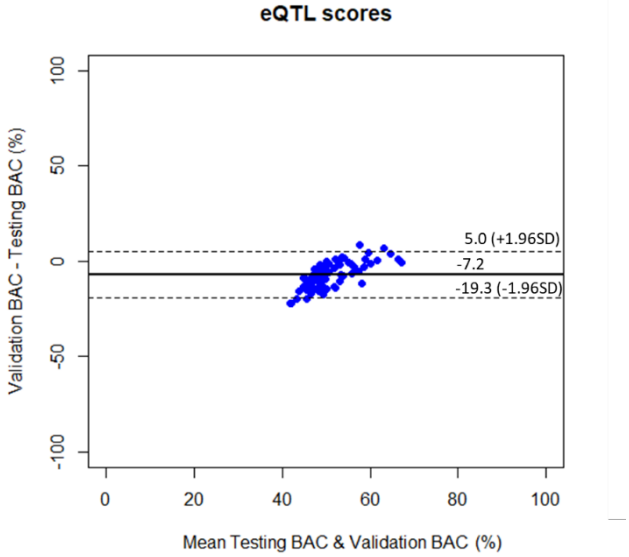


Figure 24. Bland-Altman plot with limits (dashed lines) of agreement for mean (continuous line) testing and validation balanced accuracies (BAC) of classification models trained with the eQTL scores for psychosis-associated genes expressed in the brain.

The difference between the testing and the validation BAC was statistically significant for the eQTL scores-based classifiers ($p < .001$; d = 1.2). Moreover, the validation BAC was on

average lower than the testing BAC (**Figures 24**), with a large effect size for the eQTL scores-based classifier (i.e., $d > 0.8$) (**Figure 24**).

5.3.3. Clinical interpretability of features and applicability of the best classification model

The genetic features (i.e. SNPs) with the highest importance (i.e., with magnitude 1.96 standard deviations above or below the mean importance across all features) assigned by the best classification models were the rs7013471 and the rs6504751 (see **A2.Table 2** in the **Appendix 2**).

The association analysis between the genotype of the SNPs selected by the best classification model and the clinical assessment scores (i.e., GAF and CAARMS at baseline) was not statistically significant for any of the features (i.e., FDR corrected $p > .05$; **A2.Table 2**).

The best classification models used as features (and potential clinical biomarker) SNPs previously associated with psychosis. As there is no previous consistent evidence of the positive effect of each of these features on the transition to psychosis from an ARMS, nor of the use of these same features in transition to psychosis from an ARMS classification models, the present study may be considered as exploratory. Furthermore, herein we controlled for relevant extraneous variables (i.e., matching groups – ARMS-T and ARMS-NT – for age, and sex). Therefore, this biomarker scores only 1 (out of 4) in the quality of evidence dimension of the clinical applicability assessment scale (**Table 16**). Moreover, although the best classification models showed a large effect size (DOR = 57.0), the biomarker is not assigned a score for the effect size dimension, because it was assigned with a quality of evidence score of only 1 (**Table 16**). Therefore, this biomarker should not be considered as a clinically applicable one.

5.4. Discussion

This study applied, for the first time, ML to genetic data using three types of genetic features to detect transition to psychosis from an ARMS: a) a schizophrenia PRS (previous model-based approach), b) a list of psychosis-associated SNPs, and c) a list of psychosis-associated genes for which brain tissue's eQTL scores were extracted (novel model-based approach).

5.4.1. Prediction of transition to psychosis using genetics

Genetic data was shown to predict the transition to psychosis from an ARMS marginally better than chance level when a list of psychosis-associated SNPs was use as features. Furthermore, the odds of getting a true positive with these SNP-based classifier models is, on average, 3-fold higher than getting a false positive (i.e. DOR = 3.3). This is, to the best of my knowledge, the first piece of evidence suggesting that common variants in the genome (i.e. SNPs) might predict psychosis already from a prodromal stage of the disease, albeit with a low average accuracy. SNPs-based classification models have been previously showed to classify schizophrenia (Aguiar-Pulido et al., 2010; Vivian-Griffiths et al., 2019; Yang et al., 2010), and FEP patients (Pettersson-Yeo, Benetti, Marquand, Dell'Acqua, et al., 2013) (*vs.* healthy controls) better than chance level, but not subjects at an ARMS *vs.* healthy controls or FEP patients (Pettersson-Yeo, Benetti, Marquand, Dell'acqua, et al., 2013). Furthermore, one of these studies has selected a list of SNPs from the PGC2 (Ripke et al., 2014) (Vivian-Griffiths et al., 2019), which potentially overlaps with the ones selected in this study (Pardiñas et al., 2018).

Despite the (scarce) evidence of the potential of PRS for schizophrenia (Antonucci et al., 2020; Chen et al., 2018; Vivian-Griffiths et al., 2019) to classify schizophrenia patients (*vs.* healthy controls) I were not able to predict transition to psychosis from an ARMS using this type of genetic feature. One possible explanation for the PRS negative results is that although the genetic architecture, conveyed through a PRS, has been showed to differ between schizophrenia patients and healthy controls, one cannot exclude the possibility that this genetic architecture, that is specific to schizophrenia (a fully developed psychotic disorder), might be present in all subjects at an ARMS, i.e. those who later transition to psychosis and those who do not. Moreover, the constellation of genetic variations (i.e. SNPs) that might confer susceptibility to transition to psychosis already from a prodromal stage does not necessarily has to be the same as the one for schizophrenia (when drawn in comparison to healthy controls). This serves also as an explanation for the advantage of using a less hypothesis-based approach for the selection of genetic features (as we did by pre-selecting a large list of SNPs and performing an embedded feature selection using elastic net regression). Furthermore, a PRS for transition to psychosis from an ARMS would possibly be a better predictor than the PRS for schizophrenia. However, this would require a larger sample to estimate the size of each SNP effect on the transition to psychosis, which might

be provided by multicenter projects, such as NAPLS 2 (Addington et al., 2012) and PRONIA (http://pronia.eu) over the next years.

Furthermore, eQTL scores for psychosis associated genes expressed in the brain were also not able to predict transition to psychosis from an ARMS. Only one previous study has shown the predictive value of gene expression profiling in the frontal brain region in classifying schizophrenia patients (*vs.* healthy controls) (Struyf et al., 2008). In this study, instead of actual gene expression measures I used a proxy for genetically regulated component of the expression of genes, the eQTL scores. Although I have computed eQTL scores only for the genes having a validated eQTL score model (see **4. Creating a brain-based gene expression quantitative trait loci score (eGenScore) tool**), this does not guarantee that the estimated gene expression represents (or correlates perfectly with) the real levels of the expression. Furthermore, although I have selected the initial list of genes as the ones most associated with schizophrenia (*vs.* healthy controls), this selection did not take into account the expression profile of these genes in the brain, and I have computed an eQTL score for several brain tissues. A future improvement of this step would be to test an eQTL scores-based classifier with a selection of genes a) that are highly expressed in the brain in healthy subjects, b) that are highly expressed in the brain and which expression is modulated by the schizophrenia diagnosis, and c) that are highly expressed in the brain and which expression is modulated by the transition to psychosis from an ARMS.

5.4.2. Comparison between testing and validation balanced accuracies

I further investigated the possible reasons for the poor performance of the eQTL scores-based classification models (i.e., with balanced accuracies that were not significantly better than chance level). I found a significant decrease in the validation balanced accuracies, compared to the testing ones (across bootstrapped samples) with a large effect size. At least 89% of the validation balanced accuracies were lower than the mean of the testing balanced accuracies (i.e. d = 1.2). This is suggestive of some degree of overfitting during the training/testing phase of the classification models and might explain the poor generalization when they are tested on unseen data (i.e. during the validation phase).

5.4.3. Clinical interpretability of features and applicability of the best classification model

I additionally analyzed the SNPs-based model that individually yield the best significant performance (i.e., the highest balanced accuracy with a $p < .05$) in terms of its clinical interpretability and potential applicability. Among the features contributing the most for the prediction of the transition to psychosis are rs7013471 and the rs6504751 SNPs, both contributing with a positive weigh (note that positive final score implies a classification of a subject as ARMS-T). Indeed, these two SNPs' genotype can be interpreted as to have a risk effect in psychosis, i.e., given the genotypes for these two SNPs of a subject at an ARMS, the higher the number of risk alleles (i.e. 0, 1 or 2; adenine-A for rs7013471 or guanine-G for rs6504751), the more likely this subject is to be classified as ARMS-T. Moreover, rs7013471 and rs6504751 are intron variants of the MFHAS1 (Malignant fibrous histiocytoma-amplified sequence 1) and CA10 (Carbonic anhydrase-related protein 10) genes, respectively. Although these two variants/genes have not been implicated in the transition to psychosis from an ARMS, the direction of these SNPs' genotype effect on the schizophrenia diagnosis (*vs.* healthy controls) (Pardiñas et al., 2018) is the same as the one defined by the SNPs-based classifier for the transition to psychosis from an ARMS (i.e. with the risk alleles increasing the risk for psychosis). Interestingly, CA10 has been implicated in the regulation of the neurexin expression, which is a very important protein that mediates synapse organization and facilitates synaptic transmission (Montoliu-Gaya et al., 2021). Genetic variations in neurexin genes, in turn, have been widely reported to be associated with schizophrenia (Kasem et al., 2018).

Furthermore, the statistical association between the genotype of the SNPs selected by the best classification model (including rs7013471 and rs6504751) and the clinical assessments at baseline (i.e., GAF and CAARMS) was not significant in any of the SNPs. This suggestively supports the use of genetics as a complement to the clinical assessments as they seem to convey non-overlapping information. However, before these SNPs can be considered a potential clinically applicable biomarker, at least a replication of this study must be conducted.

5.4.4. Limitations

This study has a few limitations that need to be addressed. First, the classification models performance, i.e. accuracy, may be limited by the lack of power due to the small sample size of the bootstrapped samples, i.e. the ones used to train the models. This is a critical limitation for the SNPs- and eQTL scores-based classifiers which are trained with high dimensional data, i.e. high number of genetic features. Therefore, the results reported in this study need further replication using larger samples. Second, these findings are not generalizable to populations with non-European ancestry as the sample we have used restricted ancestry to European. As the genetic architecture differs between populations with different ancestry, generalizability of this study's findings to other populations requires the exact replication of the methodology used in this study in those populations. This includes the selection of psychosis-associated SNPs and genes using the findings from a GWAS that have been conducted using a sample drawn from a population with a non-European ancestry.

In this study, I explored, for the first time, the value of genetics in predicting transition to psychosis from ARMS using machine learning. Overall, the psychosis-associated SNPs-based classifier could predict transition to psychosis with an accuracy (averaged across the bootstrapped samples) statistically better than chance level, albeit marginally. Furthermore, the performance of the PRS- and eQTL scores-based classifiers was poor, i.e. not better than chance level. These results suggest that genetic data might be a promising predictor of transition to psychosis from an ARMS. However, these findings need future replication in larger samples and in samples drawn from non-European ancestry population, as this study was conducted using only subjects with European ancestry.

6. Individual prediction of transition to psychosis using environmental data and machine learning

6.1. Introduction

Exposure to environmental risk factors has long been associated to the increased risk of developing psychosis (mainly in the form of schizophrenia) in univariate analysis (see **1.5.1. Potential environmental biomarkers** in **1. General introduction**). In particular, childhood trauma (Thompson et al., 2014) – a risk factor that has been proposed to lead to the disease vulnerability – and a cannabis consumption (Kraan et al., 2016) – a risk factor proposed to induce the onset of the disease – have been shown to increase the risk of transition to psychosis from an ARMS. Moreover, environmental risk factors have only been used in classification studies once and to classify schizophrenia patients (*vs.* healthy controls) (Antonucci et al., 2020). Therefore, to the best of my knowledge, environmental data has never been explored as a predictor of the transition to psychosis from an ARMS.

Herein, I explored, for the first time, the potential of environmental risk factors to predict transition to psychosis from an ARMS using a sample of 134 individuals and two model approaches: a previous and a novel model-based approach. In the previous model-based approach, I used a schizophrenia environmental risk score (ERS), which the proof of concept has been previously presented (Vassos et al., 2020), as a transition to psychosis predictor. In the novel model-based approach, I tested a list of individual environmental risk factor as predictors.

6.2. Materials and methods

6.2.1. Sample description

The sample used in this study was described in detail in the **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**. All the 246 individuals at an ARMS from the original pool had environmental risk factors assessment data. Over the 2-year follow-up period, 60 individuals at an ARMS developed psychosis (ARMS-T) and the remaining 186 did not (ARMS-NT). Socio-demographic (age at baseline and follow-up or transition, sex, years of education, and intelligence quotient) and clinical variables (global assessment of functioning and comprehensive assessment at risk mental state scores at baseline and follow-up or transition) of the subsample analyzed with machine learning (ML) were analyzed using a two-tailed independent *t*-test for

continuous data and a chi square test for ordinal data (**Table 25**). These statistical analyses were performed using the Statistical Package for the Social Sciences 26 (SPSS 26 for Windows, Chicago, IL, USA).

Table 25. Socio-demographic and clinical information of the ARMS sample with environmental data (with less than 20% of the environmental risk factors missing).

	ARMS-T (n = 37)	**ARMS-NT** (n = 97)	**Group comparison**
Age at baseline (years)	23.6 ± 4.8 [15.6 33.6]	21.9 ± 3.7 [14.6 33.1]	p = .027*
Age at follow-up or transition (years) [a]	25.6 ± 5.6 [17.3 39.2]	27.1 ± 4.7 [18.5 41.2]	p = .131
Sex (male/female)	22/15	50/47	p = .411
Years of education [b]	13.2 ± 2.7 [8 18]	13.3 ± 2.0 [9 18]	p = .686
IQ (z-standardized) [c]	-0.3 ± 1.0 [-2.1 2.2]	0.1 ± 1.0 [-2.1 3.5]	p = .049*
GAF at baseline [d]	55 ± 12.5 [35 90]	56.7 ± 8.6 [40 80]	p = .523
GAF at follow-up [e]	50.4 ± 19.9 [10 88]	63.2 ± 14.2 [20 94]	p = <.001*
CAARMS at baseline [f]	30.9 ± 19.4 [0 78]	28.3 ± 16.0 [0 81]	p = .478
CAARMS at follow-up [g]	29.7 ± 31.2 [0 102]	13.3 ± 14.2 [0 60]	p = <.001*

Data format: mean ± standard deviation [min max]. Information not available for [a]1 ARMS-T; [b]5 ARMS-T and 6 ARMS-NT; [c]7 ARMS-T and 13 ARMS-NT; [d]5 ARMS-T and 4 ARMS-NT; [e]5 ARMS-T and 8 ARMS-NT; [f]6 ARMS-T and 13 ARMS-NT; [g]4 ARMS-T and 8 ARMS-NT; subject. ARMS: at-risk mental state; ARMS-T: individuals at ARMS that later transitioned to psychosis; ARMS-NT: individuals at ARMS that did not transition to psychosis. *$p < .05$.

6.2.2. Environmental risk factor assessment

Each subject was assessed on, at least one of, nine environmental risk factors: 1) alcohol, 2) tobacco and 3) cannabis consumption; 4) being migrant; 5) belonging to an ethnic minority; 6) the upbringing urbanicity level; 7) the parental (i.e. mother and father) age at birth; 8) the presence of childhood trauma; and 9) the season of birth. Exposure to alcohol or tobacco was binarily defined as any exposure if the subject used to drink at least one unit a day or smoke at least one cigarette a day or as no exposure if otherwise. Exposure to cannabis was assessed in three main domains: consumption onset (late (score 1)/early (score 2) onset), frequency (infrequent (score 1)/frequent (score 2)) and duration (short- (score 1)/long (score

2)-term use). A composite score was computed as the sum of the consumption onset, frequency and duration scores (possible cannabis consumption score range: [0, 6]). Subjects belonged to an ethnic minority if they have self-reported has being as other than white or as white, but migrant. Therefore, belonging to an ethnic minority score was defined as 0 for native subjects, 1 for white migrants, 2 for self-reported black ethnicity and 3 for other self-reported ethnicity (e.g. asian, mixed). Risk for psychosis was defined as the lowest if the subject was upbrought in a village and highest if upbrought in a city, with being upbrought in a town conferring an intermediate risk (i.e. score 1 for village, 2 for town and 3 for city). Childhood trauma was assessed using the self-reported childhood experience of care and abuse questionnaire (Bifulco et al., 2005). Risk for psychosis due to exposure to childhood trauma was defined as a sum of exposure to the following events weighted by the number of years of exposure of each event: a) being bullied, b) being hit repeatedly, c) having seen or heard family violence, d) being separated from a parent for more than 1 year, and e) having been institutionalized. Finally, risk for psychosis was defined as higher if the subjects were born in the summer or autumn (score of 2) compared to being born in the winter or spring (score of 1).

6.2.3. Machine learning approach

6.2.3.1. Sample balancing and bootstrapping

The final sample used in the environment ML analysis was composed by the 37 (for the previous model-based approach) or 17 (for the novel model-based approach) ARMS-T subjects and 37 or 17 ARMS-NT subjects, respectively, randomly selected to match the ARMS-T for age and sex (see **5.2.4.1. Sample balancing and bootstrapping** in **5. Individual prediction of transition to psychosis using genetics and machine learning**). The subsampling was repeated 100 times, i.e., 100 bootstrapped samples were created, and the subsequent ML analysis was conducted for each of the bootstrapped sample.

6.2.3.2. Previous model-based approach

Vassos and colleagues have recently developed a method to compute an aggregated ERS for psychosis (Vassos et al., 2020). Herein, I tested ERS for psychosis as a predictor of the transition to psychosis from ARMS. This analysis was conducted using R software 4.0.5 (R Core Team, 2018).

6.2.3.2.1. Environmental risk score extraction

The ERS was computed for each subject in this study's sample (i.e. ARMS sample) as the sum of the exposure to each environmental risk factor that have been previously associated with the diagnosis of schizophrenia (*vs.* healthy controls; $environmental\ risk\ factor_i, i = 1, \dots, m = number\ of\ risk\ factors$) weighted by their effect size, i.e. relative risk, $RR_{environmental\ risk\ factor_i}, i = 1, \dots, m$, in that association and adjusted by the average exposure to the risk factor in the general population, $\sum_j^n RR_{i,j} \times p_{i,j}, j = 1, \dots, n = number\ of\ risk\ levels\ for\ risk\ factor\ i$, (**Eq. 15**).

$$ERS = \sum_i^m \left(\log \left(\frac{RR_{environmental\ risk\ factor_i}}{\sum_j^n RR_{environmental\ risk\ factor_i,j} \times p_{environmental\ risk\ factor_i j}} \right) \times 10 \right), \qquad (\textbf{Eq. 15})$$

In more detail, 7 environmental risk factors were included in the ERS (i.e. tobacco and cannabis consumption; belonging to an ethnic minority; the upbringing urbanicity level; the paternal age at birth; the presence of childhood trauma; and the season of birth). For each factor, the most recent meta-analysis assessing the risk that the exposure to each factor confers to psychosis was selected and its effect size was extracted. Furthermore, whenever available, crude effect sizes or with minimum adjustment were selected and considered as RR even if presented as odds ratio, incidence rate ratios or RR. Moreover, these three effect size measures are good approximations of RR because psychosis is a rare outcome (i.e. psychosis as a diagnosis *vs.* healthy controls). The RRs in the meta-analyses are defined by comparing exposed with non-exposed individuals, with the risk factor defined as a binary or an ordinal variable. Then, the effect sizes are scaled to the averaged individual, i.e. by extracting the expected proportion of individuals in each level of the risk factor in the general population. This scaling is done under the assumption that only a minority of the population would not be exposed to any risk. The next step was to multiply the logarithm of the scaled RRs by 10 and to round them by the nearest integer. Finally, ERS was computed as the sum of each of the risk factor component, where missing values, i.e. missing risk factors, were replaced by zero. **Table 26** presents the information regarding each environmental risk factor, including the meta-analysis from which the exposure risk effect size was computed, the proportion of the general population that is exposed to the risk (by level of exposure),

the exposure risk effect size scaled by this general population proportion, and the contribution of each risk factor (by level of exposure) to the ERS. The selection of the meta-analysis from which the effect sizes were extracted, definition of each factor's exposure levels, and the general population exposure to the factor was originally presented in (Vassos et al., 2020) for all the risk factors except for tobacco consumption and season of birth. For this last one, I followed the same methods as described in (Vassos et al., 2020) to select the meta-analysis, define the level of exposure and extract the proportion of the general population that is exposed to the risk (by level of exposure). Only subjects with less than 20% of missing information (i.e. missing data for less than 2 environmental risk factors) were considered for the ML analysis. Therefore, the final sample included 37 ARMS-T subjects and 97 ARMs-NT subjects.

Table 26. Contribution of each environmental risk factor to the environmental risk score (ERS). RR: relative risk.

Risk factor (meta-analysis)	**Sub-categories**	**RR from meta-analysis**	**Proportion of population** (%)[1]	**Scaled log(RR)**	**ERS component**
Upbringing urbanicity (Vassos et al., 2012)	Low	1.16	33.3	-0.14	-1.5
	Medium	1.55	33.3	-0.01	0
	High	2.07	33.3	0.11	1
Cannabis consumption (Marconi et al., 2016)	No exposure	1	70	-0.12	-1
	Little/moderate	1.41	15	0.02	0
	High exposure	2.77	15	0.32	3
Paternal age (B. Miller et al., 2011)	<40	1	92.1	-0.01	0
	40-50	1.17	7.1	0.06	0.5
	>50	1.60	0.8	0.19	2
Childhood trauma (overall) (Varese et al., 2012)	No exposure	1	73	-0.17	-1.5
	Any exposure	2.78	27	0.27	2,5
Season of birth (Davies et al., 2003)	Summer/Autumn	1	96.7	-1.00×10^{-3}	0
	Winter/Spring	1.07	3.3	0.03	0.5
Tobacco consumption (Hunter et al., 2018)	No exposure	1	70.6	-0.11	-1
	Any exposure	1.99	29.4	0.19	2
Ethnic minority (Bourque et al., 2011)	Native	1	92.4	-0.04	-0.5
	Black	4	1.3	0.56	5.5
	White	1.8	2.8	0.22	2
	Other	2	3.5	0.26	2.5

[1]Source for tobacco consumption: WHO – Prevalence of tobacco smoking 2016 (http://gamapserver.who.int/gho/interactiveharts/tobacco/use/atlas.html).

6.2.3.2.2. Logistic regression for classification

Binary classification of transition to psychosis from an ARMS (i.e. ARMS-T *vs.* ARMS-NT) was performed using logistic regression as described in the section **5.2.4.2.2. Logistic regression for classification** in **5. Individual prediction of transition to psychosis using genetics and machine learning**).

6.2.3.2.3. Cross-validation

The logistic regression was trained and tested in a simple leave-one per group-out (LPO) CV scheme (**Figure 10-inner CV cycle**; see also the section **1.2. Machine Learning** in **1. General introduction** and the section **3.2.5.2.5. Cross-validation** in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**).

6.2.3.3.Novel model-based approach

One novel prediction model for transition to psychosis from environmental data using the ARMS sample was trained and tested. Each environmental risk factor described in the section **6.2.2 Environmental risk factor assessment** was used as an individual feature in the model. Only subjects with information for all the environmental risk factors (i.e. with no missing information) were considered for this ML analysis. Therefore, the final sample included 17 ARMS-T subjects and 49 ARMs-NT subjects The analysis was conducted using R software 4.0.5 (R Core Team, 2018).

6.2.3.3.1. Elastic net for classification

Binary classification of transition to psychosis from an ARMS (i.e. ARMS-T *vs.* ARMS-NT) was performed using logistic regularized regression with elastic net as described in the section **5.2.4.3.2. Elastic net for classification** in **5. Individual prediction of transition to psychosis using genetics and machine learning**.

6.2.3.3.2. Cross-validation

The classifier was trained using the nested-CV scheme described in the section **5.2.4.3.3. Cross-validation** in **5. Individual prediction of transition to psychosis using genetics and machine learning**.

6.2.3.4. Performance measures

The classifier's performance was evaluated using the same measures as described in the section **5.2.4.4. Performance measures** in **5. Individual prediction of transition to psychosis using genetics and machine learning**.

6.2.3.5. Comparison between testing and validation balanced accuracies

The differences between the testing (i.e. from the inner CV cycle) and validation (i.e. from the outer CV cycle) BAC was assessed only for the novel model-based approach and as described in the section **5.3.2. Comparison between testing and validation balanced accuracies** in **5. Individual prediction of transition to psychosis using genetics and machine learning**.

6.2.3.6. Clinical interpretability of features and applicability of the best classification model

The clinical interpretability of the features selected by the best classification model (i.e., the one showing the best BAC across bootstrapped samples and considering only the novel model-based approach) was assessed as described in the section **3.2.5.2.9. Clinical interpretability of features and applicability of the best classification model** in **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**. The importance of the selected features was computed using the method described in the section **5.2.4.6. Clinical interpretability of features and applicability of the best classification model** in **5. Individual prediction of transition to psychosis using genetics and machine learning**. First, the beta coefficients of each selected environmental risk factors were extracted for each CV outer cycle. Then, the median of the beta coefficients per risk factor was computed and scaled by the number of times the risk factor was selected during the CV outer cycle. Moreover, the association between each selected environmental risk factor and the GAF and CAARMS score at baseline was assessed using Pearson

correlation for continuous risk factor, two-tailed *t*-test for binary risk factors or univariate analysis of variance for categorial risk factors with more than one level.

6.3. Results

6.3.2. Environment classification analysis

Overall, the BAC of the classification models trained and validated using the ERS or the environmental risk factors as features and bootstrapped sample (i.e., one of the 100 samples) ranged from 22% to 71% (**Figure 25**; **Tables 27**). Regarding the novel model-based approach (i.e., classification model trained with the individual environmental risk factors as features), the best classification performance showed a BAC of 71% (SE = 59%, SP = 82%, DOR = 6.7, p = .012). Furthermore, only 3 classification models (in a total of 100 trained models) showed a BAC higher than chance level at an individual statistically significant level. However, this statistical significance did not survive the FDR correction for multiple comparisons. Moreover, the median BAC across bootstrapped samples of the classification model trained with the environmental risk factors and the ERS were not statistically better than chance level.

Table 27. Performance measures of the environmental risk factors classification model (i.e., trained with an environmental risk score (ERS) or a list of environmental risk factors) across bootstrapped samples. Statistical significance of the median BAC across bootstrapped samples was tested using a one-tailed Wilcoxon signed rank test. *p < .05

	ERS	Environmental risk factors
SE (%)	49.6 ± 13.4 [13.5 62.2]	11.1 ± 13.2 [0 58.8]
SP (%)	52.1 ± 10.9 [18.9 67.6]	80.4 ± 12.0 [47.1 100]
BAC (%)	50.9 ± 11.8 [21.6 64.9]	45.7 ± 6.0 [29.4 70.6]
PLR	1.1 ± 0.4 [0.2 1.9]	0.5 ± 0.9 [0 7.0]
NLR	1.1 ± 0.7 [0.6 4]	1.1 ± 0.2 [0.5 1.7]
DOR	1.5 ± 0.8 [0.1 3.4]	0.6 ± 1.4 [0 11.2]
p-value	.069	>.999

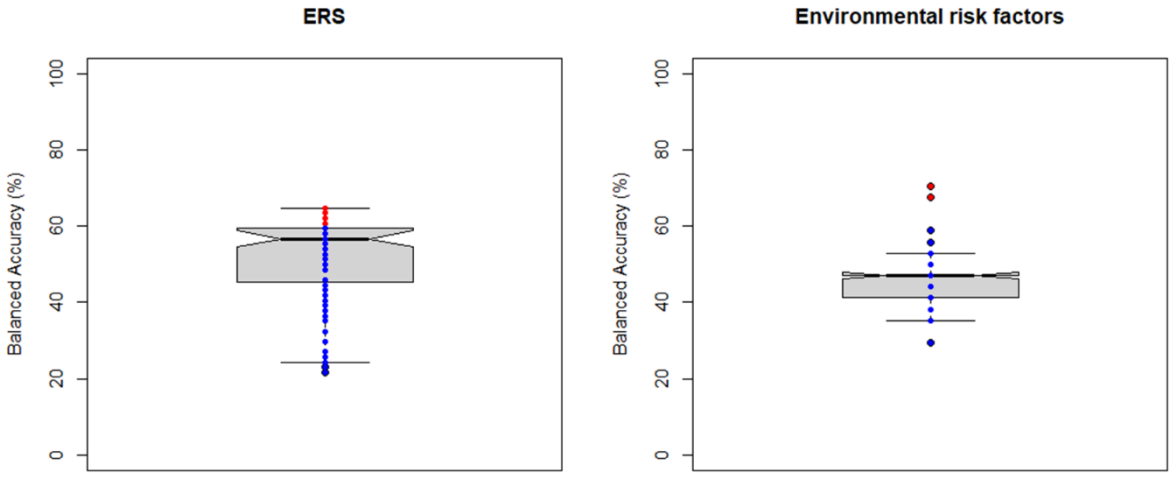


Figure 25. Balanced accuracy across bootstrapped samples for each classifier trained with the environmental risk score (ERS) or with each environmental risk factors as features. Dots represent the balanced accuracy value in each of the 100 bootstrapped samples and are red colored if the balanced accuracy is statistically significant (i.e., $p < .05$) or blue colored if it is not (i.e., $p > .05$). The statistical significance of the balanced accuracy in each bootstrapped sample was evaluated through a binomial test.

6.2.2. Comparison between testing and validation balanced accuracies

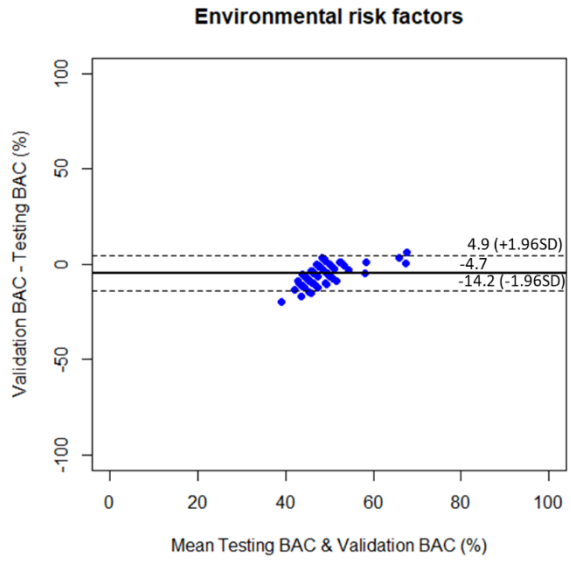


Figure 26. Bland-Altman plot with limits (dashed lines) of agreement for mean (continuous line) testing and validation balanced accuracies (BAC) of classification models trained with the individual environmental risk factors.

The difference between the testing and the validation BAC was statistically significant for the environmental risk factors-based classifier ($p < .001$; d = 1.0). Moreover, the validation BAC was on average lower than the testing BAC (**Figures 26**), with a large effect size (i.e., $d > 0.8$).

6.2.3. Clinical interpretability of features and applicability of the best classification model

The environmental features (i.e., environmental risk factors) that have been selected by the best classification model were being a migrant, belonging to an ethnic minority, having experienced trauma in childhood and the season of birth (**Table 28**).

The association analysis between the environmental risk factors selected by the best classification model and the clinical assessment scores (i.e., GAF and CAARMS at baseline) was statistically significant for migration and ethnic minority and GAF score at baseline (i.e., FDR corrected *p*-value < .05; **Table 28**).

The best classification model used as features (and potential clinical biomarker) migration, ethnic minority, childhood trauma and season of birth. As there is no previous consistent evidence of the positive effect of each of these features on the transition to psychosis from an ARMS, nor of the use of these same features in transition to psychosis from an ARMS classification models, the present study may be considered as exploratory. Furthermore, herein I controlled for relevant extraneous variables (i.e., matching groups – ARMS-T and ARMS-NT – for age, and sex). Therefore, this biomarker scores only 1 (out of 4) in the quality of evidence dimension of the clinical applicability assessment scale (**Table 16**). Moreover, although the best classification models showed a large effect size (DOR = 6.7), the biomarker is not assigned a score for the effect size dimension, because it was assigned with a quality of evidence score of only 1 (**Table 16**). Therefore, this biomarker should not be considered as a clinically applicable one.

Table 28. Interpretability analysis (feature importance and association analysis) of the environmental risk factors selected by the best classification model. Pearson correlation (*r*, childhood trauma), two-tailed t-test (*t*, Cohen's *d*, migration and season of birth) or univariate analysis of variance (*F*, η^2, ethnic minority) was computed between each environmental risk factor value and clinical assessments tools at baseline [i.e., the global assessment functioning (GAF) and comprehensive assessments of at-risk mental state (CAARMS)].

Environmental risk factor	**Feature importance**	**GAF at baseline**			**CAARMS at baseline**		
		Statistic Effect size	***p***	**FDR *p***	**Statistic Effect size**	***p***	**FDR *p***
Migration	3.25×10^{-3}	$t = 2.6$ $d = 1.1$	.016*	.033*	$t = 1.3$ $d = 0.5$	.218	.365
Ethnic minority	1.98×10^{-4}	$F = 5.1$ $\eta^2 = .37$	.007*	.027*	$F = 1.2$ $\eta^2 = .11$	.375	.375
Childhood trauma	3.62×10^{-5}	$r = -.13$	.509	.679	$r = -.39$	.035*	.140
Season of birth	-3.62×10^{-1}	$t = .17$ $d = 0.1$	.869	.869	$t = 1.1$ $d = 0.4$	.274	.365

6.3. Discussion

This study applied, for the first time, ML to environmental data using two types of features to detect transition to psychosis from an ARMS: a) a schizophrenia ERS (previous model-based approach), and b) a list of environmental risk factors (novel model-based approach).

6.3.2. Prediction of transition to psychosis using environmental risk assessment

Overall, environmental risk assessment, i.e., ERS, or individual risk factors, could not predict transition to psychosis from an ARMS with an averaged accuracy, i.e. across bootstrapped samples, better than chance level. These findings are in line with the only other report using environmental data with machine learning to classify schizophrenia (*vs.* healthy controls) (Antonucci et al., 2020). They have reported a balanced accuracy not statistically better than chance level, such as me herein. Comparing to my study, their ML model included as features a) the presence of obstetric complication during pregnancy, labor and delivery, b) the presence of developmental anomalies during the first 66 months of life, c) the parental socio-economic status, which I have not, and d) the level of urbanicity, and e) the parental age at the time of birth of the individual, which I have also included in my model (Antonucci et al., 2020). Moreover, they have not performed feature selection during the model training – as I did by using the elastic net algorithm – and, therefore, included the 5 risk factors in the model. Additionally, they trained and tested the model in a relatively large, albeit unbalanced, sample (103 patients with schizophrenia and 337 healthy controls, i.e.,

approximately 13 times bigger than mine). Although a) previous evidence suggests that environmental risk factors cannot distinguish patients with a psychotic disorder from healthy controls and b) I have used a sample balanced for possible extraneous factors, such as age and sex, albeit with smaller size, a slightly higher number of risk factors as features, and performed feature selection to reduce the number of features to train the ML model, I cannot fully exclude the lack of statistical power as a potential explanation for our negative findings.

Furthermore, ML model trained with the ERS for schizophrenia, which I have tested as an exploratory predictor to the transition to psychosis from an ARMS, showed a poor performance, i.e., a balanced accuracy similar to chance level. Indeed, ERS is a composite score of individual risk factors computed under the assumption that the risk factors are completely independent, which has been shown not to be the case (Guloksuz et al., 2018). This simplistic approach may limit the ability of the ERS to capture the real environmental architecture underlying psychosis. Moreover, an ERS for transition to psychosis from an ARMS would possibly be a better predictor than the ERS for schizophrenia (Padmanabhan et al., 2017). However, this approach would require large-scale studies to assess environmental risk factors in ARMS population and using a more homogeneous effect size measurements (e.g., using either OR or RR). Indeed, as transition to psychosis is a more common outcome than schizophrenia (i.e., with a 30% transition rate *vs.* 1% of schizophrenia diagnosis), aggregating environmental risk factors which effects sizes were measured using different metrics is no longer a valid approach.

6.3.3. Comparison between testing and validation balanced accuracies

The poor performance of the environmental risk factors-based classification model might be explained by the presence of some overfitting during the training/testing phase of the model. In fact, I found a significant decrease in the validation balanced accuracies compared to the testing ones (across bootstrapped samples) with a large effect size. At least 84% of validation balanced accuracies were lower than the mean of the testing balanced accuracies (i.e. d = 1.0). Therefore, this might explain the poor generalization when they are tested on unseen data (i.e., during the validation phase).

6.3.4. Clinical interpretability of features and applicability of the best classification model

I additionally analyzed the environmental risk scores-based model that individually yield the best significance performance (i.e., the highest balanced accuracy with a $p < .05$) in terms of its clinical interpretability and potential applicability. The features that have been selected by the ML model were, from the most important to the least (in absolute values), the season of birth, being a migrant, belonging to an ethnic minority and having experience trauma in the childhood, all except season of birth contributing with a positive weight (note that positive final score implies a classification of a subject as ARMS-T). Indeed, being born in the winter/spring, being a migrant, belonging to an ethnic minority and experiencing childhood trauma increases the likelihood of this subject being classified as ARMS-T. Although the majority of these risk factors have not been implicated in the transition to psychosis from an ARMS, the direction of these risk factors' effect (i.e., being born in the winter/spring (Davies et al., 2003), being a migrant (Bourque et al., 2011; Selten et al., 2019), and belonging to an ethnic minority (Bourque et al., 2011)) on the schizophrenia diagnosis (*vs.* healthy controls) is the same as the one defined by the environmental risk factors-based classifier for the transition to psychosis from an ARMS (i.e., with the exposure to these factors increasing the risk for psychosis). Moreover, the severity of the childhood trauma, in particular sexual abuse, has been shown to increase the risk of transition to psychosis from an ARMS (Thompson et al., 2014).

Furthermore, the statistical association between the risk factors selected by the best classification model and the clinical assessments at baseline (i.e., GAF and CAARMS) was significant for being a migrant and belonging to an ethnic minority and GAF score. All the other associations were not statistically significant. This suggests that the exposure to these risk factors has an impact on the global functioning of the ARMS subjects, with being a migrant or belonging to an ethnic minority being associated with lower global functioning. Moreover, the fact that no other risk factor – clinical assessment association was statistically significant suggestively supports the use of environment as a complement to the clinical assessments as they seem to convey almost non-overlapping information. However, before these environmental risk factors can be considered a potential clinically applicable biomarker, at least a replication of this study must be conducted.

6.3.5. Limitations

The major limitation of this study was the restricted sample size, particularly for training and testing the environmental risk factors-based classification models, which required to have non-missing data for any of the included risk factors. This potentially limits the performance of the models by the lack of power. One potential solution for this limitation would be to impute missing information (i.e., for the environmental risk factors-based classification models) using, for example, multiple imputation methods (Jakobsen et al., 2017).

In this study, I explored, for the first time, the value of environmental risk assessment in predicting transition to psychosis from ARMS using ML. Overall, the ERS- and environmental risk factors-based classifiers could not predict transition to psychosis with an accuracy (averaged across the bootstrapped samples) statistically better than chance level. These results suggest that environmental data alone cannot be used as a predictor of transition to psychosis from an ARMS. However, these findings need future replication in larger samples.

7. General discussion

The main aim of this project was to predict transition to psychosis from an ARMS using machine learning applied to multimodal quantitative data – i.e., neuroimaging (structural magnetic resonance imaging, sMRI), genetics (genome-wide genotypes), and environment (environmental risk assessment) – collected when the subjects first sought clinical help (i.e., at baseline) and were identified with an ARMS. This is, to the best of my knowledge, the first study applying a multimodal approach to predict the development of a psychotic disorder already from a prodromal stage. Three main and two secondary objectives were fulfilled under the scope of this project. In particular, the predictive value of each modality was individually tested by assessing the accuracy of the modality-specific ML models (i.e., neuroimaging – **3. Individual prediction of transition to psychosis using structural neuroimaging and machine learning**; genetics – **5. Individual prediction of transition to psychosis using genetics and machine learning**, and environment – **6. Individual prediction of psychosis using environmental data and machine learning**). Additionally, two small studies were conducted in order to optimize the analysis of the neuroimaging and genetic data. One compared two pipelines for segmenting structural magnetic resonance neuroimages – the classical unified segmentation integrated in the Statistical Parametric Mapping 12 tool (SPM12) (Ashburner & Friston, 2005), a standard neuroimaging processing tool, and the more recent and advanced segmentation pipeline of the Computational Anatomy Toolbox (CAT12), an SPM12 add-on (**2. Comparing SPM12 with CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study**). The aim of this study within the context of the main project was to select the best brain image segmentation pipeline to process the sMRI data from the ARMS subjects. The other study reported the development of the eGenScore tool, a brain-based gene expression quantitative trait loci (eQTL) score tool. This tool allows the extraction of a score that aggregated single nucleotide polymorphisms (SNPs) that are statistically associated with a given gene expression. Moreover, this tool is brain specific, i.e., it computes eQTL scores for genes that are expression in the brain (**4. Creating a brain-based gene expression quantitative trait loci score (eGenScore) tool**). Within the scope of the main project, the eGenScore tool allows the extraction of a proxy for gene expression in the brain using a statistical method that otherwise would be invasive (i.e., by sampling brain tissue and which is currently done in post mortem studies). Finally, one additional objective, i.e. the train and test of a multimodality ML model, was set at the beginning of this project: a) only the

modalities showing a ML model performance, i.e. accuracy, statistically better than chance level, i.e. 50%, in predicting transition to psychosis from an ARMS would be included in a multimodal ML model; and b) at least two modalities would have to show a statistically significant individual performance in the transition to psychosis prediction. These criteria were not fulfilled due to the fact that only the genetics-based ML model showed a statistically significant accuracy in predicting transition to psychosis from an ARMS at better than chance level. Such criteria were chosen to be conservative given the already small sample size available for the training of the multimodal ML model, i.e. only 6 ARMS-T and 23 ARMS-NT (only this subset of subjects had data for the three data modalities, simultaneously). The decrease in sample size, remarkably impairs the prediction power of the model, i.e. its accuracy. Without previous evidence of the ability to predict transition to psychosis from an ARMS by modality supporting its integration in a multimodal ML model, negative results from this multimodal model would be highly difficult to explain, as they could be theoretically explained by the increase of noise in the model due to the inclusion of features that did show previous predictive ability or by the lack of predictive power due to the very small sample size.

In specific to the unimodal prediction of transition to psychosis form an ARMS, neuroimaging has been shown as the most promising modality, with several previous studies showing its value in predicting functional outcome (de Wit et al., 2017; Kambeitz-Ilankovic et al., 2016; Koutsouleris et al., 2018; Stefania Tognin et al., 2014) and transition to psychosis (Das et al., 2018; Koutsouleris et al., 2009, 2012; Koutsouleris, Riecher-Rössler, et al., 2015; Zarogianni et al., 2017, 2019) from an ARMS. I therefore tried to replicate the previous positive findings in a relatively larger and independent sample and using an updated ML pipeline. Furthermore, I explored, for the first time, the use of whole brain and regional white matter volume, regional cortical thickness, and surface-based regional brain gyrification, sulci depth and complexity indexes with ML to predict transition to psychosis. Moreover, to the best of my knowledge, genetics and environment have never been explored as predictors of the transition to psychosis from an ARMS. Therefore, I explored, for the first time, several types of genetic and environmental predictors, such as a) a schizophrenia polygenic risk score (PRS), which has been previously shown to distinguish FEP patients from healthy controls (Vassos et al., 2016); b) a set of SNPs that have been shown to be associated with schizophrenia in the most recent genome-wide meta-analysis study (Pardiñas

et al., 2018), c) several brain tissue-specific eQTL scores, extracted with the eGenScore tool, of a set of genes that have been identified as schizophrenia-associated in the same genome-wide meta-analysis study (Pardiñas et al., 2018), and d) a schizophrenia environmental risk score (ERS), which the proof of concept has been previously presented (Vassos et al., 2020). In the training/testing of all classification models, I carefully controlled for possible confounding effects such as age, sex and scanning acquisition protocols (in the case of the use of neuroimaging predictors) to avoid overoptimistic results. Furthermore, for each classification model I have run sample balancing and bootstrapping, i.e., a semi-random subsampling and class – ARMS subjects that have later transitioned to psychosis (ARMS-T) and ARMS subjects who have not (ARMS-NT) – balancing, for a generalizability measure of the classification model.

Overall, only genetics, in the form of a set of psychosis-associated SNPs, was shown to predict the transition to psychosis from an ARMS better than chance, albeit marginally. Interestingly, the two genetic composite score approaches explored in this project, i.e., the PRS and the eQTL score – both composed by an aggregation of SNPs – were not able to predict transition to psychosis. Furthermore, despite the large amount of evidence of the effect of exposure to environmental risk factors on the development of psychosis (measured essentially as the schizophrenia diagnosis *vs.* healthy controls) (Jaaro-Peled & Sawa, 2020; Schmitt et al., 2014), features drawn from this modality – ERS and individual risk factors – were not able to predict transition to psychosis with an accuracy better than chance level. This is a result that is, indeed, in line with the only report using environmental data with ML to classify schizophrenia (*vs.* healthy controls) (Antonucci et al., 2020). It showed that the environmental classifier was not able to classify schizophrenia better than chance level. Moreover, and unexpectedly, I could not replicate previous findings showing the usefulness of structural MRI in predicting transition to psychosis from an ARMS using ML. These results suggest that a) genetic data might be a promising predictor of transition to psychosis from an ARMS; b) environment or structural MRI data alone cannot predict transition to psychosis at better than chance-level; and, furthermore, c) the prediction value of structural MRI data from a prodromal stage of psychosis, stated by previous evidence, should be reconsidered.

Only two studies combining more than one modality (and at least two of them being structural neuroimaging, genetics, or environment) and ML for classifying schizophrenia patients (*vs.* healthy controls) at an individual models have been reported so far (Antonucci et al., 2020; Struyf et al., 2008). Both studies have reported increased accuracies in classifying schizophrenia when combining, in the same ML model, genetic and environmental data. These findings are, indeed, in line with the current understanding of the etiology of psychosis, i.e. that it arises from a gene-environment interplay (Zwicker et al., 2018). Furthermore, Antonucci and colleagues incorporated this interplay in their ML strategy by training a classifier with patients with schizophrenia *vs.* healthy controls based on neuropsychological assessments of subsamples generated based upon genetic and environmental stratification (Antonucci et al., 2020). Interestingly, and in line with this project's findings, when training a modality-specific classification model, i.e., one genetics- and one environment-based models, only the genetic-based classifier – which was trained with a set of risk scores for the top 116 SNPs that have been reported in the PGC2 (Ripke et al., 2014) – showed an accuracy statistically better than chance level (Antonucci et al., 2020). The environmental data alone was not able to classify the patients with schizophrenia (Antonucci et al., 2020).

7.1. Limitations and suggestions for future work

This project was limited by several factors. First, and foremost, the small size of the ARMS sample limits the performance of the classification models. Although the sample size needed to achieve a performance better than chance level has been informed by previous successful studies showing higher accuracies in predicting transition to psychosis from an ARMS (Das et al., 2018; Koutsouleris et al., 2009, 2012; Koutsouleris, Riecher-Rössler, et al., 2015; Zarogianni et al., 2017, 2019), this project's results demonstrated that results from studies using small sample sizes should be interpreted with caution. Indeed, this is a critical limitation when dealing with high dimensional data, such as neuroimaging and genetics – which I have used herein. Although I have taken measures to avoid overfitting and an overestimation of the classification models' performance such as artificially increasing the sampling through bootstrapping and employing cross-validation strategies, this might not be enough to overcome this limitation and, ultimately, I cannot assure that negative findings were due to lack of power to obtain a good performance or due to a true lack of association between the predictors and the transition to psychosis from an ARMS. Additionally, this is

one of the reasons why replication studies in independent datasets are very important in the machine learning literature – which was what motivated me to perform this study. As a final note, a power analysis for this study design would have been the most informative way to define the sample size needed to achieve an accuracy in predicting transition to psychosis from an ARMS better than chance level. However, this is not a trivial task in machine learning analysis and there is no established method to perform this analysis as there is for univariate analysis (for examples of studies exploring innovative ways of computing sample size for classification problems see (Dobbin & Simon, 2007; Figueroa et al., 2012)) and, therefore, it was not performed.

Second, in order to dilute possible confounding effects in the developed classification models I have restricted the samples used to train the models a) to be class balanced, i.e. with the same number of ARMS-T and ARMS-NT subjects; b) matched for age, sex, scanning acquisition protocols for neuroimaging data; c) to include subjects with European ancestry only for genetic data; and d) to limit the proportion of missing data for environment data. Although this artificially homogenizes the study sample thus avoiding the presence of statistical confounders, it deems the sample to be less representative of the ARMS population. Third, overall, the findings of this study are only valid to young help-seeking individuals, i.e. that are clinically screened for ARMS criteria, and whose ARMS diagnosis was based on having a schizotypal personality disorder or on the subject's familial-high risk coupled with functioning decline and on the CAARMS (Yung et al., 2005), which mainly evaluates positive symptoms.

Regarding future directions, a few aspects of this project could be improved, and some additional work is left for future studies. First, this study's findings need to be replicated in significantly larger samples which will be provided by multicenter projects, such as NAPLS 2 (Addington et al., 2012), PSYSCAN (Stefania Tognin et al., 2020), and PRONIA (http://pronia.eu) over the next years. Second, the following alternative ML strategies could be tested:

a) using a stacking ensemble ML approach to predict the clinical outcomes of ARMS subjects. In this method machine learning models would be trained in two layers: a bottom-level layer where a machine learning model per type of modality (i.e., modality specific model – e.g., neuroimaging, genetics or environment) or type of feature (i.e., feature specific model – e.g., cortical thickness, regional grey matter

volume, SNPs, eQTL scores, ERS, environmental risk factors) would be trained to predict the clinical outcome using the subjects' modality/feature-specific data; and a top-level layer, where the output of the bottom-level layer's models, i.e., the predicted clinical outcome, would be fed into a ML model and the final predicted clinical outcome would be outputted by this model. In this way, each modality contribution to the prediction would be weighted in a data-driven way, with higher chances of small contributions being detected by the ML algorithms. This would be an alternative to the criteria employed in this project of including in a multimodal ML approach only modalities showing, individually, a performance statistically better than chance level.

b) using a sample stratification approach. Following the etiological hypothesis of psychosis that it arises from a gene-environment interplay (Zwicker et al., 2018) and the recent report of Antonucci and colleagues (Antonucci et al., 2020), an alternative multimodal ML approach to predict the clinical outcome of ARMS would be to stratify the sample according to genetic and environmental data and then train a ML model using data from other modalities. In detail, a genetic- and environment-specific ML model would be trained and tested and only subjects with the set of genetic and environmental features yielding the best prediction scores would be selected to train a prediction model using data from other modalities (e.g., neuroimaging, or cognitive data). This approach had been recently shown to reliably distinguish patients with schizophrenia from healthy controls (Antonucci et al., 2020).

Third, alternative clinical outcomes would be interesting to explore based, for example, on the functional outcome of the ARMS subjects. This has been already explored in a few studies using neuroimaging (de Wit et al., 2017; Kambeitz-Ilankovic et al., 2016; Koutsouleris et al., 2018; Stefania Tognin et al., 2014). Indeed, from a clinical intervention point of view predicting reliably the functional outcome of the prodromal patients, i.e. making prediction based on symptoms instead of clinical diagnosis, may open new avenues for further clinical trials regarding intervention and treatment effectiveness in this prodromal stage of psychosis (Millan et al., 2016).

7.2. Contributions to the current state of knowledge

Several contributions to the current state of knowledge were made under the scope of this project. First, it is the first published study exploring individual prediction of psychosis from an ARMS using exclusively quantitative and multimodal data (i.e., as predictors). Overall, I found that neuroimaging, contrary to what has been previously reported, could not predict transition to psychosis from an ARMS. I have employed several machine learning strategies aiming to replicate the previous positive findings but could not replicate them. This strongly suggests that previous evidence should be carefully reconsidered. Moreover, and to the best of my knowledge, I explored for the first time the value of genetics and environment in the prediction of psychosis already from a prodromal stage. I found that when using an a priori-defined set of single nucleotide polymorphisms (SNPs) based on their statistical association with schizophrenia diagnosis (*vs.* healthy controls) I could predict transition to psychosis marginally better than chance level. Although this result needs to be replicated in larger samples, it already suggests that some genetic features might be promising predictors of the transition to psychosis. Additionally, environmental data could not predict transition to psychosis from an ARMS. In summary, the methodological approach following within this project serves as a proof of concept to be replicated and improved in further studies, particularly when larger ARMS samples become available (e.g., with the end of data collection in current multicenter projects).

Second, two parallel and smaller studies were also conducted during this project. The first one was run in the context of the structural magnetic resonance imaging (sMRI) data processing. In summary I have compared two segmentation pipelines of brain sMRI – the SPM12 toolbox, and a SPM12 add-on, the CAT12 toolbox – by studying their impact on the effect of age and Alzheimer's disease (AD) diagnosis on brain volume and the accuracy of brain volume detecting the diagnosis of AD. Overall, I found that the choice of pipeline modulated the effect of age on brain volume measures and of diagnosis on hippocampi gray matter volumes, and this last effect was dependent on the scanner field strength. Moreover, I found that pipeline had no impact on the accuracy of any brain volume measure detecting AD diagnosis. These findings indicate that other studies should take these pipeline effects on age and AD diagnosis, into account, for improved comparability in previous literature. Additionally, I encourage future studies to use CAT12 as this is a more advanced and computationally efficient brain segmentation tool. Furthermore, under the context of the main project this study helped on our decision of using CAT12 tool to preprocess the sMRI

images of the ARMs subjects. As a final note this study was already published (V. Tavares et al., 2020).

The second parallel study comprised the development of the eGenScore, a tool that extracts expression quantitative trait loci (eQTL) scores from genome-wide genotyped data (i.e., SNPs). eGenScore is brain and gene specific, i.e., it extracts an eQTL score per gene and per brain tissue. At the time of the start of this parallel study no method for the estimation of gene expression level in the brain was available and access to this data is extremely invasive and currently done only in post-mortem studies. Therefore, there was a clear need for the development of this kind of tools. Under the context of the main project this tool allowed the extraction of potentially neurobiologically more meaningful genetic features – i.e., eQTL scores for genes that are mainly expressed in the brain – and that can be though as proxies for actual gene expression levels. Furthermore, PrediXcan, a tool similar to the eGenScore, has been developed in parallel to mine (Gamazon et al., 2015; Huckins et al., 2019). The main difference between the two tools are a) the method they use to estimate the gene expression levels and b) the databases they use to train these gene expression prediction models – I use a brain-specific database (BrainEAC) and PrediXcan uses a generic database (GTEx). Therefore, the comparison of the performance of the gene expression estimation, i.e. the eQTL scores, across methods (i.e. BrainEAC *vs.* PrediXcan) as well as across training datasets (BrainEAC *vs.* GTEx) would be important to explore and has been explored after the formulation of this thesis (Vânia Tavares et al., 2021).

References

Addington, J., Cadenhead, K. S., Cornblatt, B. A., Mathalon, D. H., McGlashan, T. H., Perkins, D. O., Seidman, L. J., Tsuang, M. T., Walker, E. F., Woods, S. W., Addington, J. A., & Cannon, T. D. (2012). North American Prodrome Longitudinal Study (NAPLS 2): Overview and recruitment. *Schizophrenia Research*, *142*(1–3), 77–82. https://doi.org/10.1016/j.schres.2012.09.012

Aguet, F., Barbeira, A. N., Bonazzola, R., Jo, B., Kasela, S., Liang, Y., Parsana, P., Aguet, F., Battle, A., Brown, A., Castel, S. E., Engelhardt, B. E., Hormozdiari, F., Im, H. K., Kim-Hellmuth, S., Oliva, M., Stranger, B. E., & Wen, X. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776

Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y. S., Parsana, P., Segrè, A. V., Strober, B. J., Zappala, Z., Cummings, B. B., Gelfand, E. T., Hadley, K., Huang, K. H., Lek, M., Li, X., Nedzel, J. L., … Zhu, J. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. https://doi.org/10.1038/nature24277

Aguiar-Pulido, V., Seoane, J. A., Rabuñal, J. R., Dorado, J., Pazos, A., & Munteanu, C. R. (2010). Machine learning techniques for single nucleotide polymorphism - disease classification models in schizophrenia. *Molecules*, *15*(7), 4875–4889. https://doi.org/10.3390/molecules15074875

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR)* (Vol. 1). American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890423349

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V). In *American Psychiatric Association* (5th ed., Vol. 5). American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Andreou, C., & Borgwardt, S. (2020). Structural and functional imaging markers for susceptibility to psychosis. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-020-0679-7

Antonucci, L. A., Pergola, G., Pigoni, A., Dwyer, D., Kambeitz-Ilankovic, L., Penzel, N., Romano, R., Gelao, B., Torretta, S., Rampino, A., Trojano, M., Caforio, G., Falkai, P., Blasi, G., Koutsouleris, N., & Bertolino, A. (2020). A Pattern of Cognitive Deficits Stratified for Genetic and Environmental Risk Reliably Classifies Patients With Schizophrenia From Healthy Control Subjects. *Biological Psychiatry*, *87*(8), 697–707. https://doi.org/10.1016/j.biopsych.2019.11.007

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, *145*, 137–165. https://doi.org/10.1016/j.neuroimage.2016.02.079

Ardlie, K. G., Deluca, D. S., Segre, A. V, Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M., Ward, L. D., Kheradpour, P., Iriarte, B., Meng, Y., Palmer, C. D., Esko, T., Winckler, W., Hirschhorn, J. N., Kellis, M., … Dermitzakis, E. T. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113. https://doi.org/10.1016/j.neuroimage.2007.07.007

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851. https://doi.org/https://doi.org/10.1016/j.neuroimage.2005.02.018

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Barbeira, A. N., Bonazzola, R., Gamazon, E. R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F., Liu, B., Rao, A., Hamel, A. R., Pividori, M. D., Aguet, F., Bastarache, L., Jordan, D. M., Verbanck, M., Do, R., … Im, H. K. (2021). Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biology*, *22*(1), 49. https://doi.org/10.1186/s13059-020-02252-4

Beards, S., Gayer-Anderson, C., Borges, S., Dewey, M. E., Fisher, H. L., & Morgan, C. (2013). Life events and psychosis: A review and meta-analysis. *Schizophrenia Bulletin*, *39*(4), 740–747. https://doi.org/10.1093/schbul/sbt065

Benetti, S., Pettersson-Yeo, W., Hutton, C., Catani, M., Williams, S. C., Allen, P., Kambeitz-Ilankovic, L. M., McGuire, P., & Mechelli, A. (2013). Elucidating neuroanatomical alterations in the at risk mental state and first episode psychosis: A combined voxel-based morphometry and voxel-based cortical thickness study. *Schizophrenia Research*, *150*(2–3), 505–511. https://doi.org/10.1016/j.schres.2013.08.030

Bifulco, A., Bernazzani, O., Moran, P. M., & Jacobs, C. (2005). The childhood experience of care and abuse questionnaire (CECA.Q): Validation in a community series. *British Journal of Clinical Psychology*, *44*(4), 563–581. https://doi.org/10.1348/014466505X35344

Bois, C., Whalley, H. C., McIntosh, A. M., & Lawrie, S. M. (2015). Structural magnetic resonance imaging markers of susceptibility and transition to schizophrenia: A review of familial and clinical high risk population studies. *Journal of Psychopharmacology*, *29*(2), 144–154. https://doi.org/10.1177/0269881114541015

Bonnici, H. M., William, T., Moorhead, J., Stanfield, A. C., Harris, J. M., Owens, D. G., Johnstone, E. C., & Lawrie, S. M. (2007). Pre-frontal lobe gyrification index in schizophrenia, mental retardation and comorbid groups: An automated study. *NeuroImage*, *35*(2), 648–654. https://doi.org/10.1016/j.neuroimage.2006.11.031

Borges, S., Gayer-Anderson, C., & Mondelli, V. (2013). A systematic review of the activity of the hypothalamic-pituitary-adrenal axis in first episode psychosis. *Psychoneuroendocrinology*, *38*(5), 603–611. https://doi.org/10.1016/j.psyneuen.2012.12.025

Borgwardt, S. J., McGuire, P. K., Aston, J., Gschwandtner, U., Pflüger, M. O., Stieglitz, R. D., Radue, E. W., & Riecher-Rössler, A. (2008). Reductions in frontal, temporal and parietal volume associated with the onset of psychosis. *Schizophrenia Research*, *106*(2–3), 108–114. https://doi.org/10.1016/j.schres.2008.08.007

Bourque, F., Van Der Ven, E., & Malla, A. (2011). A meta-analysis of the risk for

psychotic disorders among first- and second-generation immigrants. *Psychological Medicine*, *41*(5), 897–910. https://doi.org/10.1017/S0033291710001406

Bramon, E., Pirinen, M., Strange, A., Lin, K., Freeman, C., Bellenguez, C., Su, Z., Band, G., Pearson, R., Jankowski, J., Langford, C., Deloukas, P., Hunt, S., Gray, E., Dronov, S., Potter, S. C., Tashakkori-Ghanbaria, A., Viswanathan, A. C., Bumpstead, S. J., … Spencer, C. C. A. (2014). A genome-wide association analysis of a broad psychosis phenotype identifies three loci for further investigation. *Biological Psychiatry*, *75*(5), 386–397. https://doi.org/10.1016/j.biopsych.2013.03.033

Breier, A., Buchanan, R. W., Elkashef, A., Munson, R. C., Kirkpatrick, B., & Gellad, F. (1992). Brain Morphology and Schizophrenia: A Magnetic Resonance Imaging Study of Limbic, Prefrontal Cortex, and Caudate Structures. *Archives of General Psychiatry*, *49*(12), 921–926. https://doi.org/10.1001/archpsyc.1992.01820120009003

Broome, M. R., Woolley, J. B., Johns, L. C., Valmaggia, L. R., Tabraham, P., Gafoor, R., Bramon, E., & McGuire, P. K. (2005). Outreach and support in south London (OASIS): implementation of a clinical service for prodromal psychosis and the at risk mental state. *European Psychiatry*, *20*(5), 372–378. https://doi.org/10.1016/j.eurpsy.2005.03.001

Brown, A. S. (2011). The environment and susceptibility to schizophrenia. *Progress in Neurobiology*, *93*(1), 23–58. https://doi.org/10.1016/j.pneurobio.2010.09.003

Buchert, R., Lange, C., Suppa, P., Apostolova, I., Spies, L., Teipel, S., Dubois, B., Hampel, H., & Grothe, M. J. (2018). Magnetic resonance imaging-based hippocampus volume for prediction of dementia in mild cognitive impairment: Why does the measurement method matter so little? *Alzheimer's and Dementia*, *14*(7), 976–978. https://doi.org/10.1016/j.jalz.2018.03.006

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*, 121–167.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234. https://doi.org/10.1038/nmeth.4642

Cachia, A., Paillère-Martinot, M. L., Galinowski, A., Januel, D., de Beaurepaire, R., Bellivier, F., Artiges, E., Andoh, J., Bartrés-Faz, D., Duchesnay, E., Rivière, D., Plaze, M., Mangin, J. F., & Martinot, J. L. (2008). Cortical folding abnormalities in schizophrenia patients with resistant auditory hallucinations. *NeuroImage*, *39*(3), 927–935. https://doi.org/10.1016/j.neuroimage.2007.08.049

Cannon, M., Jones, P. B., & Murray, R. M. (2002). Obstetric Complications and Schizophrenia: Historical and Meta-Analytic Review. *American Journal of Psychiatry*, *159*(7), 1080–1092. https://doi.org/10.1176/appi.ajp.159.7.1080

Cannon, T. D., Chung, Y., He, G., Sun, D., Jacobson, A., van Erp, T. G. M., McEwen, S., Addington, J., Bearden, C. E., Cadenhead, K., Cornblatt, B., Mathalon, D. H., McGlashan, T., Perkins, D., Jeffries, C., Seidman, L. J., Tsuang, M., Walker, E., Woods, S. W., & Heinssen, R. (2015). Progressive Reduction in Cortical Thickness as Psychosis Develops: A Multisite Longitudinal Neuroimaging Study of Youth at Elevated Clinical Risk. *Biological Psychiatry*, *77*(2), 147–157. https://doi.org/10.1016/j.biopsych.2014.05.023

Cardno, A. G., Marshall, E. J., Coid, B., Macdonald, A. M., Ribchester, T. R., Davies, N.

J., Venturi, P., Jones, L. A., Lewis, S. W., Sham, P. C., Gottesman, I. I., Farmer, A. E., McGuffin, P., Reveley, A. M., & Murray, R. M. (1999). Heritability Estimates for Psychotic Disorders. *Archives of General Psychiatry*, *56*(2), 162. https://doi.org/10.1001/archpsyc.56.2.162

Castellani, U., Rossato, E., Murino, V., Bellani, M., Rambaldelli, G., Perlini, C., Tomelleri, L., Tansella, M., & Brambilla, P. (2012). Classification of schizophrenia using feature-based morphometry. *Journal of Neural Transmission*, *119*(3), 395–404. https://doi.org/10.1007/s00702-011-0693-7

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 1–16. https://doi.org/10.1186/s13742-015-0047-8

Chen, J., Wu, J., Mize, T., Shui, D., & Chen, X. (2018). Prediction of Schizophrenia Diagnosis by Integration of Genetically Correlated Conditions and Traits. *Journal of Neuroimmune Pharmacology*, *13*(4), 532–540. https://doi.org/10.1007/s11481-018-9811-8

Chung, Y., & Cannon, T. D. (2015). Brain imaging during the transition from psychosis prodrome to Schizophrenia. *Journal of Nervous and Mental Disease*, *203*(5), 336–341. https://doi.org/10.1097/NMD.0000000000000286

Chung, Y., Haut, K. M., He, G., van Erp, T. G. M., McEwen, S., Addington, J., Bearden, C. E., Cadenhead, K., Cornblatt, B., Mathalon, D. H., McGlashan, T., Perkins, D., Seidman, L. J., Tsuang, M., Walker, E., Woods, S. W., & Cannon, T. D. (2017). Ventricular enlargement and progressive reduction of cortical gray matter are linked in prodromal youth who develop psychosis. *Schizophrenia Research*, *189*, 169–174. https://doi.org/10.1016/j.schres.2017.02.014

Chung, Y., Jacobson, A., He, G., van Erp, T. G. M., McEwen, S., Addington, J., Bearden, C. E., Cadenhead, K., Cornblatt, B., Mathalon, D. H., McGlashan, T., Perkins, D., Seidman, L. J., Tsuang, M., Walker, E., Woods, S. W., Heinssen, R., & Cannon, T. D. (2015). Prodromal Symptom Severity Predicts Accelerated Gray Matter Reduction and Third Ventricle Expansion among Clinically High-Risk Youth Developing Psychotic Disorders. *Molecular Neuropsychiatry*, *1*(1), 13–22. https://doi.org/10.1159/000371887

Correll, C. U., & Schooler, N. R. (2020). Negative symptoms in schizophrenia: A review and clinical guide for recognition, assessment, and treatment. *Neuropsychiatric Disease and Treatment*, *16*, 519–534. https://doi.org/10.2147/NDT.S225643

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018

Crow, J. F. (1997). The high spontaneous mutation rate: Is it a health risk? *Proceedings of the National Academy of Sciences of the United States of America*, *94*(16), 8380–8386. https://doi.org/10.1073/pnas.94.16.8380

Csernansky, J. G., Schindler, M. K., Splinter, N. R., Wang, L., Gado, M., Selemon, L. D., Rastogi-Cruz, D., Posener, J. A., Thompson, P. A., & Miller, M. I. (2004). Abnormalities of Thalamic Volume and Shape in Schizophrenia. *American Journal of Psychiatry*, *161*(5), 896–902. https://doi.org/10.1176/appi.ajp.161.5.896

Curtis, D. (2018). Polygenic risk score for schizophrenia is more strongly associated with

ancestry than with schizophrenia. *Psychiatric Genetics*, *28*(5), 85–89. https://doi.org/10.1097/YPG.0000000000000206

Dahnke, R., Yotter, R. A., & Gaser, C. (2013). Cortical thickness and central surface estimation. *NeuroImage*, *65*, 336–348. https://doi.org/10.1016/j.neuroimage.2012.09.050

Dahnke, R., Ziegler, G., Grosskreutz, J., & Gaser, C. (2013). *Retrospective Quality Assurance of MR Images*. https://doi.org/10.13140/RG.2.2.25494.91200

Das, T., Borgwardt, S., Hauke, D. J., Harrisberger, F., Lang, U. E., Riecher-Rössler, A., Palaniyappan, L., & Schmidt, A. (2018). Disorganized gyrification network properties during the transition to psychosis. *JAMA Psychiatry*, *75*(6), 613–622. https://doi.org/10.1001/jamapsychiatry.2018.0391

Davatzikos, C., Shen, D., Gur, R. C., Wu, X., Liu, D., Fan, Y., Hughett, P., Turetsky, B. I., & Gur, R. E. (2005). Whole-Brain Morphometric Study of Schizophrenia Revealing a Spatially Complex Set of Focal Abnormalities. *Archives of General Psychiatry*, *62*(11), 1218. https://doi.org/10.1001/archpsyc.62.11.1218

Davies, G. J., Welham, J., Torrey, E. F., & McGrath, J. (2003). Season of birth effect and latitude: A systematic review and meta-analysis of Northern hemisphere schizophrenia studies. *Schizophrenia Research*, *41*(1), 62. https://doi.org/10.1016/s0920-9964(00)90437-7

Dazzan, P., Soulsby, B., Mechelli, A., Wood, S. J., Velakoulis, D., Phillips, L. J., Yung, A. R., Chitnis, X., Lin, A., Murray, R. M., McGorry, P. D., McGuire, P. K., & Pantelis, C. (2012). Volumetric abnormalities predating the onset of schizophrenia and affective psychoses: An MRI study in subjects at ultrahigh risk of psychosis. *Schizophrenia Bulletin*, *38*(5), 1083–1091. https://doi.org/10.1093/schbul/sbr035

de Wit, S., Ziermans, T. B., Nieuwenhuis, M., Schothorst, P. F., van Engeland, H., Kahn, R. S., Durston, S., & Schnack, H. G. (2017). Individual prediction of long-term outcome in adolescents at ultra-high risk for psychosis: Applying machine learning techniques to brain imaging data. *Human Brain Mapping*, *38*(2), 704–714. https://doi.org/10.1002/hbm.23410

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837–845.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Despotović, I., Goossens, B., & Philips, W. (2015). MRI segmentation of the human brain: Challenges, methods, and applications. *Computational and Mathematical Methods in Medicine*, *2015*. https://doi.org/10.1155/2015/450341

Dobbin, K. K., & Simon, R. M. (2007). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics*, *8*(1), 101–117. https://doi.org/10.1093/biostatistics/kxj036

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*, *21*(16), 3439–3440. https://doi.org/10.1093/bioinformatics/bti525

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Falahati, F., Ferreira, D., Soininen, H., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Lovestone, S., Eriksdotter, M., Wahlund, L. O., Simmons, A., & Westman, E. (2016). The Effect of Age Correction on Multivariate Classification in Alzheimer's Disease, with a Focus on the Characteristics of Incorrectly and Correctly Classified Subjects. *Brain Topography*, *29*(2), 296–307. https://doi.org/10.1007/s10548-015-0455-1

Falkai, P., Honer, W. G., Kamer, T., Dustert, S., Vogeley, K., Schneider-Axmann, T., Dani, I., Wagner, M., Rietschel, M., Müller, D. J., Schulze, T. G., Gaebel, W., Cordes, J., Schönell, H., Schild, H. H., Block, W., Träber, F., Steinmetz, H., Maier, W., & Tepest, R. (2007). Disturbed frontal gyrification within families affected with schizophrenia. *Journal of Psychiatric Research*, *41*(10), 805–813. https://doi.org/10.1016/j.jpsychires.2006.07.018

Fan, Y., Shen, D., & Davatzikos, C. (2005). Classification of structural images via high-dimensional image warping, robust feature extraction, and SVM. *Medical Image Computing and Computer-Assisted Intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, *8*(Pt 1), 1–8. https://doi.org/10.1007/11566465_1

Fan, Y., Shen, D., Gur, R. C., Gur, R. E., & Davatzikos, C. (2007). COMPARE: Classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging*, *26*(1), 93–105. https://doi.org/10.1109/TMI.2006.886812

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*, 8. https://doi.org/10.1186/1472-6947-12-8

Fisher, R. A. (1992). *Statistical Methods for Research Workers* (pp. 66–70). https://doi.org/10.1007/978-1-4612-4380-9_6

Fjell, A. M., & Walhovd, K. B. (2010). Structural brain changes in aging: courses, causes and cognitive consequences. *Reviews in the Neurosciences*, *21*(3), 187–221. http://www.ncbi.nlm.nih.gov/pubmed/20879692

Foley, C., Corvin, A., & Nakagome, S. (2017). Genetics of Schizophrenia: Ready to Translate? *Current Psychiatry Reports*, *19*(9), 1–9. https://doi.org/10.1007/s11920-017-0807-5

Fotenos, A., Snyder, A., Girton, L., Morris, J., & Buckner, R. (2005). Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. *Neurology*, *64*(6), 1032–1039. https://doi.org/10.1212/01.WNL.0000154530.72969.11

Fraser, M. A., Shaw, M. E., & Cherbuin, N. (2015). A systematic review and meta-analysis of longitudinal hippocampal atrophy in healthy human ageing. *NeuroImage*, *112*,

364–374. https://doi.org/10.1016/j.neuroimage.2015.03.035

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., Ruderfer, D. M., Oh, E. C., Topol, A., Shah, H. R., Klei, L. L., Kramer, R., Pinto, D., Gümüş, Z. H., Cicek, A. E., Dang, K. K., Browne, A., Lu, C., Xie, L., … Sklar, P. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience*, *19*(11), 1442–1453. https://doi.org/10.1038/nn.4399

Fusar-Poli, P., Broome, M. R., Woolley, J. B., Johns, L. C., Tabraham, P., Bramon, E., Valmaggia, L., Williams, S. C., & McGuire, P. (2011). Altered brain function directly related to structural abnormalities in people at ultra high risk of psychosis: Longitudinal VBM-fMRI study. *Journal of Psychiatric Research*, *45*(2), 190–198. https://doi.org/10.1016/j.jpsychires.2010.05.012

Fusar-Poli, Paolo, Bechdolf, A., Taylor, M. J., Bonoldi, I., Carpenter, W. T., Yung, A. R., & McGuire, P. (2013). At risk for schizophrenic or affective psychoses? A meta-analysis of DSM/ICD diagnostic outcomes in individuals at high clinical risk. *Schizophrenia Bulletin*, *39*(4), 923–932. https://doi.org/10.1093/schbul/sbs060

Fusar-Poli, Paolo, Borgwardt, S., Bechdolf, A., Addington, J., Riecher-Rössler, A., Schultze-Lutter, F., Keshavan, M., Wood, S., Ruhrmann, S., Seidman, L. J., Valmaggia, L., Cannon, T., Velthorst, E., De Haan, L., Cornblatt, B., Bonoldi, I., Birchwood, M., McGlashan, T., Carpenter, W., … Yung, A. (2013). The Psychosis High-Risk State. *JAMA Psychiatry*, *70*(1), 107. https://doi.org/10.1001/jamapsychiatry.2013.269

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., & Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*(9), 1091–1098. https://doi.org/10.1038/ng.3367

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational and Behavioral Statistics*, *18*(3), 271–279. https://doi.org/10.2307/1165136

Gifford, G., Crossley, N., Fusar-Poli, P., Schnack, H. G., Kahn, R. S., Koutsouleris, N., Cannon, T. D., & McGuire, P. (2015). Using neuroimaging to help predict the onset of psychosis. *NeuroImage*, *145*, 209–217. https://doi.org/10.1016/j.neuroimage.2016.03.075

Goghari, V. M., Rehm, K., Carter, C. S., & MacDonald, A. W. (2007). Sulcal thickness as a vulnerability indicator for schizophrenia. *British Journal of Psychiatry*, *191*(3), 229–233. https://doi.org/10.1192/bjp.bp.106.034595

Gratten, J., Wray, N. R., Keller, M. C., & Visscher, P. M. (2014). Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nature Neuroscience*, *17*(6), 782–790. https://doi.org/10.1038/nn.3708

Greenstein, D., Malley, J. D., Weisinger, B., Clasen, L., & Gogtay, N. (2012). Using Multivariate Machine Learning Methods and Structural MRI to Classify Childhood Onset Schizophrenia and Healthy Controls. *Frontiers in Psychiatry*, *3*(June), 1–12. https://doi.org/10.3389/fpsyt.2012.00053

Guloksuz, S., Rutten, B. P. F., Pries, L. K., Ten Have, M., De Graaf, R., Van Dorsselaer, S., Klingenberg, B., Van Os, J., & Ioannidis, J. P. A. (2018). The complexities of evaluating the exposome in psychiatry: A data-driven illustration of challenges and some propositions for amendments. *Schizophrenia Bulletin*, *44*(6), 1175–1179. https://doi.org/10.1093/schbul/sby118

Gurillo, P., Jauhar, S., Murray, R. M., & MacCabe, J. H. (2015). Does tobacco use cause psychosis? Systematic review and meta-analysis. *The Lancet Psychiatry*, *2*(8), 718–725. https://doi.org/10.1016/S2215-0366(15)00152-2

Haijma, S. V., Van Haren, N., Cahn, W., Koolschijn, P. C. M. P., Hulshoff Pol, H. E., & Kahn, R. S. (2013). Brain volumes in schizophrenia: A meta-analysis in over 18 000 subjects. *Schizophrenia Bulletin*, *39*(5), 1129–1138. https://doi.org/10.1093/schbul/sbs118

Hammers, A., Allom, R., Koepp, M. J., Free, S. L., Myers, R., Lemieux, L., Mitchell, T. N., Brooks, D. J., & Duncan, J. S. (2003). Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human Brain Mapping*, *19*(4), 224–247. https://doi.org/10.1002/hbm.10123

Harris, J. M., Moorhead, T. W. J., Miller, P., McIntosh, A. M., Bonnici, H. M., Owens, D. G. C., Johnstone, E. C., & Lawrie, S. M. (2007). Increased Prefrontal Gyrification in a Large High-Risk Cohort Characterizes Those Who Develop Schizophrenia and Reflects Abnormal Prefrontal Development. *Biological Psychiatry*, *62*(7), 722–729. https://doi.org/10.1016/j.biopsych.2006.11.027

Harris, J. M., Yates, S., Miller, P., Best, J. J. K., Johnstone, E. C., & Lawrie, S. M. (2004). Gyrification in first-episode schizophrenia: A morphometric study. *Biological Psychiatry*, *55*(2), 141–147. https://doi.org/10.1016/S0006-3223(03)00789-3

Haukvik, U. K., Tamnes, C. K., Söderman, E., & Agartz, I. (2018). Neuroimaging hippocampal subfields in schizophrenia and bipolar disorder: A systematic review and meta-analysis. *Journal of Psychiatric Research*, *104*(July), 217–226. https://doi.org/10.1016/j.jpsychires.2018.08.012

Hellwege, J. N., Keaton, J. M., Giri, A., Gao, X., Velez Edwards, D. R., & Edwards, T. L. (2017). Population Stratification in Genetic Association Studies. *Current Protocols in Human Genetics*, *95*(October), 1.22.1-1.22.23. https://doi.org/10.1002/cphg.48

Hernandez, L. M., Kim, M., Hoftman, G. D., Haney, J. R., de la Torre-Ubieta, L., Pasaniuc, B., & Gandal, M. J. (2020). Transcriptomic Insight Into the Polygenic Mechanisms Underlying Psychiatric Disorders. *Biological Psychiatry*, *5*, 1–11. https://doi.org/10.1016/j.biopsych.2020.06.005

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(23), 9362–9367. https://doi.org/10.1073/pnas.0903103106

Ho, B.-C., Andreasen, N. C., Nopoulos, P., Arndt, S., Magnotta, V., & Flaum, M. (2003). Progressive Structural Brain Abnormalities and Their Relationship to Clinical Outcome. *Archives of General Psychiatry*, *60*(6), 585. https://doi.org/10.1001/archpsyc.60.6.585

Hoffman, G. E., Bendl, J., Voloudakis, G., Montgomery, K. S., Sloofman, L., Wang, Y. C., Shah, H. R., Hauberg, M. E., Johnson, J. S., Girdhar, K., Song, L., Fullard, J. F., Kramer, R., Hahn, C. G., Gur, R., Marenco, S., Lipska, B. K., Lewis, D. A., Haroutunian, V., … Roussos, P. (2019). CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. *Scientific Data*, *6*(1), 180. https://doi.org/10.1038/s41597-019-0183-6

Hoopes, L. (2008). Introduction to the gene expression and regulation topic room. *Nature Education*, *1*(1), 160.

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, *5*(6), e1000529. https://doi.org/10.1371/journal.pgen.1000529

Hubbard, A. E., Ahern, J., Fleischer, N. L., Laan, M. Van Der, Lippman, S. A., Jewell, N., Bruckner, T., & Satariano, W. A. (2010). To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, *21*(4), 467–474. https://doi.org/10.1097/EDE.0b013e3181caeb90

Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, *47*(1), 64–79. https://doi.org/10.1198/004017004000000563

Hubert, M., Rousseeuw, P., & Verdonck, T. (2009). Robust PCA for skewed data and its outlier map. *Computational Statistics & Data Analysis*, *53*(6), 2264–2274. https://doi.org/10.1016/j.csda.2008.05.027

Huckins, L. M., Dobbyn, A., Ruderfer, D. M., Hoffman, G., Wang, W., Pardiñas, A. F., Rajagopal, V. M., Als, T. D., T. Nguyen, H., Girdhar, K., Boocock, J., Roussos, P., Fromer, M., Kramer, R., Domenici, E., Gamazon, E. R., Purcell, S., Johnson, J. S., Shah, H. R., … Stahl, E. A. (2019). Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nature Genetics*, *51*(4), 659–674. https://doi.org/10.1038/s41588-019-0364-4

Hulshoff Pol, H. E., Schnack, H. G., Bertens, M. G. B. C., Van Haren, N. E. M., Van der Tweel, I., Staal, W. G., Baaré, W. F. C., & Kahn, R. S. (2002). Volume changes in gray matter in patients with schizophrenia. *American Journal of Psychiatry*, *159*(2), 244–250. https://doi.org/10.1176/appi.ajp.159.2.244

Hunter, A., Murray, R., Asher, L., & Leonardi-Bee, J. (2018). The Effects of Tobacco Smoking, and Prenatal Tobacco Smoke Exposure, on Risk of Schizophrenia: A Systematic Review and Meta-Analysis. *Nicotine & Tobacco Research*, *August*, 1–8. https://doi.org/10.1093/ntr/nty160

Iwabuchi, S. J., Liddle, P. F., & Palaniyappan, L. (2013). Clinical Utility of Machine-Learning Approaches in Schizophrenia: Improving Diagnostic Confidence for Translational Neuroimaging. *Frontiers in Psychiatry*, *4*(August), 1–9. https://doi.org/10.3389/fpsyt.2013.00095

Jaaro-Peled, H., & Sawa, A. (2020). Neurodevelopmental Factors in Schizophrenia. *Psychiatric Clinics of North America*, *43*(2), 263–274. https://doi.org/10.1016/j.psc.2020.02.010

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should

multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. *BMC Medical Research Methodology*, *17*(1), 1–10. https://doi.org/10.1186/s12874-017-0442-1

Janousova, E., Schwarz, D., & Kasparek, T. (2015). Combining various types of classifiers and features extracted from magnetic resonance imaging data in schizophrenia recognition. *Psychiatry Research - Neuroimaging*, *232*(3), 237–249. https://doi.org/10.1016/j.pscychresns.2015.03.004

Jou, R. J., Hardan, A. Y., & Keshavan, M. S. (2005). Reduced cortical folding in individuals at high risk for schizophrenia: A pilot study. *Schizophrenia Research*, *75*(2–3), 309–313. https://doi.org/10.1016/j.schres.2004.11.008

Jung, S., Lee, A., Bang, M., & Lee, S. H. (2019). Gray matter abnormalities in language processing areas and their associations with verbal ability and positive symptoms in first-episode patients with schizophrenia spectrum psychosis. *NeuroImage: Clinical*, *24*(June), 102022. https://doi.org/10.1016/j.nicl.2019.102022

Kahn, R. S., Sommer, I. E., Murray, R. M., Meyer-Lindenberg, A., Weinberger, D. R., Cannon, T. D., O'Donovan, M., Correll, C. U., Kane, J. M., Van Os, J., & Insel, T. R. (2015). Schizophrenia. *Nature Reviews Disease Primers*, *1*(November). https://doi.org/10.1038/nrdp.2015.67

Kambeitz-Ilankovic, L., Meisenzahl, E. M., Cabral, C., von Saldern, S., Kambeitz, J., Falkai, P., Möller, H. J., Reiser, M., & Koutsouleris, N. (2016). Prediction of outcome in the psychosis prodrome using neuroanatomical pattern classification. *Schizophrenia Research*, *173*(3), 159–165. https://doi.org/10.1016/j.schres.2015.03.005

Kambeitz, J., Cabral, C., Sacchet, M. D., Gotlib, I. H., Zahn, R., Serpa, M. H., Walter, M., Falkai, P., & Koutsouleris, N. (2018). Reply to: Sample Size, Model Robustness, and Classification Accuracy in Diagnostic Multivariate Neuroimaging Analyses. *Biological Psychiatry*, *84*(11), e83—e84. https://doi.org/10.1016/j.biopsych.2018.01.023

Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., Falkai, P., & Koutsouleris, N. (2015). Detecting neuroimaging biomarkers for schizophrenia: A meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology*, *40*(7), 1742–1751. https://doi.org/10.1038/npp.2015.22

Karageorgiou, E., Schulz, S. C., Gollub, R. L., Andreasen, N. C., Ho, B. C., Lauriello, J., Calhoun, V. D., Bockholt, H. J., Sponheim, S. R., & Georgopoulos, A. P. (2011). Neuropsychological testing and structural magnetic resonance imaging as diagnostic biomarkers early in the course of schizophrenia and related psychoses. *Neuroinformatics*, *9*(4), 321–333. https://doi.org/10.1007/s12021-010-9094-6

Kasai, K., Shenton, M. E., Salisbury, D. F., Hirayasu, Y., Onitsuha, T., Spencer, M. H., Yurgelun-Todd, D. A., Kikinis, R., Jolesz, F. A., & McCarley, R. W. (2003). Progressive decrease of left Heschl gyrus and planum temporale gray matter volume in first-episode schizophrenia: A longitudinal magnetic resonance imaging study. *Archives of General Psychiatry*, *60*(8), 766–775. https://doi.org/10.1001/archpsyc.60.8.766

Kasem, E., Kurihara, T., & Tabuchi, K. (2018). Neurexins and neuropsychiatric disorders.

Neuroscience Research, *127*, 53–60. https://doi.org/10.1016/j.neures.2017.10.012

Kasparek, T., Thomaz, C. E., Sato, J. R., Schwarz, D., Janousova, E., Marecek, R., Prikryl, R., Vanicek, J., Fujita, A., & Ceskova, E. (2011). Maximum-uncertainty linear discrimination analysis of first-episode schizophrenia subjects. *Psychiatry Research - Neuroimaging*, *191*(3), 174–181. https://doi.org/10.1016/j.pscychresns.2010.09.016

Kawasaki, Y., Suzuki, M., Kherif, F., Takahashi, T., Zhou, S. Y., Nakamura, K., Matsui, M., Sumiyoshi, T., Seto, H., & Kurachi, M. (2007). Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. *NeuroImage*, *34*(1), 235–242. https://doi.org/10.1016/j.neuroimage.2006.08.018

Keshavan, M. S., & Hogarty, G. E. (1999). Brain maturational processes and delayed onset in schizophrenia. *Development and Psychopathology*, *11*(3), 525–543. https://doi.org/10.1017/S0954579499002199

Kirov, G., Rees, E., Walters, J. T. R., Escott-Price, V., Georgieva, L., Richards, A. L., Chambert, K. D., Davies, G., Legge, S. E., Moran, J. L., McCarroll, S. A., O'Donovan, M. C., & Owen, M. J. (2014). The penetrance of copy number variations for schizophrenia and developmental delay. *Biological Psychiatry*, *75*(5), 378–385. https://doi.org/10.1016/j.biopsych.2013.07.022

Kobrynski, L. J., & Sullivan, K. E. (2007). Velocardiofacial syndrome, DiGeorge syndrome: the chromosome 22q11.2 deletion syndromes. *Lancet*, *370*(9596), 1443–1452. https://doi.org/10.1016/S0140-6736(07)61601-8

Kotlicka-Antczak, M., Pawełczyk, A., Pawełczyk, T., Strzelecki, D., Żurner, N., & Karbownik, M. S. (2018). A history of obstetric complications is associated with the risk of progression from an at risk mental state to psychosis. *Schizophrenia Research*, *197*, 498–503. https://doi.org/10.1016/j.schres.2017.10.039

Koutsouleris, N., Borgwardt, S., Meisenzahl, E. M., Bottlender, R., Möller, H.-J., & Riecher-Rössler, A. (2012). Disease Prediction in the At-Risk Mental State for Psychosis Using Neuroanatomical Biomarkers: Results From the FePsy Study. *Schizophrenia Bulletin*, *38*(6), 1234–1246. https://doi.org/10.1093/schbul/sbr145

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D. B., Paolini, M., Chisholm, K., Kambeitz, J., Haidl, T., Schmidt, A., Gillam, J., Schultze-Lutter, F., Falkai, P., Reiser, M., Riecher-Rössler, A., Upthegrove, R., Hietala, J., Salokangas, R. K. R., … Borgwardt, S. (2018). Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or with Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. *JAMA Psychiatry*, *75*(11), 1156–1172. https://doi.org/10.1001/jamapsychiatry.2018.2165

Koutsouleris, N., Meisenzahl, E. M., Borgwardt, S., Riecher-Rössler, A., Frodl, T., Kambeitz, J., Köhler, Y., Falkai, P., Möller, H.-J., Reiser, M., & Davatzikos, C. (2015). Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. *Brain*, *138*(7), 2059–2073. https://doi.org/10.1093/brain/awv111

Koutsouleris, N., Meisenzahl, E. M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., Schmitt, G., Zetzsche, T., Decker, P., Reiser, M., Möller, H.-J., & Gaser, C. (2009). Use of neuroanatomical pattern classification to identify subjects in

at-risk mental states of psychosis and predict disease transition. *Archives of General Psychiatry*, *66*(7), 700–712. https://doi.org/10.1001/archgenpsychiatry.2009.62

Koutsouleris, N., Riecher-Rössler, A., Meisenzahl, E. M., Smieskova, R., Studerus, E., Kambeitz-Ilankovic, L., Von Saldern, S., Cabral, C., Reiser, M., Falkai, P., & Borgwardt, S. (2015). Detecting the Psychosis Prodrome Across High-Risk Populations Using Neuroanatomical Biomarkers. *Schizophrenia Bulletin*, *41*(2), 471–482. https://doi.org/10.1093/schbul/sbu078

Kraan, T., Velthorst, E., Koenders, L., Zwaart, K., Ising, H. K., Van Den Berg, D., De Haan, L., & Van Der Gaag, M. (2016). Cannabis use and transition to psychosis in individuals at ultra-high risk: Review and meta-analysis. *Psychological Medicine*, *46*(4), 673–681. https://doi.org/10.1017/S0033291715002329

Kuo, S. S., & Pogue-Geile, M. F. (2019). Variation in fourteen brain structure volumes in schizophrenia: A comprehensive meta-analysis of 246 studies. *Neuroscience and Biobehavioral Reviews*, *98*(November 2018), 85–94. https://doi.org/10.1016/j.neubiorev.2018.12.030

Lane, C. A., Hardy, J., & Schott, J. M. (2018). Alzheimer's disease. *European Journal of Neurology*, *25*(1), 59–70. https://doi.org/10.1111/ene.13439

Lappalainen, T., Sammeth, M., Friedländer, M. R., 'T Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., Van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., … Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506–511. https://doi.org/10.1038/nature12531

Lee, K. W., Woon, P. S., Teo, Y. Y., & Sim, K. (2012). Genome wide association studies (GWAS) and copy number variation (CNV) studies of the major psychoses: What have we learnt? *Neuroscience and Biobehavioral Reviews*, *36*(1), 556–571. https://doi.org/10.1016/j.neubiorev.2011.09.001

Leucht, S., Leucht, C., Huhn, M., Chaimani, A., Mavridis, D., Helfer, B., Samara, M., Rabaioli, M., Bächer, S., Cipriani, A., Geddes, J. R., Salanti, G., & Davis, J. M. (2017). Sixty years of placebo-controlled antipsychotic drug trials in acute schizophrenia: Systematic review, Bayesian meta-analysis, and meta-regression of efficacy predictors. *American Journal of Psychiatry*, *174*(10), 927–942. https://doi.org/10.1176/appi.ajp.2017.16121358

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype Imputation. *Annual Review of Genomics and Human Genetics*, *10*(1), 387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242

Loh, P. R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., De Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., O'Donovan, M. C., Neale, B. M., Patterson, N., & Price, A. L. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics*, *47*(12), 1385–1392. https://doi.org/10.1038/ng.3431

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The

Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653

Lu, X., Yang, Y., Wu, F., Gao, M., Xu, Y., Zhang, Y., Yao, Y., Du, X., Li, C., Wu, L., Zhong, X., Zhou, Y., Fan, N., Zheng, Y., Xiong, D., Peng, H., Escudero, J., Huang, B., Li, X., … Wu, K. (2016). Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images. *Medicine (United States)*, *95*(30). https://doi.org/10.1097/MD.0000000000003973

Luders, E., Thompson, P. M., Narr, K. L., Toga, A. W., Jancke, L., & Gaser, C. (2006). A curvature-based approach to estimate local gyrification on the cortical surface. *NeuroImage*, *29*(4), 1224–1230. https://doi.org/10.1016/j.neuroimage.2005.08.049

Madeira, N., Duarte, J. V., Martins, R., Costa, G. N., Macedo, A., & Castelo-Branco, M. (2020). Morphometry and gyrification in bipolar disorder and schizophrenia: A comparative MRI study. *NeuroImage: Clinical*, *26*(January). https://doi.org/10.1016/j.nicl.2020.102220

Marconi, A., Di Forti, M., Lewis, C. M., Murray, R. M., & Vassos, E. (2016). Meta-Analysis of the association between the level of cannabis use and risk of psychosis. *Schizophrenia Bulletin*, *42*(5), 1262–1269. https://doi.org/10.1093/schbul/sbw003

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, *22*(12), 2677–2684. https://doi.org/10.1162/jocn.2009.21407

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. *Journal of Cognitive Neuroscience*, *19*(9), 1498–1507. https://doi.org/10.1162/jocn.2007.19.9.1498

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2), e1608. https://doi.org/10.1002/mpr.1608

McIntosh, A. M., Moorhead, T. W. J., McKirdy, J., Hall, J., Sussmann, J. E. D., Stanfield, A. C., Harris, J. M., Johnstone, E. C., & Lawrie, S. M. (2009). Prefrontal gyral folding and its cognitive correlates in bipolar disorder and schizophrenia. *Acta Psychiatrica Scandinavica*, *119*(3), 192–198. https://doi.org/10.1111/j.1600-0447.2008.01286.x

Millan, M. J., Andrieux, A., Bartzokis, G., Cadenhead, K., Dazzan, P., Fusar-Poli, P., Gallinat, J., Giedd, J., Grayson, D. R., Heinrichs, M., Kahn, R., Krebs, M. O., Leboyer, M., Lewis, D., Marin, O., Marin, P., Meyer-Lindenberg, A., McGorry, P., McGuire, P., … Weinberger, D. (2016). Altering the course of schizophrenia: Progress and perspectives. *Nature Reviews Drug Discovery*, *15*(7), 485–515. https://doi.org/10.1038/nrd.2016.28

Miller, B., Messias, E., Miettunen, J., Alaräisänen, A., Järvelin, M. R., Koponen, H., Räsänen, P., Isohanni, M., & Kirkpatrick, B. (2011). Meta-analysis of paternal age and schizophrenia risk in male versus female offspring. *Schizophrenia Bulletin*, *37*(5), 1039–1047. https://doi.org/10.1093/schbul/sbq011

Miller, T. J., McGlashan, T. H., Rosen, J. L., Cadenhead, K., Ventura, J., McFarlane, W., Perkins, D. O., Pearlson, G. D., & Woods, S. W. (2003). Prodromal Assessment With the Structured Interview for Prodromal Syndromes and the Scale of Prodromal Symptoms: Predictive Validity, Interrater Reliability, and Training to Reliability. *Schizophrenia Bulletin*, *29*(4), 703–715. https://doi.org/10.1093/oxfordjournals.schbul.a007040

Mittal, V. A., Willhite, R., Daley, M., Bearden, C. E., Niendam, T., Ellman, L. M., & Cannon, T. D. (2009). Obstetric complications and risk for conversion to psychosis among individuals at high clinical risk. *Early Intervention in Psychiatry*, *3*(3), 226–230. https://doi.org/10.1111/j.1751-7893.2009.00137.x

Montoliu-Gaya, L., Tietze, D., Kaminski, D., Mirgorodskaya, E., Tietze, A. A., & Sterky, F. H. (2021). CA10 regulates neurexin heparan sulfate addition via a direct binding in the secretory pathway. *EMBO Reports*, *22*(4), e51349. https://doi.org/10.15252/embr.202051349

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*, *43*(11), 2412–2412. https://doi.org/10.1212/WNL.43.11.2412-a

Mourao-Miranda, J., Reinders, A. A. T. S., Rocha-Rego, V., Lappin, J., Rondina, J., Morgan, C., Morgan, K. D., Fearon, P., Jones, P. B., Doody, G. A., Murray, R. M., Kapur, S., & Dazzan, P. (2012). Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study. *Psychological Medicine*, *42*(5), 1037–1047. https://doi.org/10.1017/S0033291711002005

Murphy, K. C., Jones, L. A., & Owen, M. J. (1999). High rates of schizophrenia in adults with velo-cardio-facial syndrome. *Archives of General Psychiatry*, *56*(10), 940–945. https://doi.org/10.1001/archpsyc.56.10.940

Nagy, Z., Westerberg, H., & Klingberg, T. (2004). Maturation of white matter is associated with the development of cognitive functions during childhood. *Journal of Cognitive Neuroscience*, *16*(7), 1227–1233. https://doi.org/10.1162/0898929041920441

Nakamura, K., Kawasaki, Y., Suzuki, M., Hagino, H., Kurokawa, K., Takahashi, T., Niu, L., Matsui, M., Seto, H., & Kurachi, M. (2004). Multiple structural brain measures obtained by three-dimensional magnetic resonance imaging to distinguish between schizophrenia patients and normal subjects. *Schizophrenia Bulletin*, *30*(2), 393–404. https://doi.org/10.1093/oxfordjournals.schbul.a007087

Nakamura, M., Salisbury, D. F., Hirayasu, Y., Bouix, S., Pohl, K. M., Yoshida, T., Koo, M.-S., Shenton, M. E., & McCarley, R. W. (2007). Neocortical Gray Matter Volume in First-Episode Schizophrenia and First-Episode Affective Psychosis: A Cross-Sectional and Longitudinal MRI Study. *Biological Psychiatry*, *62*(7), 773–783. https://doi.org/10.1016/j.biopsych.2007.03.030

Nanda, P., Tandon, N., Mathew, I. T., Giakoumatos, C. I., Abhishekh, H. A., Clementz, B. A., Pearlson, G. D., Sweeney, J., Tamminga, C. A., & Keshavan, M. S. (2014). Local gyrification index in Probands with psychotic disorders and their first-degree relatives. *Biological Psychiatry*, *76*(6), 447–455. https://doi.org/10.1016/j.biopsych.2013.11.018

Navari, S., & Dazzan, P. (2009). Do antipsychotic drugs affect brain structure? A systematic and critical review of MRI findings. *Psychological Medicine*, *39*(11),

1763–1777. https://doi.org/10.1017/S0033291709005315

Nelson, H. E. (1982). The National Adult Reading Test (NART): Test Manual. *Windsor, UK: NFER-Nelson*, *124*(3), 0–25.

Nieuwenhuis, M., van Haren, N. E. M., Hulshoff Pol, H. E., Cahn, W., Kahn, R. S., & Schnack, H. G. (2012). Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *NeuroImage*, *61*(3), 606–612. https://doi.org/10.1016/j.neuroimage.2012.03.079

Nowak, I., Sabariego, C., Świtaj, P., & Anczewska, M. (2016). Disability and recovery in schizophrenia: a systematic review of cognitive behavioral therapy interventions. *BMC Psychiatry*, *16*(1), 228. https://doi.org/10.1186/s12888-016-0912-8

Nurnberger Jr, J. I., Blehar, M. C., Kaufmann, C. A., York-Cooler, C., Simpson, S. G., Harkavy-Friedman, J., Severe, J. B., Malaspina, D., & Reich, T. (1994). Diagnostic Interview for Genetic Studies: Rationale, Unique Features, and Training. *Archives of General Psychiatry*, *51*(11), 849–859. https://doi.org/10.1001/archpsyc.1994.03950110009002

Oliver, D., Radua, J., Reichenberg, A., Uher, R., & Fusar-Poli, P. (2019). Psychosis Polyrisk Score (PPS) for the Detection of Individuals At-Risk and the Prediction of Their Outcomes. *Frontiers in Psychiatry*, *10*(MAR). https://doi.org/10.3389/fpsyt.2019.00174

Oliver, D., Spada, G., Englund, A., Chesney, E., Radua, J., Reichenberg, A., Uher, R., McGuire, P., & Fusar-Poli, P. (2020). Real-world digital implementation of the Psychosis Polyrisk Score (PPS): A pilot feasibility study. *Schizophrenia Research*, *xxxx*, 4–11. https://doi.org/10.1016/j.schres.2020.04.015

Ota, M., Sato, N., Ishikawa, M., Hori, H., Sasayama, D., Hattori, K., Teraishi, T., Obu, S., Nakata, Y., Nemoto, K., Moriguchi, Y., Hashimoto, R., & Kunugi, H. (2012). Discrimination of female schizophrenia patients from healthy women using multiple structural brain measures obtained with voxel-based morphometry. *Psychiatry and Clinical Neurosciences*, *66*(7), 611–617. https://doi.org/10.1111/j.1440-1819.2012.02397.x

Padmanabhan, J. L., Shah, J. L., Tandon, N., & Keshavan, M. S. (2017). The "polyenviromic risk score": Aggregating environmental risk factors predicts conversion to psychosis in familial high-risk subjects. *Schizophrenia Research*, *181*, 17–22. https://doi.org/10.1016/j.schres.2016.10.014

Pantelis, C., Velakoulis, D., McGorry, P. D., Wood, S. J., Suckling, J., Phillips, L. J., Yung, A. R., Bullmore, E. T., Brewer, W., Soulsby, B., Desmond, P., & McGuire, P. K. (2003). Neuroanatomical abnormalities before and after onset of psychosis: A cross-sectional and longitudinal MRI comparison. *Lancet*, *361*(9354), 281–288. https://doi.org/10.1016/S0140-6736(03)12323-9

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S. E., Bishop, S., Cameron, D., Hamshere, M. L., Han, J., Hubbard, L., Lynham, A., Mantripragada, K., Rees, E., MacCabe, J. H., McCarroll, S. A., Baune, B. T., Breen, G., … Walters, J. T. R. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics*, *50*(3), 381–389. https://doi.org/10.1038/s41588-018-0059-

Pardo, P. J., Georgopoulos, A. P., Kenny, J. T., Stuve, T. A., Findling, R. L., & Schulz, S. C. (2006). Classification of adolescent psychotic disorders using linear discriminant analysis. *Schizophrenia Research*, *87*(1–3), 297–306. https://doi.org/10.1016/j.schres.2006.05.007

Pettersson-Yeo, W., Benetti, S., Marquand, A. F., Dell'Acqua, F., Williams, S. C. R., Allen, P., Prata, D., McGuire, P., & Mechelli, A. (2013). Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. *Psychological Medicine*, *43*(12), 2547–2562. https://doi.org/10.1017/S003329171300024X

Pettersson-Yeo, W., Benetti, S., Marquand, A. F., Dell'acqua, F., Williams, S. C., Allen, P., Prata, D., McGuire, P., & Mechelli, A. (2013). Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. *Psychol Med*, *43*(12), 2547–2562. https://doi.org/10.1017/S003329171300024X

Pharoah, F., Mari, J. J., Rathbone, J., & Wong, W. (2010). Family intervention for schizophrenia. *Cochrane Database of Systematic Reviews*, *4*. https://doi.org/10.1002/14651858.CD000088.pub3

Phillips, L. J., Yung, A. R., & McGorry, P. D. (2000). Identification of young people at risk of psychosis: validation of Personal Assessment and Crisis Evaluation Clinic intake criteria. *The Australian and New Zealand Journal of Psychiatry*, *34 Suppl*, S164-9. https://doi.org/10.1080/000486700239

Pina-Camacho, L., Garcia-Prieto, J., Parellada, M., Castro-Fornieles, J., Gonzalez-Pinto, A. M., Bombin, I., Graell, M., Paya, B., Rapado-Castro, M., Janssen, J., Baeza, I., Pozo, F. Del, Desco, M., & Arango, C. (2015). Predictors of schizophrenia spectrum disorders in early-onset first episodes of psychosis: a support vector machine model. *European Child and Adolescent Psychiatry*, *24*(4), 427–440. https://doi.org/10.1007/s00787-014-0593-0

Pinaya, W. H. L., Gadelha, A., Doyle, O. M., Noto, C., Zugman, A., Cordeiro, Q., Jackowski, A. P., Bressan, R. A., & Sato, J. R. (2016). Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Scientific Reports*, *6*(November), 1–9. https://doi.org/10.1038/srep38897

Pinaya, W. H. L., Mechelli, A., & Sato, J. R. (2019). Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. *Human Brain Mapping*, *40*(3), 944–954. https://doi.org/10.1002/hbm.24423

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2018). *{nlme}: Linear and Nonlinear Mixed Effects Models*. https://cran.r-project.org/package=nlme

Prata, D., Mechelli, A., & Kapur, S. (2014). Clinically meaningful biomarkers for psychosis: A systematic and quantitative review. *Neuroscience & Biobehavioral Reviews*, *45*, 134–141. https://doi.org/10.1016/j.neubiorev.2014.05.010

Preti, A., & Cella, M. (2010). Randomized-controlled trials in people at ultra high risk of psychosis: A review of treatment effectiveness. *Schizophrenia Research*, *123*(1), 30–36. https://doi.org/10.1016/j.schres.2010.07.026

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Radulescu, E., Ganeshan, B., Shergill, S. S., Medford, N., Chatwin, C., Young, R. C. D., & Critchley, H. D. (2014). Grey-matter texture abnormalities and reduced hippocampal volume are distinguishing features of schizophrenia. *Psychiatry Research - Neuroimaging*, *223*(3), 179–186. https://doi.org/10.1016/j.pscychresns.2014.05.014

Rajapakse, J. C., Giedd, J. N., & Rapoport, J. L. (1997). Statistical approach to segmentation of single-channel cerebral mr images. *IEEE Transactions on Medical Imaging*, *16*(2), 176–186. https://doi.org/10.1109/42.563663

Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., Coin, L., De Silva, R., Cookson, M. R., Singleton, A. B., Hardy, J., Ryten, M., & Weale, M. E. (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature Neuroscience*, *17*(10), 1418–1428. https://doi.org/10.1038/nn.3801

Rapoport, J. L., Giedd, J. N., & Gogtay, N. (2012). Neurodevelopmental model of schizophrenia: Update 2012. *Molecular Psychiatry*, *17*(12), 1228–1238. https://doi.org/10.1038/mp.2012.23

Riecher-Rössler, A., Aston, J., Ventura, J., Merlo, M., Borgwardt, S., Gschwandtner, U., & Stieglitz, R. D. (2008). Das Basel Screening Instrument für Psychosen (BSIP): Entwicklung, aufbau, reliabilität und validität. *Fortschritte Der Neurologie Psychiatrie*, *76*(4), 207–216. https://doi.org/10.1055/s-2008-1038155

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H., Pers, T. H., Agartz, I., Agerbo, E., Albus, M., Alexander, M., Amin, F., Bacanu, S. A., Begemann, M., Belliveau, R. A., … O'Donovan, M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421–427. https://doi.org/10.1038/nature13595

Ripke, S., Sanders, A. R., Kendler, K. S., Levinson, D. F., Sklar, P., Holmans, P. A., Lin, D. Y., Duan, J., Ophoff, R. A., Andreassen, O. A., Scolnick, E., Cichon, S., St. Clair, D., Corvin, A., Gurling, H., Werge, T., Rujescu, D., Blackwood, D. H. R., Pato, C. N., … Gejman, P. V. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics*, *43*(10), 969–976. https://doi.org/10.1038/ng.940

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Rozycki, M., Satterthwaite, T. D., Koutsouleris, N., Erus, G., Doshi, J., Wolf, D. H., Fan, Y., Gur, R. E., Gur, R. C., Meisenzahl, E. M., Zhuo, C., Yin, H., Yan, H., Yue, W., Zhang, D., & Davatzikos, C. (2018). Multisite Machine Learning Analysis Provides a Robust Structural Imaging Signature of Schizophrenia Detectable Across Diverse Patient Populations and Within Individuals. *Schizophrenia Bulletin*, *44*(5), 1035–1044. https://doi.org/10.1093/schbul/sbx137

Ruigrok, A. N. V., Salimi-Khorshidi, G., Lai, M. C., Baron-Cohen, S., Lombardo, M. V., Tait, R. J., & Suckling, J. (2014). A meta-analysis of sex differences in human brain structure. *Neuroscience and Biobehavioral Reviews*, *39*, 34–50. https://doi.org/10.1016/j.neubiorev.2013.12.004

Sallet, P. C., Elkis, H., Alves, T. M., Oliveira, J. R., Sassi, E., Campi De Castro, C., Busatto, G. F., & Gattaz, W. F. (2003). Reduced cortical folding in schizophrenia: An MRI morphometric study. *American Journal of Psychiatry*, *160*(9), 1606–1613. https://doi.org/10.1176/appi.ajp.160.9.1606

Salvador, R., Radua, J., Canales-Rodrõâguez, E. J., Solanes, A., Sarroa, S., Goikolea, J. M., Valiente, A., Montea, G. C., Natividad, M. D. C., Guerrero-Pedraza, A., Moro, N., Fernaândez-Corcuera, P., Amann, B. L., Maristany, T., Vieta, E., Mckenna, P. J., & Pomarol-Clote, E. (2017). Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. *PLoS ONE*, *12*(4), 1–24. https://doi.org/10.1371/journal.pone.0175683

Sanfelici, R., Dwyer, D. B., Antonucci, L. A., & Koutsouleris, N. (2020). Individualized Diagnostic and Prognostic Models for Patients With Psychosis Risk Syndromes: A Meta-analytic View on the State of the Art. *Biological Psychiatry*, *88*(4), 349–360. https://doi.org/10.1016/j.biopsych.2020.02.009

Schmitt, A., Malchow, B., Hasan, A., & Falkai, P. (2014). The impact of environmental factors in severe psychiatric disorders. *Frontiers in Neuroscience*, *8*(8 FEB), 1–10. https://doi.org/10.3389/fnins.2014.00019

Schnack, H. G., & Kahn, R. S. (2016). Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Frontiers in Psychiatry*, *7*(MAR). https://doi.org/10.3389/fpsyt.2016.00050

Schnack, H. G., Nieuwenhuis, M., van Haren, N. E. M., Abramovic, L., Scheewe, T. W., Brouwer, R. M., Hulshoff Pol, H. E., & Kahn, R. S. (2014). Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. *NeuroImage*, *84*, 299–306. https://doi.org/10.1016/j.neuroimage.2013.08.053

Scholkopf, B, Guyon, I., & Weston, J. (2003). Statistical learning and kernel methods in bioinformatics. *Artificial Intelligence and Heuristic Methods in Bioinformatics*, *1*(2), 1–21. https://clopinet.com/isabelle/Papers/kerbioinfo.pdf

Scholkopf, Bernhard, Williamson, R. C., & Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, *1245*, 1207–1245.

Schrimpf, L. A., Aggarwal, A., & Lauriello, J. (2018). Psychosis. *CONTINUUM: Lifelong Learning in Neurology*, *24*, 845–860. https://doi.org/10.1212/CON.0000000000000602

Schultze-Lutter, F., Ruhrmann, S., Fusar-Poli, P., Bechdolf, A., G. Schimmelmann, B., & Klosterkotter, J. (2012). Basic Symptoms and the Prediction of First-Episode Psychosis. *Current Pharmaceutical Design*, *18*(4), 351–357. https://doi.org/10.2174/138161212799316064

Seeman, M. V. (2019). Schizophrenia Mortality: Barriers to Progress. *Psychiatric Quarterly*, *90*(3), 553–563. https://doi.org/10.1007/s11126-019-09645-0

Selten, J. P., Van Der Ven, E., & Termorshuizen, F. (2019). Migration and psychosis: A meta-analysis of incidence studies. *Psychological Medicine*. https://doi.org/10.1017/S0033291719000035

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A

practical guide to calculating Cohen's f 2, a measure of local effect size, from PROC MIXED. *Frontiers in Psychology*, *3*(APR), 1–6. https://doi.org/10.3389/fpsyg.2012.00111

Shabalin, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, *28*(10), 1353–1358. https://doi.org/10.1093/bioinformatics/bts163

Sheffield, J. M., Karcher, N. R., & Barch, D. M. (2018). Cognitive Deficits in Psychotic Disorders: A Lifespan Perspective. *Neuropsychology Review*, *28*(4), 509–533. https://doi.org/10.1007/s11065-018-9388-2

Shepherd, A. M., Laurens, K. R., Matheson, S. L., Carr, V. J., & Green, M. J. (2012). Systematic meta-review and quality assessment of the structural brain alterations in schizophrenia. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1342–1356. https://doi.org/10.1016/j.neubiorev.2011.12.015

Shrivastava, A., Bureau, Y., Rewari, N., & Johnston, M. (2013). Clinical risk of stigma and discrimination of mental illnesses: Need for objective assessment and quantification. *Indian Journal of Psychiatry*, *55*(2), 178–182. https://doi.org/10.4103/0019-5545.111459

Siafis, S., Tzachanis, D., Samara, M., & Papazisis, G. (2017). Antipsychotic Drugs: From Receptor-binding Profiles to Metabolic Side Effects. *Current Neuropharmacology*, *16*(8), 1210–1223. https://doi.org/10.2174/1570159x15666170630163616

Smieskova, R., Fusar-Poli, P., Aston, J., Simon, A., Bendfeldt, K., Lenz, C., Stieglitz, R. D., McGuire, P., Riecher-Rössler, A., & Borgwardt, S. J. (2012). Insular volume abnormalities associated with different transition probabilities to psychosis. *Psychological Medicine*, *42*(8), 1613–1625. https://doi.org/10.1017/S0033291711002716

Spalthoff, R., Gaser, C., & Nenadić, I. (2018). Altered gyrification in schizophrenia and its relation to other morphometric markers. *Schizophrenia Research*, *202*, 195–202. https://doi.org/10.1016/j.schres.2018.07.014

Sprooten, E., Papmeyer, M., Smyth, A. M., Vincenz, D., Honold, S., Conlon, G. A., Moorhead, T. W. J., Job, D., Whalley, H. C., Hall, J., McIntosh, A. M., Owens, D. C. G., Johnstone, E. C., & Lawrie, S. M. (2013). Cortical thickness in first-episode schizophrenia patients and individuals at high familial risk: A cross-sectional comparison. *Schizophrenia Research*, *151*(1–3), 259–264. https://doi.org/10.1016/j.schres.2013.09.024

Stafford, M. R., Jackson, H., Mayo-Wilson, E., Morrison, A. P., & Kendall, T. (2013). Early interventions to prevent psychosis: systematic review and meta-analysis. *BMJ (Clinical Research Ed.)*, *346*(January), f185. https://doi.org/10.1136/bmj.f185

Struyf, J., Dobrin, S., & Page, D. (2008). Combining gene expression, demographic and clinical data in modeling disease: A case study of bipolar disorder and schizophrenia. *BMC Genomics*, *9*. https://doi.org/10.1186/1471-2164-9-531

Sui, J., He, H., Yu, Q., Chen, J., Rogers, J., Pearlson, G. D., Mayer, A., Bustillo, J., Canive, J., & Calhoun, V. D. (2013). Combination of Resting State fMRI, DTI, and sMRI Data to Discriminate Schizophrenia by N-way MCCA + jICA. *Frontiers in Human Neuroscience*, *7*(May), 1–14. https://doi.org/10.3389/fnhum.2013.00235

Sullivan, P. F., Kendler, K. S., & Neale, M. C. (2003). Schizophrenia as a Complex Trait. *Archives of General Psychiatry*, *60*(12), 1187. https://doi.org/10.1001/archpsyc.60.12.1187

Sun, D., van Erp, T. G. M., Thompson, P. M., Bearden, C. E., Daley, M., Kushan, L., Hardt, M. E., Nuechterlein, K. H., Toga, A. W., & Cannon, T. D. (2009). Elucidating a Magnetic Resonance Imaging-Based Neuroanatomic Biomarker for Psychosis: Classification Analysis Using Probabilistic Brain Atlas and Machine Learning Algorithms. *Biological Psychiatry*, *66*(11), 1055–1060. https://doi.org/10.1016/j.biopsych.2009.07.019

Suzuki, M., Nohara, S., Hagino, H., Kurokawa, K., Yotsutsuji, T., Kawasaki, Y., Takahashi, T., Matsui, M., Watanabe, N., Seto, H., & Kurachi, M. (2002). Regional changes in brain gray and white matter in patients with schizophrenia demonstrated with voxel-based analysis of MRI. *Schizophrenia Research*, *55*(1–2), 41–54. https://doi.org/10.1016/S0920-9964(01)00224-9

Tai, A. M. Y., Albuquerque, A., Carmona, N. E., Subramanieapillai, M., Cha, D. S., Sheko, M., Lee, Y., Mansur, R., & McIntyre, R. S. (2019). Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry. *Artificial Intelligence in Medicine*, *99*(April 2017), 101704. https://doi.org/10.1016/j.artmed.2019.101704

Takahashi, T., Wood, S. J., Yung, A. R., Phillips, L. J., Soulsby, B., McGorry, P. D., Tanino, R., Zhou, S. Y., Suzuki, M., Velakoulis, D., & Pantelis, C. (2009). Insular cortex gray matter changes in individuals at ultra-high-risk of developing psychosis. *Schizophrenia Research*, *111*(1–3), 94–102. https://doi.org/10.1016/j.schres.2009.03.024

Takayanagi, Y., Takahashi, T., Orikabe, L., Mozue, Y., Kawasaki, Y., Nakamura, K., Sato, Y., Itokawa, M., Yamasue, H., Kasai, K., Kurachi, M., Okazaki, Y., & Suzuki, M. (2011). Classification of first-episode schizophrenia patients and healthy subjects by automated MRI measures of regional brain volume and cortical thickness. *PLoS ONE*, *6*(6), 1–10. https://doi.org/10.1371/journal.pone.0021047

Tavares, V., Prata, D., & Ferreira, H. A. (2020). Comparing SPM12 and CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study. *Journal of Neuroscience Methods*, *334*. https://doi.org/10.1016/j.jneumeth.2019.108565

Tavares, Vânia, Monteiro, J., Vassos, E., Coleman, J., & Prata, D. (2021). Evaluation of Genotype-Based Gene Expression Model Performance: A Cross-Framework and Cross-Dataset Study. *Genes*, *12*(10), 1531. https://doi.org/10.3390/genes12101531

Tavares, Vânia, Prata, D., & Ferreira, H. A. (2020). Comparing SPM12 and CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study. *Journal of Neuroscience Methods*, *334*, 108565. https://doi.org/10.1016/J.JNEUMETH.2019.108565

Teipel, S. J., Grothe, M., Lista, S., Toschi, N., Garaci, F. G., & Hampel, H. (2013). Relevance of Magnetic Resonance Imaging for Early Detection and Diagnosis of Alzheimer Disease. *Medical Clinics of North America*, *97*(3), 399–424. https://doi.org/10.1016/j.mcna.2012.12.013

Thompson, A. D., Nelson, B., Yuen, H. P., Lin, A., Amminger, G. P., McGorry, P. D., Wood, S. J., & Yung, A. R. (2014). Sexual trauma increases the risk of developing psychosis in an ultra high-risk "prodromal" population. *Schizophrenia Bulletin*, *40*(3), 697–706. https://doi.org/10.1093/schbul/sbt032

Toepper, M. (2017). Dissociating Normal Aging from Alzheimer's Disease: A View from Cognitive Neuroscience. *Journal of Alzheimer's Disease : JAD*, *57*(2), 331–352. https://doi.org/10.3233/JAD-161099

Tognin, S, Riecher-Rössler, A., Meisenzahl, E. M., Wood, S. J., Hutton, C., Borgwardt, S. J., Koutsouleris, N., Yung, A. R., Allen, P., Phillips, L. J., McGorry, P. D., Valli, I., Velakoulis, D., Nelson, B., Woolley, J., Pantelis, C., McGuire, P., & Mechelli, A. (2014). Reduced parahippocampal cortical thickness in subjects at ultra-high risk for psychosis. *Psychological Medicine*, *44*(3), 489–498. https://doi.org/10.1017/S0033291713000998

Tognin, Stefania, Pettersson-Yeo, W., Valli, I., Hutton, C., Woolley, J., Allen, P., McGuire, P., & Mechelli, A. (2014). Using Structural Neuroimaging to Make Quantitative Predictions of Symptom Progression in Individuals at Ultra-High Risk for Psychosis. *Frontiers in Psychiatry*, *4*(JAN), 1–9. https://doi.org/10.3389/fpsyt.2013.00187

Tognin, Stefania, Van Hell, H. H., Merritt, K., Winter-Van Rossum, I., Bossong, M. G., Kempton, M. J., Modinos, G., Fusar-Poli, P., Mechelli, A., Dazzan, P., Maat, A., De Haan, L., Crespo-Facorro, B., Glenthøj, B., Lawrie, S. M., McDonald, C., Gruber, O., Van Amelsvoort, T., Arango, C., … Morgan, S. (2020). Towards precision medicine in psychosis: Benefits and challenges of multimodal multicenter studies - PSYSCAN: Translating neuroimaging findings from research into clinical practice. *Schizophrenia Bulletin*, *46*(2), 432–441. https://doi.org/10.1093/schbul/sbz067

Trifu, S., Kohn, B., Vlasie, A., & Patrichi, B.-E. (2020). Genetics of schizophrenia (Review). *Experimental and Therapeutic Medicine*, 59–70. https://doi.org/10.3892/etm.2020.8973

Turner, D. T., McGlanaghy, E., Cuijpers, P., Van Der Gaag, M., Karyotaki, E., & MacBeth, A. (2018). A Meta-Analysis of Social Skills Training and Related Interventions for Psychosis. *Schizophrenia Bulletin*, *44*(3), 475–491. https://doi.org/10.1093/schbul/sbx146

Valli, I., Marquand, A. F., Mechelli, A., Raffin, M., Allen, P., Seal, M. L., & McGuire, P. (2016). Identifying individuals at high risk of psychosis: Predictive utility of support vector machine using structural and functional MRI data. *Frontiers in Psychiatry*, *7*(APR), 1–8. https://doi.org/10.3389/fpsyt.2016.00052

Van Der Gaag, M., Smit, F., Bechdolf, A., French, P., Linszen, D. H., Yung, A. R., McGorry, P., & Cuijpers, P. (2013). Preventing a first episode of psychosis: Meta-analysis of randomized controlled prevention trials of 12month and longer-term follow-ups. *Schizophrenia Research*, *149*(1–3), 56–62. https://doi.org/10.1016/j.schres.2013.07.004

Van Haren, N. E. M., Schnack, H. G., Cahn, W., Van Den Heuvel, M. P., Lepage, C., Collins, L., Evans, A. C., Hulshoff Pol, H. E., & Kahn, R. S. (2011). Changes in cortical thickness during the course of illness in schizophrenia. *Archives of General Psychiatry*, *68*(9), 871–880. https://doi.org/10.1001/archgenpsychiatry.2011.88

Varese, F., Smeets, F., Drukker, M., Lieverse, R., Lataster, T., Viechtbauer, W., Read, J., Van Os, J., & Bentall, R. P. (2012). Childhood adversities increase the risk of psychosis: A meta-analysis of patient-control, prospective-and cross-sectional cohort studies. *Schizophrenia Bulletin*, *38*(4), 661–671. https://doi.org/10.1093/schbul/sbs050

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, *145*(August 2015), 166–179. https://doi.org/10.1016/j.neuroimage.2016.10.038

Vassos, E., Di Forti, M., Coleman, J., Iyegbe, C., Prata, D., Euesden, J., O'Reilly, P., Curtis, C., Kolliakou, A., Patel, H., Newhouse, S., Traylor, M., Ajnakina, O., Mondelli, V., Marques, T. R., Gardner-Sood, P., Aitchison, K. J., Powell, J., Atakan, Z., … al., et. (2016). An Examination of Polygenic Score Risk Prediction in Individuals with First Episode Psychosis. *Biological Psychiatry*, *0*(0), 135–145. https://doi.org/10.1016/j.biopsych.2016.06.028

Vassos, E., Di Forti, M., Coleman, J., Iyegbe, C., Prata, D., Euesden, J., O'Reilly, P., Curtis, C., Kolliakou, A., Patel, H., Newhouse, S., Traylor, M., Ajnakina, O., Mondelli, V., Marques, T. R., Gardner-Sood, P., Aitchison, K. J., Powell, J., Atakan, Z., … Breen, G. (2017). An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis. *Biological Psychiatry*, *81*(6), 470–477. https://doi.org/10.1016/j.biopsych.2016.06.028

Vassos, E., Pedersen, C. B., Murray, R. M., Collier, D. A., & Lewis, C. M. (2012). Meta-Analysis of the Association of Urbanicity With Schizophrenia. *Schizophrenia Bulletin*, *38*(6), 1118–1123. https://doi.org/10.1093/schbul/sbs096

Vassos, E., Sham, P., Kempton, M., Trotta, A., Stilo, S. A., Gayer-Anderson, C., Di Forti, M., Lewis, C. M., Murray, R. M., & Morgan, C. (2020). The Maudsley environmental risk score for psychosis. *Psychological Medicine*, *50*(13), 2213–2220. https://doi.org/10.1017/S0033291719002319

Vieira, S., Gong, Q. Y., Pinaya, W. H. L., Scarpazza, C., Tognin, S., Crespo-Facorro, B., Tordesillas-Gutierrez, D., Ortiz-García, V., Setien-Suero, E., Scheepers, F. E., van Haren, N. E. M., Marques, T. R., Murray, R. M., David, A., Dazzan, P., McGuire, P., & Mechelli, A. (2020). Using machine learning and structural neuroimaging to detect first episode psychosis: Reconsidering the evidence. *Schizophrenia Bulletin*, *46*(1), 17–26. https://doi.org/10.1093/schbul/sby189

Vivian-Griffiths, T., Baker, E., Schmidt, K. M., Bracher-Smith, M., Walters, J., Artemiou, A., Holmans, P., O'Donovan, M. C., Owen, M. J., Pocklington, A., & Escott-Price, V. (2019). Predictive modeling of schizophrenia from genomic data: Comparison of polygenic risk score with kernel support vector machines approach. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *180*(1), 80–85. https://doi.org/10.1002/ajmg.b.32705

Voineskos, A. N., Mulsant, B. H., Dickie, E. W., Neufeld, N. H., Rothschild, A. J., Whyte, E. M., Meyers, B. S., Alexopoulos, G. S., Hoptman, M. J., Lerch, J. P., & Flint, A. J. (2020). Effects of Antipsychotic Medication on Brain Structure in Patients with Major Depressive Disorder and Psychotic Features: Neuroimaging Findings in the Context of a Randomized Placebo-Controlled Clinical Trial. *JAMA Psychiatry*, *77*(7), 674–683. https://doi.org/10.1001/jamapsychiatry.2020.0036

Vu, M. A. T., Adalı, T., Ba, D., Buzsáki, G., Carlson, D., Heller, K., Liston, C., Rudin, C., Sohal, V. S., Widge, A. S., Mayberg, H. S., Sapiro, G., & Dzirasa, K. (2018). A shared vision for machine learning in neuroscience. *Journal of Neuroscience*, *38*(7), 1601–1607. https://doi.org/10.1523/JNEUROSCI.0508-17.2018

Wachinger, C., Rieckmann, A., & Pölsterl, S. (2021). Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, *67*. https://doi.org/10.1016/j.media.2020.101879

Wagner, G., Koch, K., Schachtzabel, C., Schultz, C. C., Gaser, C., Reichenbach, J. R., Sauer, H., Bär, K. J., & Schlösser, R. G. (2013). Structural basis of the fronto-thalamic dysconnectivity in schizophrenia: A combined DCM-VBM study. *NeuroImage: Clinical*, *3*, 95–105. https://doi.org/10.1016/j.nicl.2013.07.010

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., Björkegren, J. L. M., Im, H. K., Pasaniuc, B., Rivas, M. A., & Kundaje, A. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, *51*(4), 592–599. https://doi.org/10.1038/s41588-019-0385-z

Walterfang, M., McGuire, P. K., Yung, A. R., Phillips, L. J., Velakoulis, D., Wood, S. J., Suckling, J., Bullmore, E. T., Brewer, W., Soulsby, B., Desmond, P., McGorry, P. D., & Pantelis, C. (2008). White matter volume changes in people who develop psychosis. *British Journal of Psychiatry*, *193*(3), 210–215. https://doi.org/10.1192/bjp.bp.107.043463

Watanabe, K., Stringer, S., Frei, O., Mirkov, M. U., Leeuw, C. de, Polderman, T. J. C., Sluis, S. van der, Andreassen, O. A., Neale, B. M., & Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics 2019 51:9*, *51*(9), 1339–1348. https://doi.org/10.1038/s41588-019-0481-0

Wheeler, D. G., & Harper, C. G. (2007). Localised reductions in gyrification in the posterior cingulate: Schizophrenia and controls. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *31*(2), 319–327. https://doi.org/10.1016/j.pnpbp.2006.09.009

White, T., & Hilgetag, C. C. (2011). Gyrification and neural connectivity in schizophrenia. *Development and Psychopathology*, *23*(1), 339–352. https://doi.org/10.1017/S0954579410000842

Witthaus, H., Brüne, M., Kaufmann, C., Bohner, G., Özgürdal, S., Gudlowski, Y., Heinz, A., Klingebiel, R., & Juckel, G. (2008). White matter abnormalities in subjects at ultra high-risk for schizophrenia and first-episode schizophrenic patients. *Schizophrenia Research*, *102*(1–3), 141–149. https://doi.org/10.1016/j.schres.2008.03.022

Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A. E., Dudbridge, F., & Middeldorp, C. M. (2014). Research Review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, *55*(10), 1068–1087. https://doi.org/10.1111/jcpp.12295

Xiao, Y., Yan, Z., Zhao, Y., Tao, B., Sun, H., Li, F., Yao, L., Zhang, W., Chandan, S., Liu, J., Gong, Q., Sweeney, J. A., & Lui, S. (2019). Support vector machine-based classification of first episode drug-naïve schizophrenia patients and healthy controls using structural MRI. *Schizophrenia Research*, *214*, 11–17.

https://doi.org/10.1016/j.schres.2017.11.037

Yang, H., Liu, J., Sui, J., Pearlson, G., & Calhoun, V. D. (2010). A Hybrid Machine Learning Method for Fusing fMRI and Genetic Data: Combining both Improves Classification of Schizophrenia. *Frontiers in Human Neuroscience*, *4*(October), 192. https://doi.org/10.3389/fnhum.2010.00192

Yijun Sun, Todorovic, S., & Goodison, S. (2010). Local-Learning-Based Feature Selection for High-Dimensional Data Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1610–1626. https://doi.org/10.1109/TPAMI.2009.190

Yotter, Rachel A., Nenadic, I., Ziegler, G., Thompson, P. M., & Gaser, C. (2011). Local cortical surface complexity maps from spherical harmonic reconstructions. *NeuroImage*, *56*(3), 961–973. https://doi.org/10.1016/j.neuroimage.2011.02.007

Yotter, Rachel A., Thompson, P. M., & Gaser, C. (2011). Algorithms to Improve the Reparameterization of Spherical Mappings of Brain Surface Meshes. *Journal of Neuroimaging*, *21*(2), 1–14. https://doi.org/10.1111/j.1552-6569.2010.00484.x

Yotter, Rachel Aine, Dahnke, R., Thompson, P. M., & Gaser, C. (2011). Topological correction of brain surface meshes using spherical harmonics. *Human Brain Mapping*, *32*(7), 1109–1124. https://doi.org/10.1002/hbm.21095

Yung, A. R., Yung, A. R., Pan Yuen, H., Mcgorry, P. D., Phillips, L. J., Kelly, D., Dell'olio, M., Francey, S. M., Cosgrave, E. M., Killackey, E., Stanford, C., Godfrey, K., & Buckby, J. (2005). Mapping the Onset of Psychosis: The Comprehensive Assessment of At-Risk Mental States. *Australian & New Zealand Journal of Psychiatry*, *39*(11–12), 964–971. https://doi.org/10.1080/j.1440-1614.2005.01714.x

Zanetti, M. V., Schaufelberger, M. S., Doshi, J., Ou, Y., Ferreira, L. K., Menezes, P. R., Scazufca, M., Davatzikos, C., & Busatto, G. F. (2013). Neuroanatomical pattern classification in a population-based sample of first-episode schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *43*, 116–125. https://doi.org/10.1016/j.pnpbp.2012.12.005

Zarogianni, E., Storkey, A. J., Borgwardt, S., Smieskova, R., Studerus, E., Riecher-Rössler, A., & Lawrie, S. M. (2019). Individualized prediction of psychosis in subjects with an at-risk mental state. *Schizophrenia Research*, *214*, 18–23. https://doi.org/10.1016/j.schres.2017.08.061

Zarogianni, E., Storkey, A. J., Johnstone, E. C., Owens, D. G. C., & Lawrie, S. M. (2017). Improved individualized prediction of schizophrenia in subjects at familial high risk, based on neuroanatomical data, schizotypal and neurocognitive features. *Schizophrenia Research*, *181*, 6–12. https://doi.org/10.1016/j.schres.2016.08.027

Zhang, T., & Davatzikos, C. (2013). Optimally-Discriminative Voxel-Based Morphometry significantly increases the ability to detect group differences in schizophrenia, mild cognitive impairment, and Alzheimer's disease. *NeuroImage*, *79*, 94–110. https://doi.org/10.1016/j.neuroimage.2013.04.063

Zhou, S. Y., Suzuki, M., Hagino, H., Takahashi, T., Kawasaki, Y., Nohara, S., Yamashita, I., Seto, H., & Kurachi, M. (2003). Decreased volume and increased asymmetry of the anterior limb of the internal capsule in patients with schizophrenia. *Biological Psychiatry*, *54*(4), 427–436. https://doi.org/10.1016/S0006-3223(03)00007-6

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zwicker, A., Denovan-Wright, E. M., & Uher, R. (2018). Gene-environment interplay in the etiology of psychosis. *Psychological Medicine*, *48*(12), 1925–1936. https://doi.org/10.1017/S003329171700383X

Appendix 1

A1.Table 1. Statistical significance of the balanced accuracy (BAC) for each bootstrapped sample and each tested combination of regional feature type [i.e., regional-based grey (ROIGM) or white (ROIWM) matter volume; or surface-based regional cortical thickness, gyrification, sulci and complexity indexes (ROISurface)], no feature selection and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; or 5-fold CV]. The statistical significance of the balanced accuracy in each bootstrapped sample was evaluated through permutation testing at a significance level of 5% (i.e., FDR-corrected $p < .05$).

	ROIGM			**ROIWM**			**ROISurface**		
	BAC (%)	***p***	**FDR-*p***	**BAC** (%)	***p***	**FDR-*p***	**BAC** (%)	***p***	**FDR-*p***
LSO CV scheme									
Bootstrapped sample 1	48	.718	.919	43	.827	.919	54	.245	.919
Bootstrapped sample 2	57	.252	.756	59	.077	.400	50	.562	.843
Bootstrapped sample 3	57	.146	.876	50	.559	.897	43	.917	.981
Bootstrapped sample 4	54	.300	.942	54	.314	.942	48	.713	.962
Bootstrapped sample 5	63	.020*	.120	52	.374	.673	63	.012*	.120
LPO CV scheme									
Bootstrapped sample 1	50	.640	.919	43	.859	.919	57	.162	.919
Bootstrapped sample 2	50	.650	.897	48	.698	.897	50	.552	.843
Bootstrapped sample 3	52	.369	.897	57	.238	.897	61	.028*	.504
Bootstrapped sample 4	59	.115	.942	52	.426	.962	52	.459	.962
Bootstrapped sample 5	46	.761	.913	54	.294	.673	52	.348	.673
5-fold CV scheme									
Bootstrapped sample 1	52	.358	.919	54	.297	.919	57	.157	.919
Bootstrapped sample 2	57	.059	.400	46	.760	.912	61	.099	.400
Bootstrapped sample 3	46	.803	.964	48	.698	.897	50	.576	.897
Bootstrapped sample 4	39	.957	.962	46	.819	.962	48	.621	.962
Bootstrapped sample 5	50	.545	.818	57	.166	.598	54	.256	.673

*p < .05, **p < .01, ***p < .001.

A1.Table 2. Statistical significance of the balanced accuracy for each bootstrapped sample and each tested combination of regional feature type [i.e., regional-based grey (ROIGM) or white (ROIWM) matter volume; or surface-based regional cortical thickness, gyrification, sulci and complexity indexes (ROISurface)], forward feature selection and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; or 5-fold CV]. The statistical significance of the balanced accuracy in each bootstrapped sample was evaluated through permutation testing at a significance level of 5% (i.e., FDR-corrected $p < .05$).

	ROIGM			**ROIWM**			**ROISurface**		
	BAC (%)	***p***	**FDR-*p***	**BAC** (%)	***p***	**FDR-*p***	**BAC** (%)	***p***	**FDR-*p***
LSO CV scheme									
Bootstrapped sample 1	52	.438	.919	41	.908	.919	43	.919	.919
Bootstrapped sample 2	52	.340	.843	37	.998	>.999	50	.549	.843
Bootstrapped sample 3	50	.538	.897	54	.302	.897	52	.977	.981
Bootstrapped sample 4	43	.850	.962	48	.816	.962	43	.959	.962
Bootstrapped sample 5	52	.470	.769	46	.826	.929	67	.016*	.120
LPO CV scheme									
Bootstrapped sample 1	46	.813	.919	43	.891	.919	43	.873	.919
Bootstrapped sample 2	54	.381	.843	37	>.999	>.999	61	.111	.400
Bootstrapped sample 3	59	.114	.876	48	.683	.897	48	.681	.897
Bootstrapped sample 4	65	.059	.942	54	.281	.942	57	.193	.942
Bootstrapped sample 5	54	.336	.673	41	.970	.970	59	.088	.396
5-fold CV scheme									
Bootstrapped sample 1	41	.915	.919	46	.735	.919	63	.065	.919
Bootstrapped sample 2	41	.898	>.999	52	.431	.843	67	.004**	.072
Bootstrapped sample 3	37	.981	.981	48	.678	.897	48	.647	.897
Bootstrapped sample 4	41	.962	.962	50	.658	.962	50	.509	.962
Bootstrapped sample 5	43	.918	.970	50	.613	.849	48	.713	.913

*p < .05, **p < .01, ***p < .001.

A1.Table 3. Statistical significance of the balanced accuracy for each bootstrapped sample and each tested combination of voxel-wise feature type [i.e., voxel-based grey (VBGM) or white (VBWM) matter volume maps], principal component analysis and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; or 5-fold CV]. The statistical significance of the balanced accuracy in each bootstrapped sample was evaluated through permutation testing at a significance level of 5% (i.e., FDR-corrected $p < .05$).

	VBGM			**VBWM**		
	BAC (%)	***p***	**FDR-*p***	**BAC** (%)	***p***	**FDR-*p***
LSO CV scheme						
Bootstrapped sample 1	52	.324	.599	52	.011*	.066
Bootstrapped sample 2	48	.895	.979	50	.826	.979
Bootstrapped sample 3	46	.859	.979	50	.669	.979
Bootstrapped sample 4	43	.956	.978	50	.278	.978
Bootstrapped sample 5	43	.968	.968	46	.785	.968
LPO CV scheme						
Bootstrapped sample 1	63	.985	.985	52	.284	.599
Bootstrapped sample 2	46	.979	.979	54	.670	.979
Bootstrapped sample 3	48	.969	.979	48	.669	.979
Bootstrapped sample 4	54	.978	.978	50	.641	.978
Bootstrapped sample 5	46	.938	.968	54	.942	.968
5-fold CV scheme						
Bootstrapped sample 1	41	.399	.599	46	.764	.917
Bootstrapped sample 2	37	.336	.979	59	.089	.534
Bootstrapped sample 3	41	.686	.979	43	.979	.979
Bootstrapped sample 4	41	.528	.978	41	.908	.978
Bootstrapped sample 5	43	.322	.968	46	.779	.968

*p < .05, **p < .01, ***p < .001.

A1.Table 4. Statistical significance of the difference between the testing (i.e., from the inner CV cycle) and the validation (i.e., from the outer CV cycle) balanced accuracy for each tested combination of regional feature type [i.e., regional-based grey (ROIGM) or white (ROIWM) matter volume; or surface-based regional cortical thickness, gyrification, sulci and complexity indexes (ROISurface)], feature selection [i.e., no feature selection (NoFS) or forward feature selection (FFS)] and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; or 5-fold CV]. The difference between the testing and validation BAC was considered statistically significant at a level of 5% (i.e., $p < .05$). Effects sizes were computed as the Cohen's *d* between validation and testing balanced accuracies.

	ROIGM		**ROIWM**		**ROISurface**	
	p	*d*	*p*	*d*	*p*	*d*
LSO CV scheme						
NoFS	.296	0.10	.002**	-0.21	.052	-0.44
FFS	.001***	-1.34	.019*	-0.19	.007**	-0.59
LPO CV scheme						
NoFS	.903	-0.11	< .001***	-0.21	.095	-0.34
FFS	< .001***	-1.50	.043*	-0.04	< .001***	-1.27
5-fold CV scheme						
NoFS	.552	-0.15	.003**	-0.16	.936	0.03
FFS	< .001***	-1.44	.110	-0.16	< .001***	-1.00

*p < .05, **p < .01, ***p < .001.

A1.Table 5. Statistical significance of the difference between the testing (i.e., from the inner CV cycle) and the validation (i.e., from the outer CV cycle) balanced accuracy for each tested combination of voxel-wise feature type [i.e., voxel-based grey (VBGM) or white (VBWM) matter volume maps], principal component analysis and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; or 5-fold CV]. The difference between the testing and validation BAC was considered statistically significant at a level of 5% (i.e., $p < .05$). Effects sizes were computed as the Cohen's *d* between validation and testing balanced accuracies.

	VBGM		**VBWM**	
	p	*d*	*p*	*d*
LSO CV scheme	.198	-0.45	.004**	-0.12
LPO CV scheme	.018*	-0.53	.267	-0.25
5-fold CV scheme	.004**	-0.13	.050*	-0.43

*p < .05, **p < .01, ***p < .001.

A1.Table 6. Interpretability analysis (feature importance and correlation analysis) of the surface-based regional cortical thickness features selected by one of the two best classification models (best model 1). Pearson correlation was computed between each feature value and clinical assessments tools at baseline [i.e., the global assessment functioning (GAF) and comprehensive assessments of at-risk mental state (CAARMS)].

Brain region	**Feature importance**	**GAF at baseline**			**CAARMS at baseline**		
		r	**p**	**FDR p**	**r**	**p**	**FDR p**
Left frontal pole	-6.32	-.07	.667	.967	.22	.173	.797
Left lingual gyrus	0.79	.28	.060	.735	.02	.920	.956
Left middle temporal gyrus	14.45	.12	.415	.967	-.07	.653	.906
Left postcentral gyrus	-1.68	-.01	.967	.967	.14	.380	.797
Left precuneus cortex	1.87	.11	.479	.967	.16	.319	.797
Left superior parietal cortex	0.23	.07	.644	.967	-.14	.380	.797
Left transverse temporal cortex	5.00	-.01	.927	.967	-.15	.370	.797
Right banks superior temporal sulcus	-2.32	.01	.942	.967	.18	.272	.797
Right entorhinal cortex	8.26	.15	.316	.967	.04	.793	.933
Right frontal pole	3.90	.04	.792	.967	.17	.307	.797
Right insula	5.26	.07	.642	.967	.17	.308	.797
Right medial orbital frontal cortex	-3.30	.01	.965	.967	-.11	.516	.895
Right paracentral lobule	-0.92	-.01	.927	.967	-.18	.254	.797
Right pars orbitalis	1.11	-.14	.340	.967	.08	.622	.906
Right pars triangularis	-4.92	.05	.738	.967	.14	.391	.797
Right pericalcarine cortex	3.12	.05	.746	.967	-.10	.523	.895
Right supramarginal gyrus	2.80	.15	.322	.967	.26	.102	.797

A1.Table 7. Interpretability analysis (feature importance and correlation analysis) of the surface-based regional gyrification index features selected by one of the two best classification models (best model 1). Pearson correlation was computed between each feature value and clinical assessments tools at baseline [i.e., the global assessment functioning (GAF) and comprehensive assessments of at-risk mental state (CAARMS)].

Brain region	**Feature importance**	**GAF at baseline**			**CAARMS at baseline**		
		r	**p**	**FDR p**	**r**	**p**	**FDR p**
Left caudal middle frontal gyrus	0.59	.23	.117	.967	.07	.683	.906
Left cuneus cortex	1.27	.15	.313	.967	-.20	.206	.797
Left entorhinal cortex	-8.23	.06	.683	.967	-.17	.284	.797
Left fusiform gyrus	-1.68	.09	.572	.967	.06	.719	.930
Left isthmus–cingulate cortex	6.77	.27	.069	.735	-.15	.363	.797
Left medial orbital frontal cortex	-185.82	.09	.568	.967	-.07	.663	.906
Left pars triangularis	-1.22	.37	.010*	.553	.20	.212	.797
Right caudal anterior-cingulate cortex	9.93	.11	.448	.967	.03	.840	.938
Right caudal middle frontal gyrus	24.67	.14	.367	.967	-.03	.867	.938
Right fusiform gyrus	-7.74	-.05	.739	.967	-.24	.132	.797
Right inferior parietal cortex	-6.56	-.03	.842	.967	.12	.460	.895
Right lateral occipital cortex	-37.83	-.06	.695	.967	.05	.740	.933
Right lateral orbital frontal cortex	-8.12	.29	.049*	.735	.15	.353	.797
Right rostral middle frontal gyrus	53.11	.03	.858	.967	.01	.944	.962

*p < .05, **p < .01, ***p < .001.

A1.Table 8. Interpretability analysis (feature importance and correlation analysis) of the surface-based regional sulci depth features selected by one of the two best classification models (best model 1). Pearson correlation was computed between each feature value and clinical assessments tools at baseline [i.e., the global assessment functioning (GAF) and comprehensive assessments of at-risk mental state (CAARMS)].

Brain region	**Feature importance**	**GAF at baseline**			**CAARMS at baseline**		
		r	**p**	**FDR p**	**r**	**p**	**FDR p**
Left lateral occipital cortex	-8.67	-.07	.642	.967	-.04	.800	.933
Left lingual gyrus	8.69	-.06	.671	.967	-.09	.596	.906
Left medial orbital frontal cortex	1.94	.04	.778	.967	.08	.633	.906
Left parahippocampal gyrus	18.70	.05	.735	.967	-.04	.810	.933
Left pars orbitalis	0.50	-.01	.964	.967	.04	.792	.933
Left pericalcarine cortex	-54.13	.07	.664	.967	-.19	.228	.797
Left posterior-cingulate cortex	9.87	-.12	.438	.967	.16	.339	.797
Left rostral anterior cingulate cortex	-4.33	-.17	.250	.967	-.11	.509	.895
Left rostral middle frontal gyrus	-2.80	-.12	.437	.967	-.25	.116	.797
Left superior parietal cortex	-2.29	.12	.413	.967	.00	.984	.984
Right lateral orbital frontal cortex	2.00	-.01	.941	.967	-.09	.566	.906
Right paracentral lobule	-2.07	-.01	.941	.967	.02	.893	.946
Right precentral gyrus	-3.98	.03	.840	.967	-.03	.857	.938
Right rostral anterior cingulate cortex	1.73	-.06	.670	.967	.22	.173	.797
Right rostral middle frontal gyrus	-4.58	-.03	.855	.967	.08	.631	.906
Right temporal pole	1.52	-.05	.731	.967	-.18	.277	.797

A1.Table 9. Interpretability analysis (feature importance and correlation analysis) of the surface-based regional complexity index features selected by one of the two best classification models (best model 1). Pearson correlation was computed between each feature value and clinical assessments tools at baseline [i.e., the global assessment functioning (GAF) and comprehensive assessments of at-risk mental state (CAARMS)].

Brain region	**Feature importance**	**GAF at baseline**			**CAARMS at baseline**		
		r	**p**	**FDR p**	**r**	**p**	**FDR p**
Left banks superior temporal sulcus	-71.98	-.01	.961	.967	-.12	.475	.895
Left caudal anterior-cingulate cortex	1.95	.31	.037	.735	.15	.356	.797
Left inferior parietal cortex	-2.02	-.03	.832	.967	-.26	.105	.797
Left isthmus–cingulate cortex	9.68	.08	.587	.967	-.17	.288	.797
Right banks superior temporal sulcus	6.45	.16	.274	.967	.07	.683	.906
Right cuneus cortex	2.00	-.01	.939	.967	-.23	.147	.797

A1.Table 10. Interpretability analysis (feature importance and correlation analysis) of the left surface-based regional cortical thickness features selected by one of the two best classification models (best model 2). Pearson correlation was computed between each feature value and clinical assessments tools at baseline [i.e., the global assessment functioning (GAF) and comprehensive assessments of at-risk mental state (CAARMS)].

Brain region	**Feature importance**	**GAF at baseline**			**CAARMS at baseline**		
		r	**p**	**FDR p**	**r**	**p**	**FDR p**
Banks superior temporal sulcus	5.05	-.02	.883	.994	.08	.633	.934
Caudal anterior-cingulate cortex	32.58	-.13	.395	.954	-.14	.421	.841
Caudal middle frontal gyrus	8.78	.07	.637	.954	.01	.945	.982
Cuneus cortex	-9.51	-.01	.946	.994	.05	.750	.934
Frontal pole	8.35	.06	.670	.954	-.12	.493	.895
Insula	2.66	-.18	.223	.954	.22	.187	.766
Isthmus–cingulate cortex	-13.10	-.01	.964	.994	-.27	.105	.766
Lateral orbital frontal cortex	2.62	-.06	.669	.954	.18	.285	.766
Lingual gyrus	-0.33	-.15	.325	.954	-.09	.610	.934
Middle temporal gyrus	9.63	-.06	.681	.954	.03	.867	.966
Paracentral lobule	13.08	-.06	.678	.954	-.05	.757	.934
Parahippocampal gyrus	-16.34	-.16	.303	.954	-.05	.786	.947
Pars orbitalis	24.18	-.10	.502	.954	.17	.322	.793
Pars triangularis	11.70	-.03	.854	.994	.15	.381	.793
Precentral gyrus	47.92	.04	.786	.965	-.05	.749	.934
Precuneus cortex	26.22	.00	.986	.994	.02	.918	.966
Rostral anterior cingulate cortex	8.95	-.09	.566	.954	-.08	.638	.934
Rostral middle frontal gyrus	0.33	-.01	.939	.994	.10	.550	.934
Superior parietal cortex	10.49	.12	.433	.954	.05	.751	.934
Superior temporal gyrus	10.05	-.06	.685	.954	.21	.213	.766
Temporal pole	16.14	-.05	.734	.954	-.02	.897	.966

A1.Table 11. Interpretability analysis (feature importance and correlation analysis) of the right surface-based regional cortical thickness features selected by one of the two best classification models (best model 2). Pearson correlation was computed between each feature value and clinical assessments tools at baseline [i.e., the global assessment functioning (GAF) and comprehensive assessments of at-risk mental state (CAARMS)].

Brain region	**Feature weight**	**GAF at baseline**			**CAARMS at baseline**		
		r	**p**	**FDR p**	**r**	**p**	**FDR p**
Banks superior temporal sulcus	15.21	-.10	.512	.954	.15	.362	.793
Caudal anterior-cingulate cortex	-.24	-.15	.309	.954	.20	.232	.766
Caudal middle frontal gyrus	4.01	.06	.672	.954	.07	.701	.934
Cuneus cortex	15.97	-.01	.952	.994	.09	.589	.934
Fusiform gyrus	-29.57	-.11	.484	.954	-.02	.912	.966
Inferior parietal cortex	-3.77	.15	.306	.954	.07	.681	.934
Inferior temporal gyrus	-.04	-.03	.827	.994	.20	.227	.766
Insula	9.40	-.29	.050*	.954	.23	.179	.766
Isthmus–cingulate cortex	-15.11	-.10	.517	.954	-.11	.501	.895
Lingual gyrus	3.51	-.20	.176	.954	-.19	.272	.766
Medial orbital frontal cortex	-1.64	-.05	.753	.954	.07	.684	.934
Middle temporal gyrus	8.30	.02	.884	.994	.18	.288	.766
Parahippocampal gyrus	-29.48	-.06	.684	.954	.12	.481	.895
Pars opercularis	-11.40	-.20	.180	.954	.12	.494	.895
Pars triangularis	-18.27	-.12	.409	.954	.20	.225	.766
Postcentral gyrus	11.61	.03	.843	.994	.03	.842	.966
Posterior-cingulate cortex	11.21	-.09	.542	.954	.02	.886	.966
Precuneus cortex	1.55	-.02	.878	.994	.08	.630	.934
Rostral anterior cingulate cortex	-5.79	-.08	.610	.954	.16	.349	.793
Superior temporal gyrus	-2.36	-.05	.717	.954	.38	.020*	.466
Supramarginal gyrus	4.93	.09	.532	.954	.21	.222	.766
Transverse temporal cortex	25.75	-.04	.772	.965	.05	.748	.934

*p < .05, **p < .01, ***p < .001.

A1.Table 12. Interpretability analysis (feature importance and correlation analysis) of the left surface-based regional gyrification index features selected by one of the two best classification models (best model 2). Pearson correlation was computed between each feature value and clinical assessments tools at baseline [i.e., the global assessment functioning (GAF) and comprehensive assessments of at-risk mental state (CAARMS)].

Brain region	**Feature importance**	**GAF at baseline**			**CAARMS at baseline**		
		r	**p**	**FDR p**	**r**	**p**	**FDR p**
Banks superior temporal sulcus	-11.38	.20	.181	.954	-.16	.354	.793
Caudal anterior-cingulate cortex	16.96	.15	.315	.954	-.05	.754	.934
Cuneus cortex	9.77	-.09	.571	.954	.12	.496	.895
Entorhinal cortex	4.91	.10	.512	.954	.08	.626	.934
Fusiform gyrus	-7.58	.15	.334	.954	-.19	.271	.766
Inferior parietal cortex	-32.20	-.25	.099	.954	-.12	.468	.895
Inferior temporal gyrus	-4.87	-.02	.910	.994	-.32	.054	.597
Insula	24.16	.05	.721	.954	-.28	.095	.766
Isthmus–cingulate cortex	13.24	.30	.046*	.954	-.06	.732	.934
Lateral occipital cortex	-12.84	.00	.974	.994	-.19	.266	.766
Lateral orbital frontal cortex	-12.98	-.06	.713	.954	-.17	.325	.793
Lingual gyrus	3.31	.08	.609	.954	-.18	.273	.766
Medial orbital frontal cortex	-39.57	-.08	.600	.954	-.08	.654	.934
Parahippocampal gyrus	8.16	.05	.749	.954	.00	.988	.994
Pars opercularis	4.41	.08	.603	.954	-.25	.133	.766
Pars triangularis	-5.13	.18	.242	.954	-.26	.121	.766
Postcentral gyrus	-1.55	.18	.241	.954	.01	.973	.993
Posterior-cingulate cortex	11.65	.27	.065	.954	-.24	.158	.766
Precentral gyrus	-7.46	.13	.382	.954	-.23	.172	.766
Precuneus cortex	3.68	.00	.994	.994	.05	.753	.934
Rostral anterior cingulate cortex	0.34	.10	.517	.954	.15	.374	.793
Rostral middle frontal gyrus	-6.65	-.09	.531	.954	-.22	.197	.766
Superior frontal gyrus	1.25	.11	.464	.954	-.36	.027	.466
Supramarginal gyrus	-5.61	.05	.717	.954	-.10	.572	.934
Temporal pole	12.14	-.11	.487	.954	-.02	.894	.966
Transverse temporal cortex	-2.04	.12	.445	.954	-.03	.839	.966

*p < .05, **p < .01, ***p < .001.

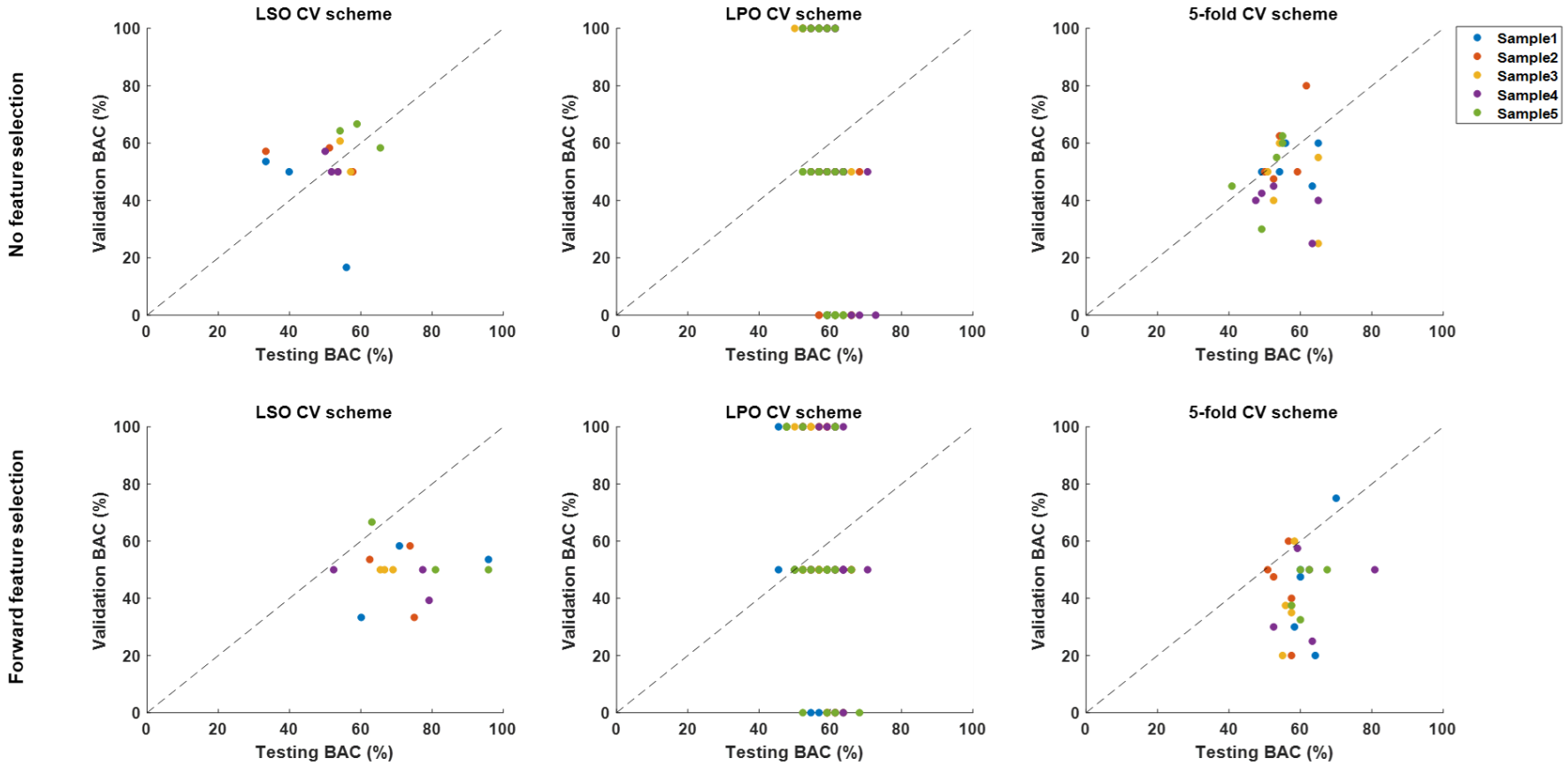
A1.Table 13. Interpretability analysis (feature importance and correlation analysis) of the right surface-based regional gyrification index features selected by one of the two best classification models (best model 2). Pearson correlation was computed between each feature value and clinical assessments tools at baseline [i.e. the global assessment functioning (GAF) and comprehensive assessments of at-risk mental state (CAARMS)].

Brain region	**Feature importance**	**GAF at baseline**			**CAARMS at baseline**		
		r	**p**	**FDR p**	**r**	**p**	**FDR p**
Banks superior temporal sulcus	-17.13	-.09	.563	.954	-.07	.664	.934
Caudal anterior-cingulate cortex	28.71	.18	.233	.954	-.09	.612	.934
Entorhinal cortex	-15.94	-.31	.037*	.954	-.22	.196	.766
Frontal pole	2.91	.15	.307	.954	-.10	.573	.934
Fusiform gyrus	-1.32	.02	.905	.994	-.15	.370	.793
Inferior parietal cortex	36.18	.09	.531	.954	-.35	.036*	.501
Inferior temporal gyrus	-7.81	.05	.733	.954	-.36	.028*	.466
Isthmus–cingulate cortex	13.18	.33	.024*	.954	-.03	.864	.966
Lateral occipital cortex	-31.77	-.06	.712	.954	.01	.953	.982
Lateral orbital frontal cortex	-7.58	.04	.791	.965	-.25	.136	.766
Lingual gyrus	18.76	.11	.474	.954	-.26	.115	.766
Medial orbital frontal cortex	16.97	.16	.277	.954	-.04	.800	.953
Paracentral lobule	13.65	.13	.380	.954	-.39	.018*	.466
Pars opercularis	17.35	.07	.635	.954	-.16	.356	.793
Pars orbitalis	-5.62	.01	.968	.994	-.40	.014*	.466
Pars triangularis	41.82	.10	.523	.954	-.41	.011*	.466
Pericalcarine cortex	4.65	.12	.418	.954	-.24	.150	.766
Posterior-cingulate cortex	-8.44	.08	.599	.954	.00	.994	.994
Precentral gyrus	-45.60	.15	.331	.954	-.19	.254	.766
Precuneus cortex	-7.89	.29	.050*	.954	-.06	.711	.934
Rostral anterior cingulate cortex	-3.01	.06	.678	.954	.20	.230	.766
Rostral middle frontal gyrus	-11.59	.05	.738	.954	-.34	.040*	.501
Superior frontal gyrus	-1.79	.05	.752	.954	-.23	.180	.766
Superior temporal gyrus	-28.67	.05	.754	.954	-.27	.105	.766
Temporal pole	-19.11	.13	.374	.954	-.05	.785	.947

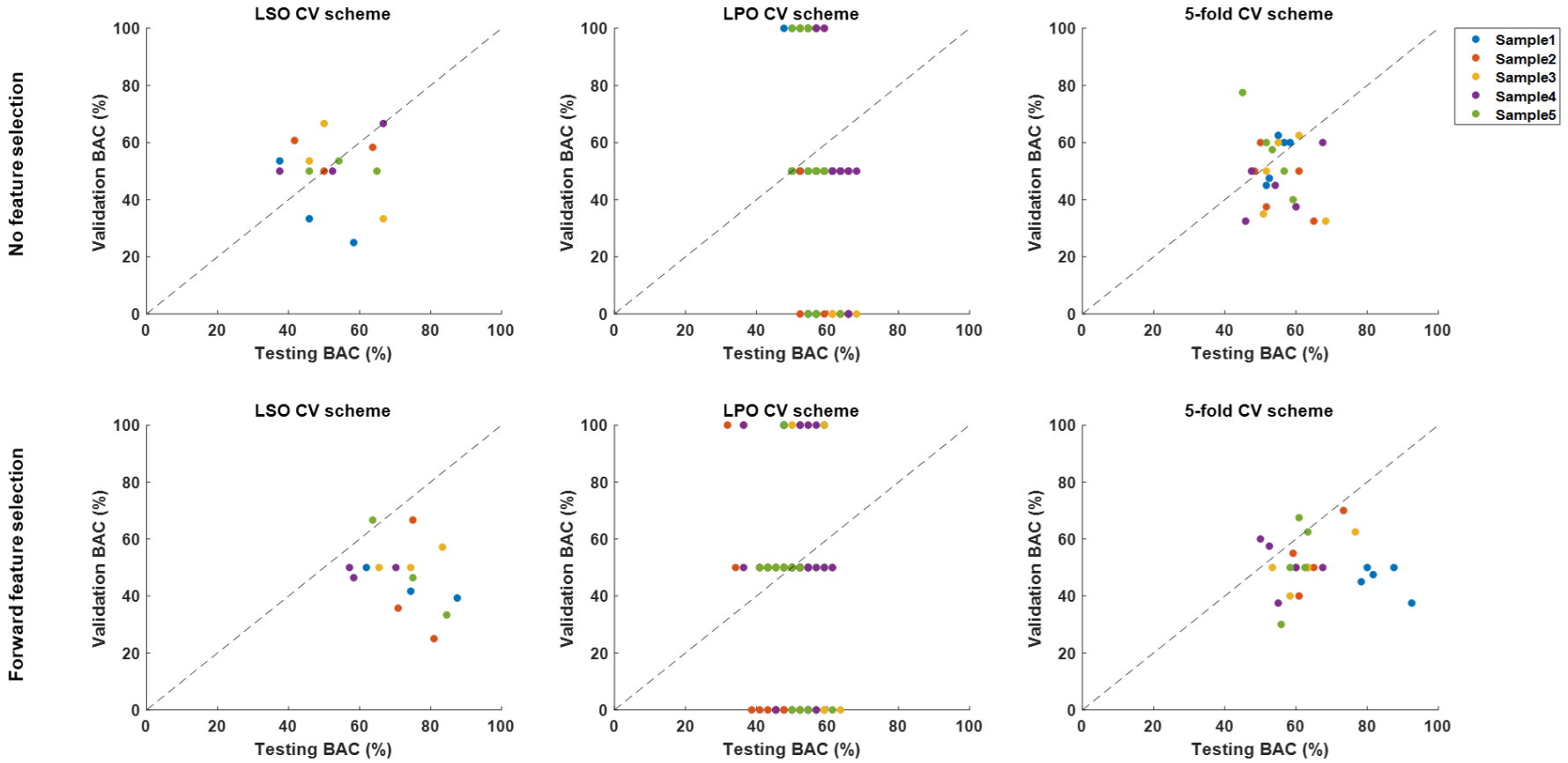
*p < .05, **p < .01, ***p < .001.

A1.Table 14. Interpretability analysis (feature importance and correlation analysis) of the surface-based regional sulci depth features selected by one of the two best classification models (best model 2). Pearson correlation was computed between each feature value and clinical assessments tools at baseline [i.e., the global assessment functioning (GAF) and comprehensive assessments of at-risk mental state (CAARMS)].

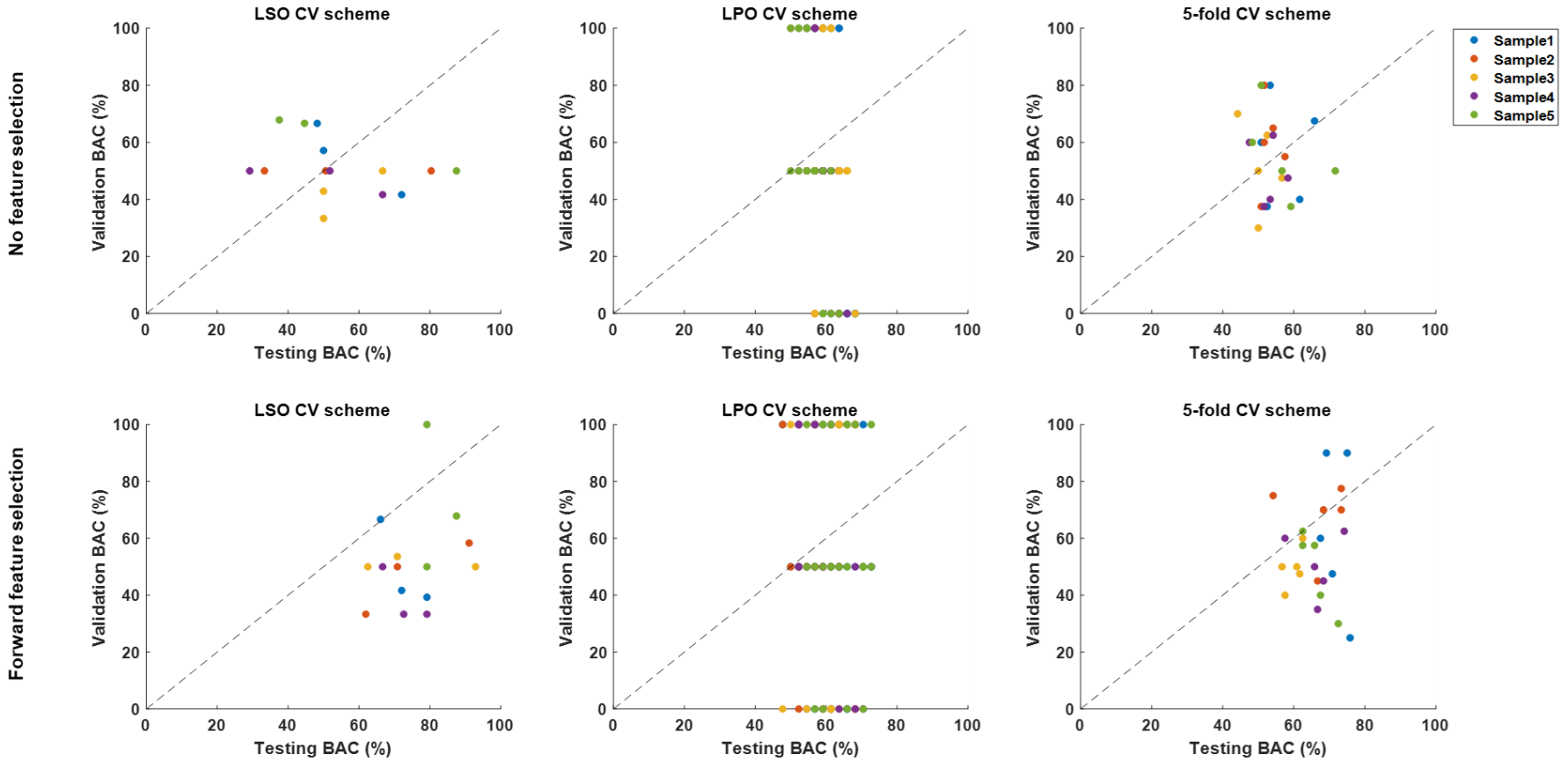
Brain region	**Feature importance**	**GAF at baseline**			**CAARMS at baseline**		
		r	**p**	**FDR p**	**r**	**p**	**FDR p**
Left banks superior temporal sulcus	-2.74	-.05	.745	.954	.15	.391	.797
Left caudal anterior-cingulate cortex	6.59	.16	.300	.954	.09	.595	.934
Left caudal middle frontal gyrus	4.61	.07	.630	.954	-.18	.291	.766
Right caudal anterior-cingulate cortex	-1.46	.14	.357	.954	.03	.838	.966
Right cuneus cortex	-12.57	-.01	.952	.994	-.15	.380	.793
Right inferior parietal cortex	3.95	.00	.979	.994	.02	.909	.966



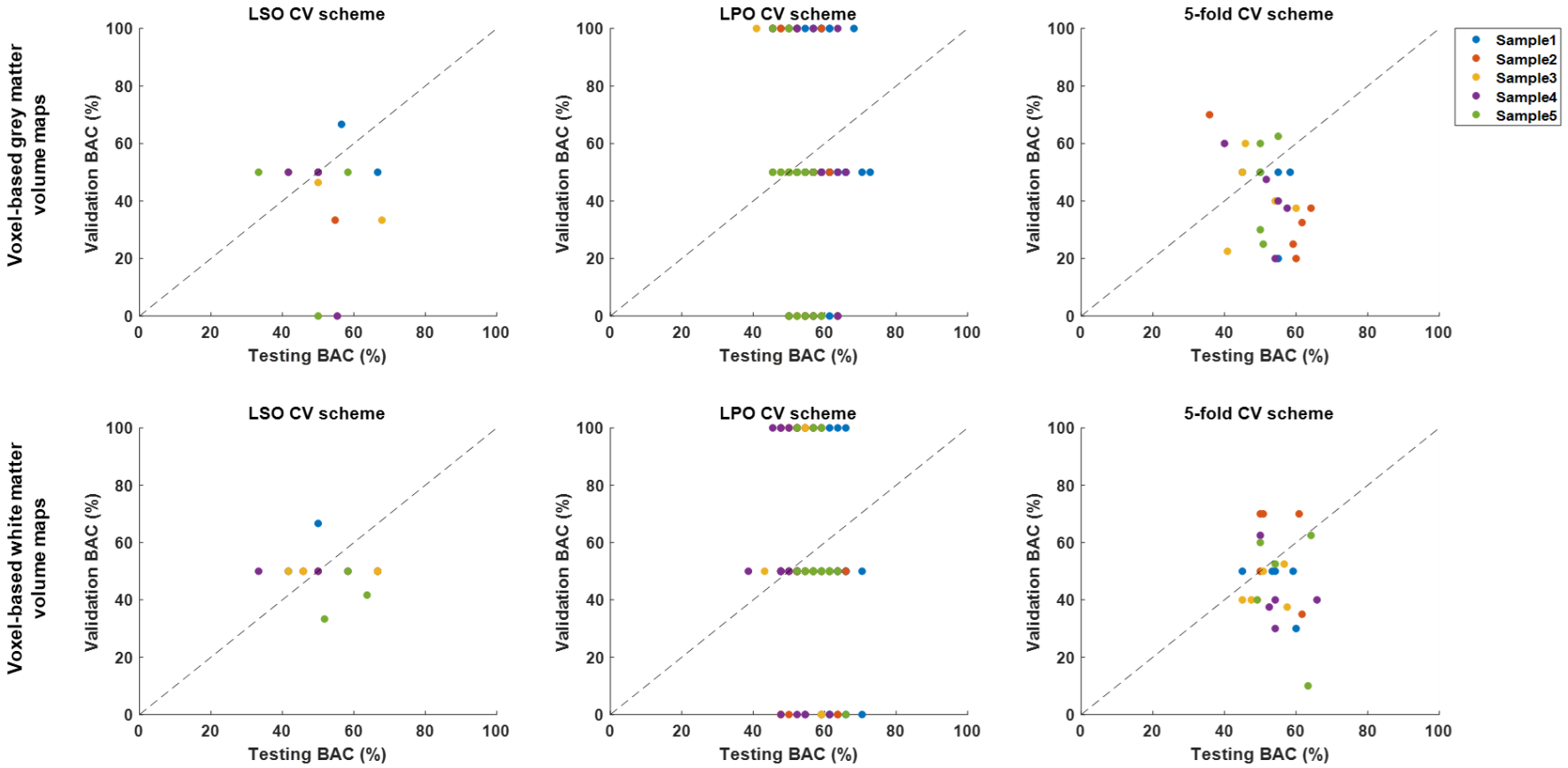
A1.Figure 1. Testing versus validation balanced accuracies (BAC) of classification models trained with regional-based grey matter volumes in combination with feature selection (i.e., no feature selection; or forward feature selection) and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; and 5-fold CV] for each bootstrapped samples (i.e., Sample1 to Sample5).



A1.Figure 2. Testing versus validation balanced accuracies (BAC) of classification models trained with regional-based white matter volumes in combination with feature selection [i.e., no feature selection (top); or forward feature selection (bottom)] and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; and 5-fold CV] for each bootstrapped samples (i.e., Sample1 to Sample5).



A1.Figure 3. Testing *versus* validation balanced accuracies (BAC) of classification models trained with surface-based regional cortical thickness, gyrification, sulci and complexity indexes in combination with feature selection [i.e., no feature selection (top); or forward feature selection (bottom)] and cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; and 5-fold CV] for each bootstrapped samples (i.e., Sample1 to Sample5).



A1.Figure 4. Testing versus validation balanced accuracies (BAC) of classification models trained with voxel-based grey (top) or white (bottom) matter volume maps in combination feature dimensionality reduction through principal component analysis with cross-validation (CV) scheme [i.e., leave-one scan acquisition protocol-out (LSO) CV; leave-one per group-out (LPO) CV; and 5-fold CV] for each bootstrapped samples (i.e., Sample1 to Sample5).

Appendix 2

A2.Table 1. Number of subjects at an at-risk mental state (ARMS) per self-reported ethnicity and prognosis (i.e. transition to psychosis, ARMS-T, or remission of symptoms, ARMS-NT) with genome-wide genotyped data.

Self-reported ethnicity	ARMS-T	ARMS-NT
White (n=77)	21	56
Black (n=41)	16	25
Asian (n=6)	2	4
Mixed (n=9)	2	7

A2.Table 2. Interpretability analysis (feature importance and correlation analysis) of the SNPs selected by the best classification models. Univariate analysis of variance was computed between each SNP's genotype and clinical assessments tools at baseline [i.e., the global assessment functioning (GAF) and comprehensive assessments of at-risk mental state (CAARMS)].

		GAF at baseline				CAARMS at baseline			
SNP	**Feature importance**	**F**	**η^2**	**p**	**FDR p**	**F**	**η^2**	**p**	**FDR p**
rs7013471	1.35×10^{-02}	0.24	.00	.625	.868	0.10	.00	.747	.979
rs6504751	9.90×10^{-03}	0.45	.01	.506	.868	0.51	.01	.478	.922
rs7601756	1.29×10^{-03}	1.85	.03	.179	.796	0.16	.00	.687	.974
rs11257841	6.66×10^{-04}	0.47	.01	.497	.868	3.33	.05	.073	.922
rs11257842	5.71×10^{-04}	0.47	.01	.497	.868	3.33	.05	.073	.922
rs7911557	4.12×10^{-04}	0.47	.01	.497	.868	3.33	.05	.073	.922
rs876983	3.13×10^{-04}	0.34	.00	.560	.868	1.65	.03	.204	.922
rs4875095	1.00×10^{-04}	0.32	.00	.575	.868	0.05	.00	.832	.979
rs2163308	1.00×10^{-04}	0.32	.00	.575	.868	0.05	.00	.832	.979
rs1036555	1.00×10^{-04}	0.32	.00	.575	.868	0.05	.00	.832	.979
rs4875096	1.00×10^{-04}	0.32	.00	.575	.868	0.05	.00	.832	.979
rs4875339	1.00×10^{-04}	0.32	.00	.575	.868	0.05	.00	.832	.979
rs2625068	7.72×10^{-05}	1.82	.03	.181	.796	0.02	.00	.902	.979
rs1384582	7.71×10^{-05}	1.82	.03	.181	.796	0.02	.00	.902	.979
rs2850219	7.71×10^{-05}	1.82	.03	.181	.796	0.02	.00	.902	.979
rs16975694	6.51×10^{-05}	1.73	.02	.193	.796	2.62	.04	.111	.922
rs7077935	5.69×10^{-05}	0.66	.01	.418	.829	0.02	.00	.880	.979
rs4692706	4.61×10^{-05}	0.42	.01	.521	.868	0.03	.00	.871	.979
rs2135777	4.61×10^{-05}	0.42	.01	.521	.868	0.03	.00	.871	.979
rs6823933	4.58×10^{-05}	0.42	.01	.521	.868	0.03	.00	.871	.979
rs6841955	4.58×10^{-05}	0.42	.01	.521	.868	0.03	.00	.871	.979
rs6840742	4.58×10^{-05}	0.42	.01	.521	.868	0.03	.00	.871	.979
rs2840190	2.64×10^{-05}	0.93	.01	.338	.807	1.51	.02	.224	.922
rs16958561	2.64×10^{-05}	0.93	.01	.338	.807	1.51	.02	.224	.922
rs9445537	2.58×10^{-05}	0.05	.00	.824	.875	1.40	.02	.241	.922
rs13204628	2.58×10^{-05}	0.05	.00	.824	.875	1.40	.02	.241	.922
rs9445536	2.58×10^{-05}	0.05	.00	.824	.875	1.40	.02	.241	.922
rs9453289	2.57×10^{-05}	0.05	.00	.824	.875	1.40	.02	.241	.922
rs9445535	2.56×10^{-05}	0.05	.00	.824	.875	1.40	.02	.241	.922
rs13196401	2.34×10^{-05}	0.17	.00	.683	.868	0.35	.01	.555	.922
rs13195937	2.34×10^{-05}	0.17	.00	.683	.868	0.35	.01	.555	.922
rs10477042	2.17×10^{-05}	0.15	.00	.702	.868	0.71	.01	.402	.922
rs966088	2.16×10^{-05}	0.05	.00	.816	.875	0.71	.01	.402	.922
rs11167608	2.15×10^{-05}	0.15	.00	.702	.868	0.71	.01	.402	.922
rs7758073	2.09×10^{-05}	0.07	.00	.796	.868	0.44	.01	.509	.922
rs7776330	2.08×10^{-05}	0.07	.00	.796	.868	0.44	.01	.509	.922

rs13192437	2.08 x 10^{-05}	0.07	.00	.796	.868	0.44	.01	.509	.922
rs5025220	2.08 x 10^{-05}	0.07	.00	.796	.868	0.44	.01	.509	.922
rs7766184	2.08 x 10^{-05}	0.07	.00	.796	.868	0.44	.01	.509	.922
rs12662042	2.08 x 10^{-05}	0.07	.00	.796	.868	0.44	.01	.509	.922
rs12663554	2.08 x 10^{-05}	0.07	.00	.796	.868	0.44	.01	.509	.922
rs12664395	2.08 x 10^{-05}	0.07	.00	.796	.868	0.44	.01	.509	.922
rs12663369	2.07 x 10^{-05}	0.07	.00	.796	.868	0.44	.01	.509	.922
rs12662765	2.06 x 10^{-05}	0.07	.00	.796	.868	0.44	.01	.509	.922
rs11755736	2.06 x 10^{-05}	0.07	.00	.796	.868	0.44	.01	.509	.922
rs479070	1.67 x 10^{-05}	0.03	.00	.866	.900	0.02	.00	.896	.979
rs509821	1.65 x 10^{-05}	0.03	.00	.866	.900	0.02	.00	.896	.979
rs575436	1.64 x 10^{-05}	0.03	.00	.866	.900	0.02	.00	.896	.979
rs9847693	1.64 x 10^{-05}	0.03	.00	.866	.900	0.02	.00	.896	.979
rs542297	1.64 x 10^{-05}	0.03	.00	.866	.900	0.02	.00	.896	.979
rs7761614	1.58 x 10^{-05}	0.00	.00	.957	.957	0.42	.01	.520	.922
rs5025221	1.58 x 10^{-05}	0.00	.00	.957	.957	0.42	.01	.520	.922
rs35104491	1.57 x 10^{-05}	0.00	.00	.957	.957	0.42	.01	.520	.922
rs6917254	1.56 x 10^{-05}	0.00	.00	.957	.957	0.42	.01	.520	.922
rs6937736	1.56 x 10^{-05}	0.00	.00	.957	.957	0.42	.01	.520	.922
rs6937950	1.55 x 10^{-05}	0.00	.00	.957	.957	0.42	.01	.520	.922
rs7769945	1.55 x 10^{-05}	0.00	.00	.957	.957	0.42	.01	.520	.922
rs6568674	1.55 x 10^{-05}	0.00	.00	.957	.957	0.42	.01	.520	.922
rs34663994	1.54 x 10^{-05}	0.00	.00	.957	.957	0.42	.01	.520	.922
rs7285549	1.50 x 10^{-05}	0.86	.01	.356	.807	0.28	.00	.597	.922
rs1930961	1.49 x 10^{-05}	0.86	.01	.356	.807	0.28	.00	.597	.922
rs5996945	1.48 x 10^{-05}	0.86	.01	.356	.807	0.28	.00	.597	.922
rs5996946	1.46 x 10^{-05}	0.86	.01	.356	.807	0.28	.00	.597	.922
rs997873	1.44 x 10^{-05}	0.86	.01	.356	.807	0.28	.00	.597	.922
rs6004671	1.43 x 10^{-05}	0.86	.01	.356	.807	0.28	.00	.597	.922
rs12515986	1.20 x 10^{-05}	1.24	.02	.269	.796	0.04	.00	.833	.979
rs9313939	1.20 x 10^{-05}	1.24	.02	.269	.796	0.04	.00	.833	.979
rs2562323	1.19 x 10^{-05}	1.24	.02	.269	.796	0.04	.00	.833	.979
rs17684043	5.06 x 10^{-06}	3.67	.05	.059	.796	0.01	.00	.938	.979
rs78641718	5.06 x 10^{-06}	3.67	.05	.059	.796	0.01	.00	.938	.979
rs12456794	5.05 x 10^{-06}	3.67	.05	.059	.796	0.01	.00	.938	.979
rs4800099	5.05 x 10^{-06}	3.67	.05	.059	.796	0.01	.00	.938	.979
rs17684097	5.03 x 10^{-06}	3.67	.05	.059	.796	0.01	.00	.938	.979
rs17684151	4.99 x 10^{-06}	3.67	.05	.059	.796	0.01	.00	.938	.979
rs79146010	4.96 x 10^{-06}	3.67	.05	.059	.796	0.01	.00	.938	.979
rs1817499	4.92 x 10^{-06}	3.67	.05	.059	.796	0.01	.00	.938	.979
rs80198146	4.88 x 10^{-06}	3.67	.05	.059	.796	0.01	.00	.938	.979
rs74641519	4.84 x 10^{-06}	2.14	.03	.147	.796	0.01	.00	.938	.979
rs1531821	4.80 x 10^{-06}	2.14	.03	.147	.796	0.01	.00	.938	.979
rs1482974	4.77 x 10^{-06}	2.14	.03	.147	.796	0.01	.00	.938	.979
rs7220627	1.32 x 10-16	0.45	.01	.506	.868	0.51	.01	.478	.922

rs1501265	5.02 x 10-17	0.45	.01	.506	.868	0.51	.01	.478	.922
rs7566584	-2.50 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs3732289	-2.58 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs3821299	-2.72 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs13415100	-3.03 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs13415025	-3.55 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs13389662	-4.26 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs7563417	-4.65 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs7562060	-5.09 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs11681245	-5.32 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs3771721	-5.45 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs1126842	-5.58 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs13432	-5.70 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs12617546	-5.82 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs4664301	-5.94 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs4665110	-5.97 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs9784044	-6.06 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs28521037	-6.17 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs1427328	-6.27 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs12692564	-6.38 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs13401087	-6.48 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs11687502	-6.58 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs11681565	-6.67 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs3863924	-6.76 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs34893664	-6.84 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs6432560	-6.85 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs13032500	-6.94 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs34392518	-7.02 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs1427329	-7.10 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs13397597	-7.18 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs12611522	-7.26 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs12620924	-7.33 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs12692563	-7.40 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs7567818	-7.48 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs6432558	-7.55 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs1834761	-7.62 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs34795915	-7.64 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs16822558	-7.69 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs16822556	-7.77 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs4664300	-7.85 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs16844269	-7.93 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs2357532	-8.02 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs2357531	-8.13 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs10194483	-8.25 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs10204867	-8.33 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922

rs12998291	-8.38 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs12692562	-8.52 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs2357529	-8.68 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs10166694	-8.86 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs13022404	-8.88 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs2081722	-9.04 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs2081723	-9.22 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs35196716	-9.28 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs2098976	-9.40 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs6432557	-9.54 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs12623283	-9.55 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs13428679	-9.66 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs7603274	-9.68 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs13403371	-9.71 x 10^{-07}	0.08	.00	.778	.868	0.22	.00	.642	.922
rs331934	-1.57 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs331549	-1.64 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs1069182	-1.75 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs992775	-1.88 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs992773	-2.01 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs331545	-2.14 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs170734	-2.25 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs161952	-2.28 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs161953	-2.30 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs161747	-2.32 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs2136164	-2.34 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs327793	-2.35 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs2617507	-2.37 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs327792	-2.39 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs331918	-2.41 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs326463	-2.42 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs466340	-2.45 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs331929	-2.46 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs161735	-2.47 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs330419	-2.49 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs331931	-2.49 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs692832	-2.50 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs326489	-2.50 x 10^{-06}	1.68	.02	.199	.796	0.38	.01	.541	.922
rs7593614	-3.56 x 10^{-06}	0.16	.00	.690	.868	0.02	.00	.893	.979
rs975567	-6.22 x 10^{-06}	0.77	.01	.382	.808	0.71	.01	.403	.922
rs416707	-6.22 x 10^{-06}	0.81	.01	.370	.808	0.43	.01	.515	.922
rs10049181	-6.80 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs62289571	-6.81 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs6443732	-6.84 x 10^{-06}	1.32	.02	.255	.796	0.00	.00	.960	.979
rs147425366	-6.88 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs16832256	-6.95 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979

rs4855014	-6.98 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs35650441	-7.10 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs1805611	-7.23 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs1806190	-7.35 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs12629509	-7.42 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs13064866	-7.46 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs62289594	-7.55 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs13075474	-7.56 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs4456860	-7.58 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs9853264	-7.61 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs35039934	-7.63 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs6801189	-7.63 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs9859557	-7.66 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs13086738	-7.68 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs918253	-7.68 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs35212830	-7.70 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs4855015	-7.71 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs11710991	-7.72 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs57671339	-7.73 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs9870874	-7.73 x 10^{-06}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs7682890	-1.47 x 10^{-05}	0.71	.01	.401	.808	1.50	.02	.226	.922
rs73244489	-1.47 x 10^{-05}	0.71	.01	.401	.808	1.50	.02	.226	.922
rs996223	-1.47 x 10^{-05}	0.71	.01	.401	.808	1.50	.02	.226	.922
rs17470919	-1.48 x 10^{-05}	0.71	.01	.401	.808	1.50	.02	.226	.922
rs3749503	-1.48 x 10^{-05}	0.71	.01	.401	.808	1.50	.02	.226	.922
rs12629507	-1.49 x 10^{-05}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs13100468	-1.49 x 10^{-05}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs13071836	-1.49 x 10^{-05}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs2047563	-1.49 x 10^{-05}	1.13	.02	.291	.796	1.66	.03	.203	.922
rs17574399	-1.49 x 10^{-05}	0.71	.01	.401	.808	1.50	.02	.226	.922
rs13065466	-1.49 x 10^{-05}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs7646443	-1.49 x 10^{-05}	1.13	.02	.291	.796	1.66	.03	.203	.922
rs9647341	-1.50 x 10^{-05}	1.13	.02	.291	.796	1.66	.03	.203	.922
rs11716190	-1.50 x 10^{-05}	1.13	.02	.291	.796	1.66	.03	.203	.922
rs2661535	-1.51 x 10^{-05}	0.71	.01	.401	.808	1.50	.02	.226	.922
rs7654182	-1.52 x 10^{-05}	0.71	.01	.401	.808	1.50	.02	.226	.922
rs2456928	-1.53 x 10^{-05}	0.71	.01	.401	.808	1.50	.02	.226	.922
rs10021919	-1.54 x 10^{-05}	0.71	.01	.401	.808	1.50	.02	.226	.922
rs17471024	-1.55 x 10^{-05}	1.06	.01	.307	.796	1.35	.02	.250	.922
rs28702422	-1.55 x 10^{-05}	1.06	.01	.307	.796	1.35	.02	.250	.922
rs6814046	-1.56 x 10^{-05}	0.71	.01	.401	.808	1.50	.02	.226	.922
rs1836047	-1.71 x 10^{-05}	1.63	.02	.206	.796	1.05	.02	.309	.922
rs4962312	-1.71 x 10^{-05}	1.63	.02	.206	.796	1.05	.02	.309	.922
rs701848	-2.14 x 10^{-05}	1.00	.01	.321	.796	1.05	.02	.310	.922
rs4466298	-2.97 x 10^{-05}	0.70	.01	.404	.808	0.44	.01	.508	.922

rs13105270	-3.88 x 10^{-05}	2.22	.03	.141	.796	1.09	.02	.301	.922
rs1460293	-3.88 x 10^{-05}	2.22	.03	.141	.796	1.09	.02	.301	.922
rs12502763	-3.89 x 10^{-05}	2.22	.03	.141	.796	1.09	.02	.301	.922
rs7673711	-3.89 x 10^{-05}	2.22	.03	.141	.796	1.09	.02	.301	.922
rs12651668	-3.90 x 10^{-05}	2.22	.03	.141	.796	1.09	.02	.301	.922
rs12642828	-3.91 x 10^{-05}	2.22	.03	.141	.796	1.09	.02	.301	.922
rs1936602	-5.08 x 10^{-05}	1.71	.02	.195	.796	0.16	.00	.690	.974
rs1936603	-5.08 x 10^{-05}	1.71	.02	.195	.796	0.16	.00	.690	.974
rs10013658	-5.37 x 10^{-05}	1.28	.02	.263	.796	0.49	.01	.488	.922
rs11733675	-5.38 x 10^{-05}	1.28	.02	.263	.796	0.49	.01	.488	.922
rs2331550	-5.39 x 10^{-05}	1.28	.02	.263	.796	0.49	.01	.488	.922
rs11132799	-5.40 x 10^{-05}	1.28	.02	.263	.796	0.49	.01	.488	.922
rs181904588	-9.72 x 10^{-05}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs2002226	-1.24 x 10^{-04}	0.17	.00	.677	.868	1.63	.03	.206	.922
rs1111060	-1.24 x 10^{-04}	0.17	.00	.677	.868	1.63	.03	.206	.922
rs12635535	-1.35 x 10^{-04}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs11708198	-1.58 x 10^{-04}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs6784620	-1.58 x 10^{-04}	0.98	.01	.325	.796	0.00	.00	.979	.979
rs11682883	-2.33 x 10^{-04}	0.43	.01	.514	.868	2.72	.04	.104	.922
rs11682835	-2.33 x 10^{-04}	0.43	.01	.514	.868	2.72	.04	.104	.922
rs10928203	-2.33 x 10^{-04}	0.43	.01	.514	.868	2.72	.04	.104	.922
rs12637938	-1.07 x 10-03	0.98	.01	.325	.796	0.00	.00	.979	.979
rs4662355	-1.52 x 10-03	0.43	.01	.514	.868	2.72	.04	.104	.922
rs10179288	-2.00 x 10-03	0.26	.00	.612	.868	1.57	.03	.215	.922

Publications

Scientific results from the present thesis were submitted to peer-reviewed publication in the following original articles:

Vânia Tavares, Diana Prata*, Hugo Alexandre Ferreira* (2020) **Comparing SPM12 unified segmentation with CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study**, Journal of Neuroscience Methods, 334, 108565. https://doi.org/10.1016/j.jneumeth.2019.108565. *These authors gave equal contribution.

Vânia Tavares, Joana Monteiro, Evangelous Vassos, Jonathan Coleman, Diana Prata (2021) **Evaluation of Genotype-Based Gene Expression Model Performance: A Cross-Framework and Cross-Dataset Study**, Genes, 12(10):1531. https://doi.org/10.3390/genes12101531

Vânia Tavares, Evangelous Vassos, Andre Marquand, James Stone, Isabel Valli, Gareth Barker, Hugo A. Ferreira, Diana P. Prata (2022) **Prediction of transition to psychosis from an at-risk mental state using structural neuroimaging, genetic and environmental data**, Frontiers in Psychiatry 13:1086038. https://doi.org/10.3389/fpsyt.2022.1086038 [accepted for publication]

Journal of Neuroscience Methods 334 (2020) 108565

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Comparing SPM12 and CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study

Vânia Tavares[a,b,*], Diana Prata[a,c,d,1], Hugo Alexandre Ferreira[a,1]

[a] Instituto de Biofísica e Engenharia Biomédica, Faculdade de Ciências da Universidade de Lisboa, Portugal
[b] Faculdade de Medicina da Universidade de Lisboa, Portugal
[c] Instituto Universitário de Lisboa (ISCTE-IUL), CIS-IUL, Lisboa, Portugal
[d] Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK

ARTICLE INFO

Keywords:
Brain tissue segmentation
SPM12
CAT12
Alzheimer's disease diagnosis
Biomarker
Aging

ABSTRACT

Background: Brain volumes have been used as research biomarkers both in health and in Alzheimer's disease (AD). In order to improve the comparability between studies and aid future analytical software platform choice in the research setting, here we compare two segmentation pipelines of structural brain magnetic resonance imaging(sMRI): the SPM12 toolbox, and a SPM12 add-on, the CAT12 toolbox.

Methods: We segmented 1.5T and 3T T1-weighted sMRI images (from the OASIS-brain database) using both pipelines and compared them in terms of their impact on: 1)the effect of age on the total grey matter(GM) and white matter(WM), and on the hippocampi GM volumes in a healthy sample(n = 238); 2)the effect of AD diagnosis on the same volume measures; and 3)the accuracy of each volume measure detecting diagnosis (100 patients with AD and 78 age- and gender-matched healthy subjects).

Results and comparison between methods: Our results demonstrated that: 1)volume estimates from SPM12 were highly correlated with the ones from CAT12, albeit absolute differences between pipelines were tissue specific; 2)the choice of pipeline modulated the effect of age on all volume measures and of diagnosis on hippocampi GM volumes computed from 3 T data; and 3)pipeline had no impact on the accuracy of any brain volume measure detecting AD diagnosis.

Conclusions: Our findings indicate that other studies should take these pipeline effects on age and AD diagnosis, into account, for improved comparability in previous literature. Additionally, we encourage future studies to use CAT12 as this is a more advanced and computationally efficient brain segmentation tool.

1. Introduction

Segmentation is a key neuroimaging processing step to study brain structure (Despotović et al., 2015). Brain tissue volumes, i.e. grey and white matter volumes, measured using structural magnetic resonance imaging (sMRI) have been widely used as biomarkers in the research context. It is well established that the whole brain volume starts declining in early adulthood and accelerates in advanced normal aging, driven essentially by the atrophy of grey matter (Fjell and Walhovd, 2010; Fotenos et al., 2005; Marcus et al., 2007; Toepper, 2017), with an accelerated degeneration, for example, of the hippocampi (Fraser et al., 2015). Understanding normal aging brain atrophy profiles is indeed a fundamental step to identify specific age-related pathologies, such as the case of Alzheimer's disease (AD) (Fjell and Walhovd, 2010). AD is the most common neurodegenerative disorder characterized by synapse and neuronal loss which leads to macroscopic brain atrophy visible with sMRI (Lane et al., 2018). It has been consistently shown that, when compared to healthy subjects, patients with AD show a global grey matter volume loss that is accelerated with disease progression, with the hippocampi showing an initial atrophy already in the prodromal phase of the disease (Toepper, 2017). In fact, the hippocampi volumetry is considered a key imaging biomarker for AD (Teipel et al., 2013). However, the lack of a standard brain segmentation protocol able to uniformize brain volume measures across research and clinical centers is preventing the transition of these biomarkers into clinical practice, i.e. usable to define the clinical diagnosis criteria. Therefore, there is a need for studies comparing brain segmentation pipelines.

Several well-established automatic preprocessing pipelines designed to process sMRI data and segment brain images are currently available. However, it is not always clear what features represent a real relative

* Corresponding author at: Instituto de Biofísica e Engenharia Biomédica, Faculdade de Ciências da Universidade de Lisboa, Portugal
E-mail address: vstavares@fc.ul.pt (V. Tavares).
[1] These authors contributed equally.

https://doi.org/10.1016/j.jneumeth.2019.108565
Received 6 December 2019; Accepted 18 December 2019
Available online 27 December 2019
0165-0270/

improvement, and findings from existing and future studies using different platforms cannot be interpreted reliably in comparison to each other. To inform future choice of pipeline by researchers and study comparability, in this study we have compared two: the classical unified segmentation integrated in the Statistical Parametric Mapping 12 tool (SPM12) (Ashburner and Friston, 2005), a standard neuroimaging processing tool, and the more recent and advanced segmentation pipeline of the Computational Anatomy Toolbox (CAT12; Christian Gaser 2018, http://www.neuro.uni-jena.de/cat/), an SPM12 add-on. The main difference between these two pipelines is in the way they initialize and update their estimation models for brain tissue classification. SPM12 bases the image segmentation on Tissue Probability Maps (TPM), which represent the prior probability of an image unit (voxel) being either grey or white matter, or non-brain tissue. On the other hand, CAT12 uses TPM only to spatially normalize the image, to perform an initial skull-stripping, and to initialize the segmentation. Then, it uses an adaptive maximum *a posteriori* (AMAP) segmentation approach (hypothesis-free approach) (Rajapakse et al., 1997), for further update of the estimation models for brain tissue classification, accounting also for partial volume effects. This allows for a more precise segmentation than SPM12 by estimating the amount of each brain tissue type – grey or white matter or non-brain tissue – that is present in each image unit. Furthermore, the AMAP approach models the parameters of the estimation as slowly varying spatial functions, which accounts for local intensity variations in the original brain image, including subject specific biological variations of brain tissues. Additionally, CAT12 is also computationally less expensive than SPM12 due to its parallel processing algorithms.

In order to improve the comparability between studies and aid future pipeline choice in the research setting, herein, we processed 1.5 T sMRI data from healthy controls and patients with AD using SPM12 and CAT12 segmentation pipelines. Differences in brain volume measures (i.e. the total grey and white matter volumes, and the left and right hippocampi grey matter volumes) using the segmented images from both pipelines were evaluated in two stages. First, the brain volume measures of healthy subjects were regressed against age and were compared with those of patients with AD. Second, the brain volume measures were tested as predictors of the presence of an AD diagnosis. We chose these brain measures to compare pipelines, given the robust existing evidence for expecting a statistically significant negative association between all brain measures with age in health (Fotenos et al., 2005; Fraser et al., 2015; Marcus et al., 2010) and between grey matter (total and of the hippocampi) with AD diagnosis (Toepper, 2017). Finally, we replicated the above described analyses in an independent 3T sMRI dataset to verify if the results were replicable across datasets from different magnetic field strengths.

2. Materials and methods

2.1. Sample description

Two datasets were selected and downloaded from the Open Access Series of Imaging Studies database (OASIS, https://www.oasis-brains.org/): "1.5T dataset" and "3T dataset". A detailed description of both datasets is given below.

2.1.1. 1.5T dataset

Two samples of subjects from a pool of 316 healthy subjects and 100 patients with AD (cross-sectional OASIS-1, (Marcus et al., 2007)) were defined as follows: "Sample1" with 238 healthy subjects and no statistically significant gender effect on age (2 independent sample *t*-test; *p*-value > .05); and "Sample2" with 100 patients with AD (all subjects available in the original pool) and 78 age- and gender-matched healthy subjects, so that no statistically significant gender and diagnosis effects on age (2 independent sample *t*-test; *p*-value > .05) or gender effect on diagnosis (Pearson Σ^2 test; *p*-value > .05) were present. All subjects had at least one 1.5T sMRI scan available. In cases when there was more than one scan available from a single session, the first sMRI scan was chosen.

2.1.2. 3T dataset

An additional sample, "Sample3", was defined from a pool of 609 healthy subjects and 489 patients with AD (longitudinal OASIS-3, (Marcus et al., 2010)) with the following criteria (see also Table S2 in the Supplementary Material for more details): 1) composed by 100 patients with AD and 78 age- and gender-matched healthy subjects (same number of subjects as in "Sample2"); 2) same number of females and males as in Sample2; 3) no statistically significant gender and diagnosis effects on age (2 independent sample *t*-test; *p*-value > .05) or gender effect on diagnosis (Pearson Σ^2 test; *p*-value > .05) were present (similar to Sample2); and 4) only subjects with at least one 3T sMRI scan were included. In cases when there was more than one scan available from a single session, the first scan was chosen. No sample composed only by healthy subjects (similar to Sample1) was selected due to the difference in the subjects' age range between the OASIS-1 and OASIS-3 (18–96 years and 42–95 years, respectively). A direct comparison of the results obtained by the analysis of these two samples would be impractical.

The statistical testing described above was performed in R (R Core Team, 2018) and the results are shown in Table 1. The subjects' IDs included in each Sample are shown in Table S1 in the Supplementary Material. All subjects were right-handed. Patients were diagnosed with AD if they scored higher than zero in the Clinical Dementia Rating (CDR) (Morris, 1993). Information regarding age and gender is also shown in Table 1.

2.2. Structural magnetic resonance imaging

Structural MRI scans were acquired with three different scanners using a structural T1-weighted (T1w) magnetization prepared rapid gradient-echo (MP-RAGE) protocol: 1.5T Magnetom Vision scanner (Siemens, Erlangen, Germany, voxel size = 1.0 × 1.0 × 1.25 mm^3; field-of-view = 256 × 256 × 128; repetition/echo/inversion times = 9.7 msec/4.0 msec/20.0 msec; flip angle = 10°); 3 T TIM Trio scanner (Siemens, Erlangen, Germany, voxel size = 1.0 × 1.0 × 1.0 mm^3; field-of-view = 176 × 256 × 256; repetition/echo/inversion times = 2.4 msec/3.2 msec/1.0 msec; flip angle = 8°; 125 scans); and 3T BioGraph mMR PET-MR scanner (Siemens, Erlangen, Germany, voxel

Table 1
Sample characteristics.

	Age (years) (mean ± SD, [range])	Female/Male
Sample1		
Healthy Subjects[a] (n = 238)	40.0 ± 22.2 [18, 90]	119/119
Sample2		
Healthy Subjects[b] (n = 78)	76.3 ± 7.6 [62, 94]	54/24
Alzheimer's Disease[b] (n = 100)	76.8 ± 7.1 [62, 96]	59/41
Sample3		
Healthy Subjects[c] (n = 78)	73.5 ± 6.7 [62 89]	54/24
Alzheimer's Disease[c] (n = 100)	74.4 ± 5.9 [64, 89]	59/41

[a] Effect of gender on age: 2 independent sample *t*-test = − 0.05, *p* = .963.
[b] Effect of gender on age: 2 independent sample *t*-test = − 0.58; *p* = .592; Effect of diagnosis on age: 2 independent sample *t*-test = − 0.40; *p* = .690; Effect of gender on the diagnosis: $\chi^2 = 1.56$, *p* = .211.
[c] Effect of gender on age: 2 independent sample *t*-test *t* = 1.2; *p* = .773; Effect of diagnosis on age: 2 independent sample *t*-test = − 1.88; *p* = .062; Effect of gender on the diagnosis: $\chi^2 = 0.01$, *p* = .905.

size = 1.2 × 1.1 × 1.1 mm^3; field-of-view = 176 × 240 × 256; repetition/echo/inversion times = 2.3 msec/3.0 msec/0.9 msec; flip angle = 9°; 53 scans).

2.3. Image processing

T1w 1.5T MRI images were first reoriented to the anterior commissure – posterior commissure line using the Display tool of SPM12 (v6909, http://www.fil.ion.ucl.ac.uk/spm/). Then, both reoriented T1w 1.5T and T1w 3T images were processed, in parallel, with SPM12 and with CAT12 (v1092, http://www.neuro.uni-jena.de/cat/) using default settings and MATLAB (9.1). First, bias field inhomogeneity correction was performed using an algorithm shared by SPM12 and CAT12. Second, images were segmented into grey matter, white matter, and cerebrospinal fluid using a classical unified segmentation approach with SPM12 (Ashburner and Friston, 2005) and the AMAP approach with CAT12 (Rajapakse et al., 1997). Third, the images were spatially normalized to a template derived from 555 healthy subjects of the IXI-database (http://www.brain-development.org) using the Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL) algorithm, which is shared by SPM12 and CAT12. Finally, total relative volumes of grey and white matter were computed by dividing the total volume of grey and white matter (the sum of all voxels classified as grey or white matter, respectively) by the total intracranial volume (TIV; the sum of all voxels classified as grey or white matter or as cerebrospinal fluid). Additionally, the grey matter relative volume of the left and right hippocampi was obtained using the Hammers region-of-interest (ROIs) atlas (Hammers et al., 2003). In particular, the Hammers atlas, which was first normalized to the same template as the segmented grey matter images (i.e. to the IXI-database template) using DARTEL algorithm, was used as a binary mask to select the ROIs corresponding to the left and right hippocampi in the segmented grey matter image. Then, the grey matter relative volume was defined as the sum of all the voxels classified as grey matter inside the respective ROI and divided by the TIV. The segmentation results (i.e. the total relative volume of grey and white matter and of left and right hippocampi's grey matter) are summarized in the Table S3 and illustrated in Fig. S1, in the Supplementary Material. Segmentation values for each subject, their associated brain images, and the scripts used in the analyses are available upon request.

2.4. Statistical analysis

The statistical analysis was divided in 3 main parts. In the first part (Part 1 below), a correlation between pipelines (SPM12 or CAT12)'s brain volume measures is performed using linear regression. The second part (Part 2 below) examines the interaction of pipeline with age and AD diagnosis using linear mixed models (LMMs). In the third part (Part 3 below), using logistic regression, the predictive accuracy of each brain volume measure in detecting the presence of AD was estimated for each pipeline separately, and then compared between them.

2.4.1. Correlation between the pipelines' brain volume measures (Part 1)

The total grey and white matter relative volumes obtained with SPM12 were linearly regressed against the ones obtained with CAT12, using only the healthy subjects from Sample1 and Sample3, separately. Additionally, the coefficient of determination (R^2) was extracted.

2.4.2. Interaction of pipeline with age and Alzheimer's diagnosis on volume measures (Part 2)

2.4.2.1. Effect of age on brain volume measures using SPM12 and CAT12 in healthy subjects (Part 2.1). The interaction of pipeline (SPM12, CAT12) with age on each brain volume measure (i.e. total grey and white matter relative volumes and grey matter relative volume of the right and left hippocampi) was tested with four LMMs in healthy subjects from Sample1, i.e. one for each brain volume measure. In these LMMs, age, pipeline, and 'age by pipeline' interaction were entered in the model as fixed effects, and subject as a random effect.

2.4.2.2. Effect of Alzheimer's disease diagnosis on brain volume measures using SPM12 and CAT12 (Part 2.2). The interaction of pipeline (SPM12, CAT12) with AD diagnosis (AD, healthy subjects) on each brain volume measure was also tested in subjects from Sample2 and Sample3, separately. Diagnosis, pipeline, and 'diagnosis by pipeline' interaction were entered in the LMM model as fixed effects, covarying for age, and subject as a random effect.

2.4.3. Detection of Alzheimer's disease diagnosis from brain volume measures using SPM12 and CAT12 (Part 3)

The predictive ability of each brain volume in detecting AD was tested for each pipeline separately in Sample2 and in Sample3 individually, by logistically regressing each brain volume measure against diagnosis, while controlling for age. The area under the receiver operating characteristic curve was then statistically compared between pipelines using the DeLong method (DeLong et al., 1988).

Overall statistical significance of each effect of interest (pipeline and 'age by pipeline' in Sample1, and 'diagnosis by pipeline' in Sample2 and in Sample3 - the main effect of pipeline was already tested in Sample1) was evaluated using the Log-Likelihood ratio test. All p-values were corrected for multiple testing using False Discovery Rate (FDR) and reported as statistically significant at an FDR-corrected p-value < .05. Cohen's f^2 (Selya et al., 2012) effect size was computed for all pipeline effects (i.e. the main effect and the 'age by pipeline' and 'diagnosis by pipeline' interaction effects). Additionally, the beta coefficients for age in each pipeline is reported as the effect size for the 'age by pipeline' interaction and Cohen's d (Gibbons et al., 1993) as the effect size for the pipeline and 'diagnosis by pipeline' contrasts (i.e. SPM12 *vs.* CAT12 and SPM12 *vs.* CAT12 in AD and in healthy subjects, respectively). Interpretation of Cohen's d was performed using Kristoffer Magnusson web tool (Interpreting Cohen's d effect size, https://rpsychologist.com/d3/cohend/). All the statistical analysis described in this section was done using R (R Core Team, 2018), using the following R packages: 1) 'stats' (R Core Team, 2018) for simple linear and logistic regression, and FDR correction of the statistical testing p-values; 'nlme' (Pinheiro et al., 2018) for LMMs fitting and statistical testing; and 'pROC' (Robin et al., 2011) for the ROC curve analysis.

3. Results

3.1. Correlation between the pipelines' brain volume measures (Part 1)

The linear regression analysis showed that the total grey and white matter relative volumes estimated with SPM12 and CAT12 are strongly correlated [Sample1 - grey matter relative volume: $p <$ .001, $R^2 = 0.90$, Fig. 1A, top left; and white matter relative volume: $p <$ .001, $R^2 = 0.71$, Fig. 1A, top right; Sample3 - grey matter relative volume: $p <$ .001, $R^2 = 0.55$, Fig. 1B, top left; and white matter relative volume: $p <$ .001, $R^2 = 0.50$, Fig. 1B, top right;]. Grey matter relative volume estimations with SPM12 showed higher values than the ones with CAT12, with this difference increasing with higher average total grey matter relative volumes. The opposite effect was shown for white matter relative volume estimations (see Bland-Altman plots in Fig. 1A and B, bottom).

3.2. Interaction of pipeline with age and Alzheimer's disease diagnosis on volume measures (Part 2)

3.2.1. Effect of age on brain volume measures using SPM12 and CAT12 in healthy subjects (Part 2.1)

We found a statistically significant effect of pipeline on all brain volume measures ($p <$ .001), with SPM12 showing lower volume

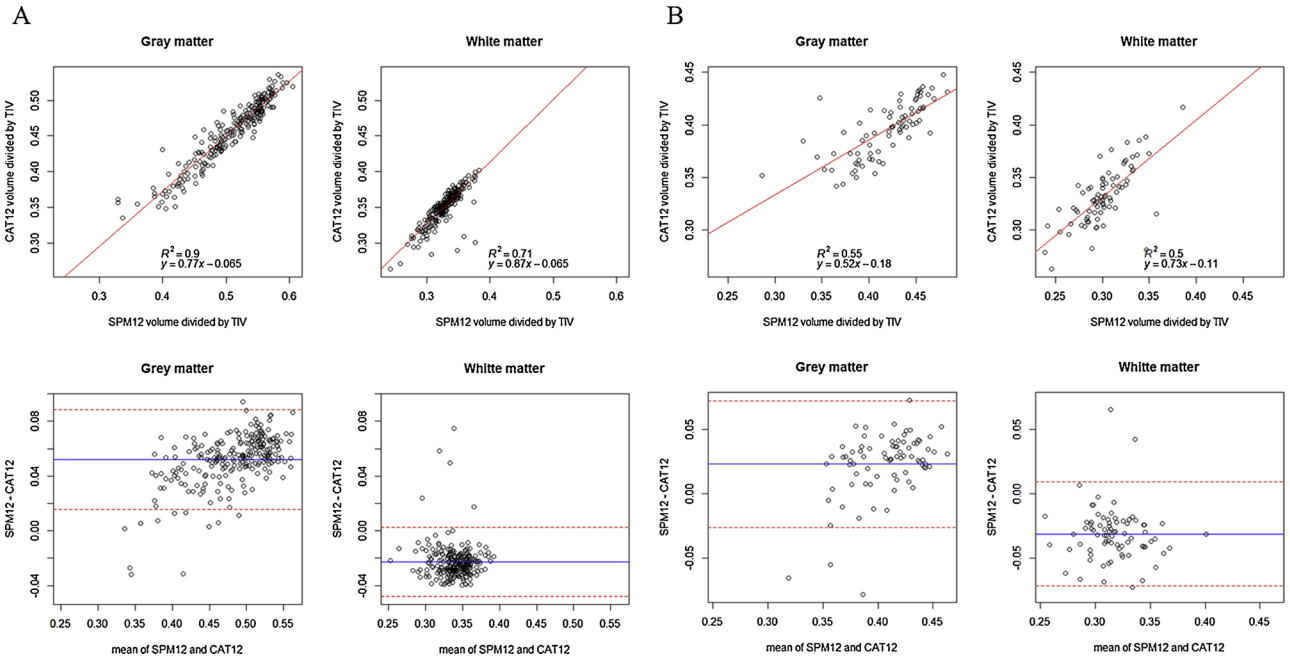


Fig. 1. Top: linear regression analysis between total grey (**left**) and white (**right**) matter volumes obtained with SPM12 and CAT12 using only the healthy subjects from Sample1 (**A**) or Sample3 (**B**). The red line represents the fitted regression line (which equation is represented in the form of CAT12 volume (y) = slope * SPM12 volume (x) + intercept; and effect size is represented by the R^2). **Bottom**: Bland-Altman plots with limits (dashed red lines) of agreement for mean (continuous blue line) total grey (**left**) and white (**right**) matter volumes. TIV: total intracranial volume.

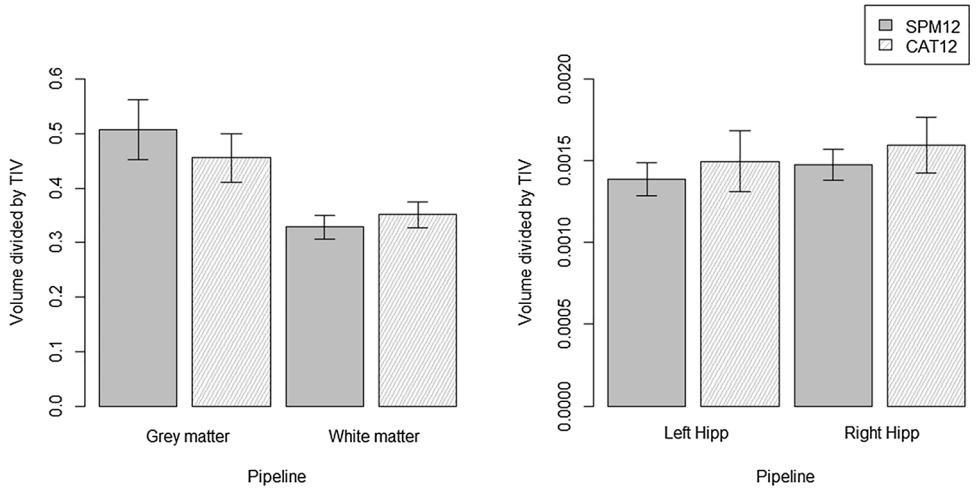


Fig. 2. Mean total grey and white matters volume (**left**), and grey matter volume of hippocampi (**right**) estimated using SPM12 and CAT12 pipelines. Only the healthy subjects from Sample1 were used. Error bars represent the standard deviation from the mean volume. TIV: total intracranial volume.

estimations in total white matter and grey matter of the hippocampi and higher volume estimations in total grey matter (grey matter: $d_{\text{SPM12-CAT12}} = 4.02$; white matter: $d_{\text{SPM12-CAT12}} = -1.25$; left hippocampus: $d_{\text{SPM12-CAT12}} = -1.56$; right hippocampus: $d_{\text{SPM12-CAT12}} = -1.64$; Fig. 2). Most importantly, we found an interaction effect of age and pipeline on all brain measures ($p < .001$), with SPM12 showing a higher decrease in total grey matter relative volume with aging than CAT12 ($\beta_{\text{age,CAT12}} = -0.0018$, $\beta_{\text{age,SPM12}} = -0.0022$); and lower decrease in total white matter and both hippocampi grey matter relative volumes (white matter: $\beta_{\text{age,CAT12}} = -0.25$, $\beta_{\text{age,SPM12}} = -0.17, \beta^2_{\text{age,CAT12}} = -0.24$, $\beta^2_{\text{age,SPM12}} = -0.17$; left hippocampus: $\beta_{\text{age,CAT12}} = -4.7 \times 10^{-6}$, $\beta_{\text{age,SPM12}} = -3.1 \times 10^{-6}$; right hippocampus: $\beta_{\text{age,CAT12}} = -4.2 \times 10^{-6}, \beta_{\text{age,SPM12}} = -2.7 \times 10^{-6}$; Fig. 3). See also Table 2 for statistics and effect sizes.

3.2.2. Effect of Alzheimer's disease diagnosis on brain volume measures using SPM12 and CAT12 (Part 2.2)

We found a statistically significant interaction effect of diagnosis and pipeline on both hippocampi grey matter relative volumes (left: $p = .013$, $d_{\text{CAT12, Alzheimer's Disease (AD) – Healthy Subjects (HS)}} = -0.87$, $d_{\text{SPM12, AD - HS}} = -0.83$; and right: $p = .003$, $d_{\text{CAT12, AD - HS}} = -0.92$, $d_{\text{SPM12, AD – HS}} = -0.85$; Fig. 4) when using Sample3. In particular,

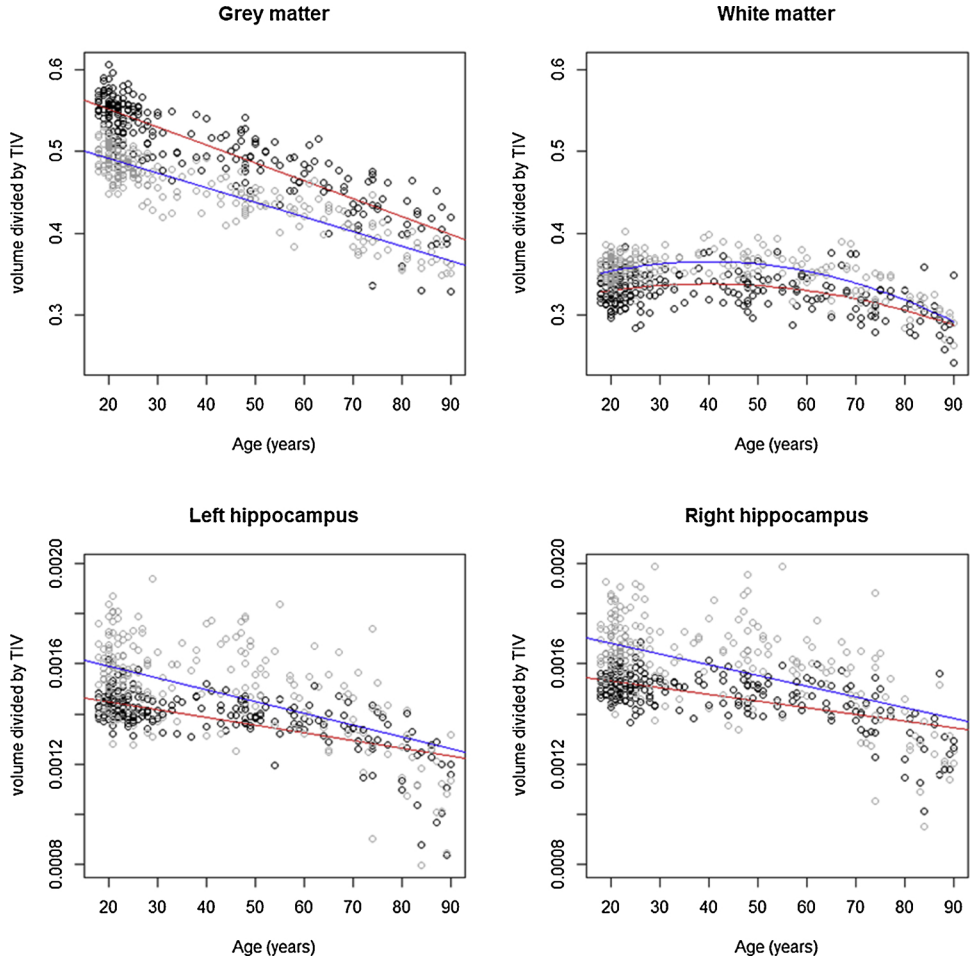


Fig. 3. Scatter plot of the grey (**top left**) and white (**top right**) matter volume estimation and grey matter volume estimation of the left (**bottom left**) and right (**bottom right**) hippocampus using SPM12 (black dots) and CAT12 (grey dots). Fitted trend lines are in red (continuous line) for SPM12 and in blue for CAT12 (dashed line). Only the healthy subjects from Sample1 were used. TIV: total intracranial volume.

patients with AD showed lower volume estimates than healthy subjects in both segmentation pipelines, but with a higher difference when CAT12 was used. The interaction effect was not significant on the total grey and white matter relative volumes when using Sample3, nor on any brain volume measures when using Sample1. See Table 2 for statistics and effect sizes.

3.3. Detection of Alzheimer's disease diagnosis from brain volume measures using SPM12 and CAT12 (Part 3)

All brain volume measures were able to statistically predict AD diagnosis above chance level (AUC $>$ 0.5; Table 3) when controlling for age. This result holds for both pipelines (SPM12 and CAT12) and samples (Sample2 and Sample3). The between-pipeline AUC statistical comparison showed no difference between the volume measures' predictive accuracy, of SPM12 versus CAT12, at a statistical significance level of 5 % (see Table 3 for statistics). The ROC for each test is represented in Fig. 5.

4. Discussion

Aiming to improve the comparability between studies and aid future platform choice in the research setting, we compared two brain image segmentation pipelines, SPM12 and CAT12, by 1) regressing the total grey and white matter relative volumes obtained with SPM12 with those obtained with CAT12; studying the association of total grey and white matter, and grey matter of left and right hippocampi relative volumes with 2) age in a healthy population; and 3) the diagnosis of AD; and 4) the effect of pipeline on the volume measures' estimation of an AD diagnosis (*vs.* healthy subjects).

4.1. Pipeline comparison in volume measurements in healthy subjects (Part 1)

Overall, the SPM12 and CAT12 grey and white matter relative volume estimates are strongly correlated in both datasets (i.e., 1.5T and 3T datasets) as expected. CAT12 total grey matter relative volume estimates are, on average, lower than the one estimated by SPM12, but the opposite is shown for grey matter of hippocampi and total white matter relative volumes. In particular, higher mean total grey matter relative volume estimates showed a higher volume difference between SPM12 and CAT12, whereas the opposite was shown for total white matter relative volume estimates. Our results seem to indicate that CAT12 overestimates (relative to SPM) the total grey matter relative volume and underestimates the total white matter relative volume as the mean volume increases. This finding should be considered when comparing volumetric studies that use SPM12 with the ones that use CAT12. Furthermore, these effects are present in both 1.5T and 3T

Table 2
Effect of age, pipeline, and interaction effect of 'age by pipeline' (SPM12 vs. CAT12) on brain volume measures using Sample1 and effect of diagnosis (patients with Alzheimer's disease (AD) *vs.* healthy subjects (HS)) and interaction effect of 'diagnosis by pipeline' on brain volume measures using Sample2 and Sample3. The overall statistical significance of the effect was tested with the Log-likelihood test (χ^2). Cohen's f^2 effect size was computed for all pipeline effects (i.e. 'age by pipeline' and 'diagnosis by pipeline' interaction effects). Additionally, the beta coefficients for age in each pipeline is reported as the effect size for the 'age by pipeline' interaction and Cohen's *d* as the effect size for the pipeline and 'diagnosis by pipeline' contrasts (i.e. SPM12 *vs.* CAT12 and SPM12 *vs.* CAT12 in AD and in HS, respectively). Statistically significant results are reported with a *p*-value < .05 (all *p*-values are FDR-corrected) and market with an asterisk.

	Grey matter			White matter			Left hippocampus			Right hippocampus		
	χ^2	*p*	Effect size	χ^2	*p*	Effect size	χ^2	*p*		χ^2	*p*	Effect size
Sample1												
pipeline	536.97	< .001*	$f^2 = 0.39$ $d_{SPM12\text{-}CAT12} = 4.02$	337.09	< .001*	$f^2 = 0.95$ $d_{SPM12\text{-}CAT12} = -1.25$	156.62	< .001*	$f^2 = 0.53$ $d_{SPM12\text{-}CAT12} = -1.56$	178.03	< .001*	$f^2 = 0.51$ $d_{SPM12\text{-}CAT12} = -1.64$
age by pipeline	60.18	< .001*	$f^2 = 0.03$ $\beta_{ageCAT12} = -0.0018$ $\beta_{ageSPM12} = -0.0022$	39.32	< .001*	$f^2 = 0.05$ $\beta_{ageCAT12} = -0.25$ $\beta_{ageSPM12} = -0.17$ $\beta_{age^2CAT12} = -0.24$ $\beta_{age^2SPM12} = -0.17$	24.16	< .001*	$f^2 = 0.09$ $\beta_{ageCAT12} = -4.7 \times 10^{-6}$ $\beta_{ageSPM12} = -3.1 \times 10^{-6}$	23.75	< .001*	$f^2 = 0.08$ $\beta_{ageCAT12} = -4.2 \times 10^{-6}$ $\beta_{ageSPM12} = -2.7 \times 10^{-6}$
Sample2												
diagnosis by pipeline	2.70	.134	$f^2 = 0.01$ $d_{CAT12,\ AD\ -\ HS} = -0.66$ $d_{SPM12,\ AD\ -\ HS} = -0.49$	0.07	.789	$f^2 = 0.00$ $d_{CAT12,\ AD\ -\ HS} = -0.63$ $d_{SPM12,\ AD\ -\ HS} = -0.61$	3.90	.134	$f^2 = 0.72$ $d_{CAT12,\ AD\ -\ HS} = -0.91$ $d_{SPM12,\ AD\ -\ HS} = -0.83$	0.65	.134	$f^2 = 0.04$ $d_{CAT12,\ AD\ -\ HS} = -0.89$ $d_{SPM12,\ AD\ -\ HS} = -0.77$
Sample3												
diagnosis by pipeline	0.15	.696	$f^2 = 0.00$ $d_{CAT12,\ AD\ -\ HS} = -0.63$ $d_{SPM12,\ AD\ -\ HS} = -0.51$	0.13	.716	$f^2 = 0.00$ $d_{CAT12,\ AD\ -\ HS} = -0.62$ $d_{SPM12,\ AD\ -\ HS} = -0.68$	6.15	.013*	$f^2 = 0.07$ $d_{CAT12,\ AD\ -\ HS} = -0.87$ $d_{SPM12,\ AD\ -\ HS} = -0.83$	8.93	.003*	$f^2 = 0.16$ $d_{CAT12,\ AD\ -\ HS} = -0.92$ $d_{SPM12,\ AD\ -\ HS} = -0.85$

datasets, suggesting that the magnetic field strength has little impact on the choice of the brain segmentation pipeline.

4.2. Pipeline comparison in the effect of age on brain volume measures in healthy subjects (Part 2.1)

Also as expected, age was negatively correlated with all grey matter measures (i.e. total and hippocampi relative volumes), whereas total white matter relative volume showed an inverted U-shape relationship with age, with an inflexion point at middle age (around 45 years of age). This result is in line with previous findings using the same dataset as us herein (Fotenos et al., 2005) and as more recently reviewed (Fjell and Walhovd, 2010; Toepper, 2017), which showed that aging leads to a decrease in both grey and white matter (although only after the forth decade of life), but with a steeper grey matter atrophy. Importantly, the choice of pipeline explained between 39 % (for total grey matter) and 95 % (for total white matter) of the variance in all volume estimates (i.e. a 'very large' effect size). Roughly 100% of the subjects showed mean SPM12 total grey matter relative volume estimates that were higher than CAT12 (d = 4.02), and 88–95% showed mean CAT12 total white matter and grey matter of the hippocampi relative volume estimates that were higher than SPM12 ($-1.25 < d < -1.64$). Additionally, the interaction of age and pipeline explained 3–9% of the variance in all brain volume measures generally (i.e. a 'small' to 'medium' effect size). Total grey matter relative volume measures estimated by SPM12 showed a steeper decrease with age than the ones estimated by CAT12, whereas the opposite was found for the other three brain measures, i.e. a percentage difference of 20.0 % for total grey matter, 41.0 % and 43.5 % for the grey matter of the left and right hippocampi, respectively, and 38.1 % (β_{age}) and 34.1 % ($\beta_{age}{}^2$) for total white matter (computed as the absolute percentage difference between the SPM12 and CAT12 beta coefficients). Although one cannot establish which segmentation pipeline translates the actual brain volume and aging association as the ground truth since the volume measures are not known, this segmentation platform comparison may be relevant for the appropriate comparison between existing and future findings using these platforms, in particular those of aging. Moreover, it is worth to note that these results hold only for 1.5T MRI scans, as the effect of segmentation pipeline choice to study the association between age and brain volume measures were not tested for using a 3T dataset.

4.3. Pipeline comparison in the effect of Alzheimer's disease diagnosis on brain volume measures (Part 2.2)

Additionally, we compared brain volume measures of patients with AD with those of age- and gender-matched healthy subjects and explored the interaction between diagnosis and pipeline. Across brain volume measures, AD lowered brain volumes as expected, particularly in the hippocampi (Toepper, 2017). Interestingly, the pipeline did not significantly modulate the diagnosis effect when using 1.5T MRI scans, with this interaction explaining roughly 0%–4% of all brain volume measures (i.e. 'small' effect size), except of the left hippocampus where 72 % of its variance was explained by the pipeline choice (i.e. a 'large' effect size). On the other hand, when testing the 3T dataset, the pipeline and diagnosis interaction effect was significant on the hippocampi grey matter relative volumes, explaining 7 and 16 % of its variance (a 'small' and 'medium' effect size, respectively), with SPM12 distinguishing patients with AD from healthy subjects slightly better than CAT12. There was an increase in effect size of 4.7 % and 7.9 % for left and right hippocampi grey matter relative volume, respectively (computed as the absolute percentage difference between the SPM12 and CAT12 Cohen's *d*), resulting in an increase of 1 % and 2 %, respectively, of patients showing mean relative volume estimates lower than healthy subjects when using SPM12, compared to CAT12.

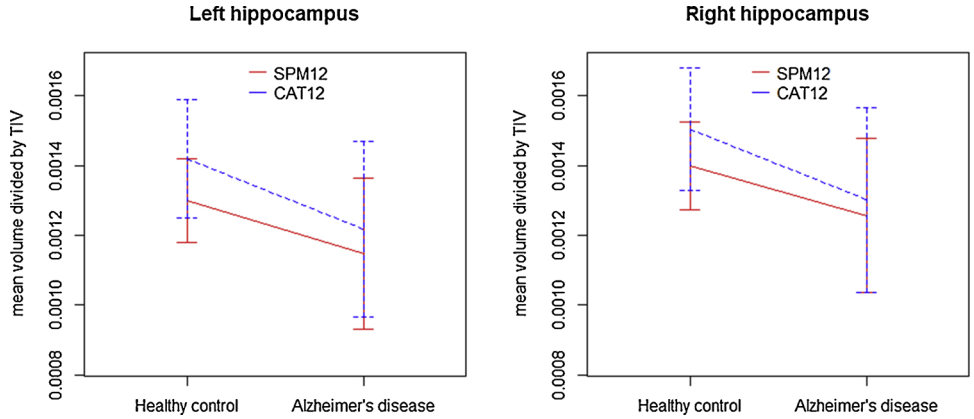


Fig. 4. Mean grey matter volume of left (**left**) and right (**right**) hippocampus for healthy subjects and patients with Alzheimer's disease from Sample3 estimated using SPM12 (continuous red line) and CAT12 (dashed blue line). Error bars represent the standard deviation from the mean volume. TIV: total intracranial volume.

Table 3
Logistic regression results between each brain volume measures and diagnosis (healthy subjects *vs.* patients with AD) using SPM12, CAT12, Sample1 and Sample2. Area under the receiver operating characteristic curve (AUC) is showed for each tested model together with the 95 % confidence interval (CI). The AUC values were statistically compared between pipelines (SPM12 and CAT12) for each sample using the DeLong method.

	$R^2_{Nagelkerke\ SPM12}$	$R^2_{Nagelkerke\ CAT12}$	**SPM12** **AUC** [95% CI]	**CAT12** **AUC** [95% CI]	Statistical comparison AUC_{SPM12} *vs.* AUC_{CAT12}
Sample2					
Grey matter	.15	.19	.71 [.63, .79]	.72 [.65, .80]	$Z = -0.40$, $p = .973$
White matter	.04	.05	.59 [.51, .67]	.60 [.52, .68]	$Z = -0.33$, $p = .973$
Left hippocampus	.27	.27	.76 [.69, .83]	.77 [.70, .84]	$Z = -0.26$, $p = .973$
Right hippocampus	.36	.34	.79 [.73, .86]	.79 [.73, .86]	$Z = -0.03$, $p = .973$
Sample3					
Grey matter	.08	.13	.65 [.57, .73]	.69 [.62, .77]	$Z = -1.53$, $p = .503$
White matter	.12	.13	.68 [.60, .76]	.68 [.60, .75]	$Z = 0.13$, $p = .899$
Left hippocampus	.24	.25	.73 [.66, .80]	.74 [.66, .81]	$Z = -0.21$, $p = .899$
Right hippocampus	.20	.23	.72 [.64, .79]	.74 [.67, .81]	$Z = -0.69$, $p = .899$

4.4. Pipeline comparison in the detection of Alzheimer's disease diagnosis using brain volume measures (Part 3)

Finally, we compared the ability of the brain volumes obtained, with each segmentation pipeline, to distinguish patients with AD from age- and gender-matched healthy subjects. Results were similar when using 1.5T and 3T datasets. Particularly, both pipelines (SPM12 and CAT12) were able to produce volumes detecting AD above chance level, with hippocampi relative volumes showing the highest accuracies, as expected. Indeed, hippocampi volume shows a steeper grey matter atrophy rate than total grey matter already at a prodromal phase of the disease (Toepper, 2017), and it is currently the best-established imaging biomarker (research-wise) for AD (Teipel et al., 2013). Furthermore, the volumes' measure ability to detect AD diagnosis did not differ significantly between pipelines which is in line with the recent demonstration that the prediction power of MRI-based brain volume measures, in particular of hippocampi, does not depend on the measurement method, but mainly on the degree of brain tissue atrophy (Buchert et al., 2018).

4.5. Limitations

Our study has a few limitations that need to be addressed. First, we did not compare our results with the ground truth segmentation method, that currently is defined as manual segmentation usually performed by one or more experts. This is a very time-consuming task and almost impractical when analyzing medium to big sample sizes, as our herein. Second, we could not analyze the age effect modulated by the segmentation pipeline on brain volume measures in a 3T dataset, due to the lack of availability of subjects with a matched age range with the 1.5T dataset. Furthermore, in order to statistically test the modulation effect of the magnetic field strength on the choice of pipeline in brain volume-based studies, a dataset composed by the subjects with both MRI scans ideally (ie. 1.5T and 3T) should be used. Third, we only tested total grey and white matter relative volume measures and grey matter relative volume measures of the hippocampi. Future studies should also investigate the impact of the segmentation pipeline choice on other brain regions, by conducting whole brain voxel-wise morphometry studies.

5. Conclusions

In this study, we attempted to compare the research usefulness, albeit not clinical usefulness, of SPM12 and CAT12 neuroimaging analytical software pipelines. Taken together, our results show that 1) SPM12 and CAT12 brain volume measure differences are tissue-dependent; 2) the choice of segmentation pipeline (i.e. SPM12 or CAT12) modulated the effect of age on all brain tissue volumes and of diagnosis, albeit only on 3T MRI-based hippocampi grey matter volumes, but 3) did not impact the accuracy of the brain tissue volumes in detecting AD diagnosis. Therefore, we encourage future volume-based studies to take into account these pipeline effects when comparing their results to other studies' findings. Furthermore, we also encourage the use of CAT12 when conducting AD studies, as this is a more advanced brain segmentation tool and computationally less expensive than SPM12.

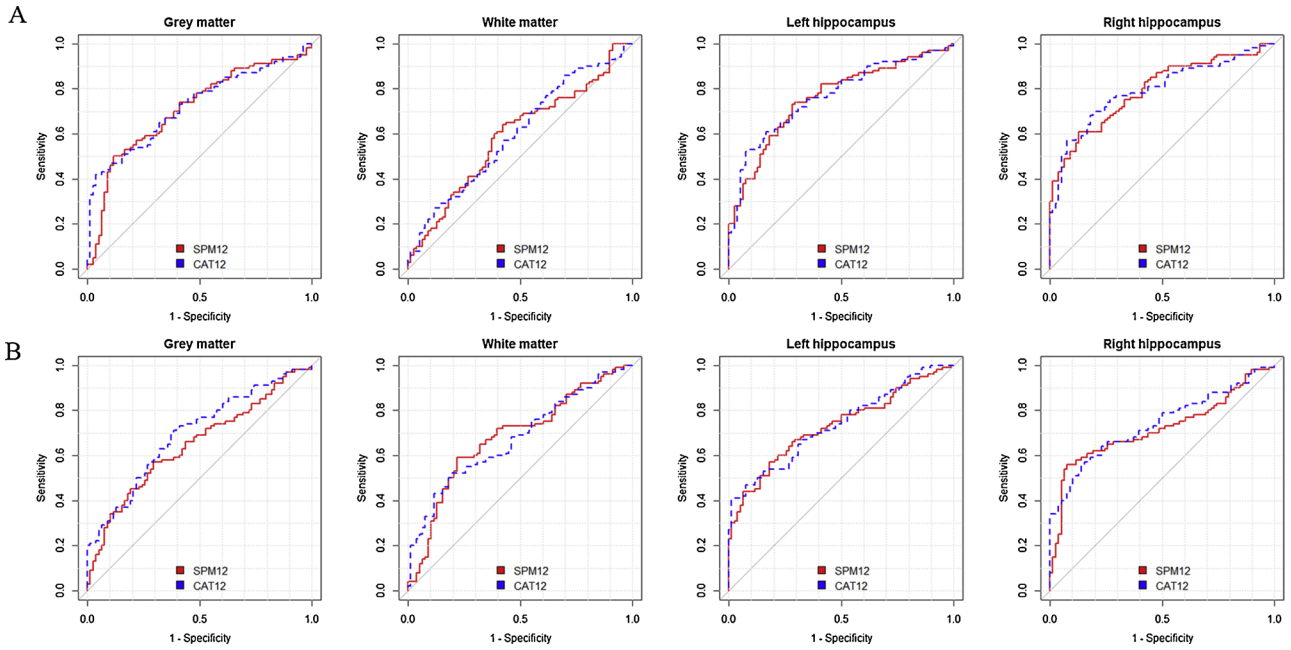


Fig. 5. Receiver characteristic curves for each pipeline (SPM12 in red and CAT12 in blue) when logistically regressing each brain volume measure against diagnosis (healthy subjects *vs.* patients with Alzheimer's Disease). Subjects used in this analysis belong to Sample2 (**A**, top row) or Sample3 (**B**, bottom row).

Funding

This work was supported by the Fundação para Ciência e Tecnologia [grant numbers: PD/BD/114460/2016; IF/00787/2014; PTDC/MEC-PSQ/30907/2017 and DSAIPA/DS/0065/2018]; the European Commission Seventh Framework Programme Marie Curie Career Integration [grant number FP7-PEOPLE-2013-CIG-631952]; the Bial Foundation 2018 grant [Ref. 292/16]; and the data for this study were acquired from the OASIS database [grant numbers P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382].

Author contributions

V. Tavares, D. Prata and H. A. Ferreira developed the study concept, design and interpretation. V. Tavares performed data analysis and interpretation and drafted the manuscript. D. Prata and H. A. Ferreira provided critical revisions to it. All authors approved the final version of the manuscript.

Declaration of Competing Interest

None.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jneumeth.2019.108565.

References

Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26 (3), 839–851. https://doi.org/10.1016/j.neuroimage.2005.02.018.

Buchert, R., Lange, C., Suppa, P., Apostolova, I., Spies, L., Teipel, S., et al., 2018. Magnetic resonance imaging-based hippocampus volume for prediction of dementia in mild cognitive impairment: why does the measurement method matter so little? Alzheimer's Dementia 14 (7), 976–978. https://doi.org/10.1016/j.jalz.2018.03.006.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44 (3), 837–845.

Despotović, I., Goossens, B., Philips, W., 2015. MRI segmentation of the human brain: challenges, methods, and applications. Comput. Math. Methods Med. 2015. https://doi.org/10.1155/2015/450341.

Fjell, A.M., Walhovd, K.B., 2010. Structural brain changes in aging: courses, causes and cognitive consequences. Rev. Neurosci. 21 (3), 187–221. Retrieved from. http://www.ncbi.nlm.nih.gov/pubmed/20879692.

Fotenos, A., Snyder, A., Girton, L., Morris, J., Buckner, R., 2005. Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. Neurology 64 (6), 1032–1039. https://doi.org/10.1212/01.WNL.0000154530.72969.11.

Fraser, M.A., Shaw, M.E., Cherbuin, N., 2015. A systematic review and meta-analysis of longitudinal hippocampal atrophy in healthy human ageing. NeuroImage 112, 364–374. https://doi.org/10.1016/j.neuroimage.2015.03.035.

Gibbons, R.D., Hedeker, D.R., Davis, J.M., 1993. Estimation of effect size from a series of experiments involving paired comparisons. J. Educ. Behav. Stat. 18 (3), 271–279. https://doi.org/10.2307/1165136.

Hammers, A., Allom, R., Koepp, M.J., Free, S.L., Myers, R., Lemieux, L., et al., 2003. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum. Brain Mapp. 19 (4), 224–247. https://doi.org/10.1002/hbm.10123.

Lane, C.A., Hardy, J., Schott, J.M., 2018. Alzheimer's disease. Eur. J. Neurol. 25 (1), 59–70. https://doi.org/10.1111/ene.13439.

Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2010. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. J. Cogn. Neurosci. 22 (12), 2677–2684. https://doi.org/10.1162/jocn.2009.21407.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19 (9), 1498–1507. https://doi.org/10.1162/jocn.2007.19.9.1498.

Morris, J.C., 1993. The clinical dementia rating (CDR): current version and scoring rules. Neurology 43 (11). https://doi.org/10.1212/WNL.43.11.2412-a. 2412–2412.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team, 2018. nlme}: Linear and Nonlinear Mixed Effects Models. Retrieved from. https://cran.r-project.org/package=nlme.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from. https://www.r-project.org/.

Rajapakse, J.C., Giedd, J.N., Rapoport, J.L., 1997. Statistical approach to segmentation of single-channel cerebral mr images. IEEE Trans. Med. Imaging 16 (2), 176–186. https://doi.org/10.1109/42.563663.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf. 12 (1), 77. https://doi.org/10.1186/1471-2105-12-77.

Selya, A.S., Rose, J.S., Dierker, L.C., Hedeker, D., Mermelstein, R.J., 2012. A practical guide to calculating Cohen's f 2, a measure of local effect size, from PROC MIXED. Front. Psychol. 3 (APR), 1–6. https://doi.org/10.3389/fpsyg.2012.00111.

Teipel, S.J., Grothe, M., Lista, S., Toschi, N., Garaci, F.G., Hampel, H., 2013. Relevance of magnetic resonance imaging for early detection and diagnosis of Alzheimer disease. Med. Clin. North Am. 97 (3), 399–424. https://doi.org/10.1016/j.mcna.2012.12.013.

Toepper, M., 2017. Dissociating normal aging from Alzheimer's disease: a view from cognitive neuroscience. J. Alzheimer's Dis. 57 (2), 331–352. https://doi.org/10.3233/JAD-161099.

genes

MDPI

Article

Evaluation of Genotype-Based Gene Expression Model Performance: A Cross-Framework and Cross-Dataset Study

Vânia Tavares [1,2], Joana Monteiro [1,3], Evangelos Vassos [4,5], Jonathan Coleman [4,5,*] and Diana Prata [1,6,*]

1 Instituto de Biofísica e Engenharia Biomédica, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal; vstavares@fc.ul.pt (V.T.); js.monteiro@campus.fct.unl.pt (J.M.)
2 Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisboa, Portugal
3 Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Almada, Portugal
4 Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London SE5 8AF, UK; evangelos.vassos@kcl.ac.uk
5 NIHR Maudsley Biomedical Research Centre, South London and Maudsley NHS Trust, London SE5 8AZ, UK
6 Instituto Universitário de Lisboa (Iscte-IUL), CIS-Iscte, 1749-016 Lisboa, Portugal
* Correspondence: jonathan.coleman@kcl.ac.uk (J.C.); diana.prata@kcl.ac.uk (D.P.); Tel.: +44-(0)-2078480629 (J.C.); +351-217-500-77 (D.P.)

Abstract: Predicting gene expression from genotyped data is valuable for studying inaccessible tissues such as the brain. Herein we present eGenScore, a polygenic/poly-variation method, and compare it with PrediXcan, a method based on regularized linear regression using elastic nets. While both methods have the same purpose of predicting gene expression based on genotype, they carry important methodological differences. We compared the performance of expression quantitative trait loci (eQTL) models to predict gene expression in the frontal cortex, comparing across these frameworks (eGenScore vs. PrediXcan) and training datasets (BrainEAC, which is brain-specific, vs. GTEx, which has data across multiple tissues). In addition to internal five-fold cross-validation, we externally validated the gene expression models using the CommonMind Consortium database. Our results showed that (1) PrediXcan outperforms eGenScore regardless of the training database used; and (2) when using PrediXcan, the performance of the eQTL models in frontal cortex is higher when trained with GTEx than with BrainEAC.

Keywords: expression quantitative trait loci; transcriptome; gene expression; genome wide association study; polygenic score

Citation: Tavares, V.; Monteiro, J.; Vassos, E.; Coleman, J.; Prata, D. Evaluation of Genotype-Based Gene Expression Model Performance: A Cross-Framework and Cross-Dataset Study. *Genes* **2021**, *12*, 1531. https://doi.org/10.3390/genes12101531

Academic Editor: Anelia D. Horvath

Received: 9 August 2021
Accepted: 26 September 2021
Published: 28 September 2021

1. Introduction

The emergence of genome-wide association studies (GWAS) has allowed the identification of associations between thousands of variants (mainly single nucleotide polymorphisms (SNP)) and susceptibility to complex diseases, such as schizophrenia [1]. However, there is still a gap between the variants and their functional role in the diseases' etiologies, in particular in regards to SNPs [2]. Indeed, nearly 90% of these genetic variations occur in non-coding deoxyribonucleic acid (DNA) sequences, and only about 4–5% of plausibly causal variants in GWAS-associated regions are coding variants, which suggests that the main mechanism by which variation in these regions acts is not by altering protein structure. In comparison, about 50% of plausibly causal variants are expression quantitative trait loci (eQTL), suggesting moderation of gene expression is an important mechanism of action [3,4]. As such, it is crucial to consider and efficiently utilize variants correlated with gene expression, i.e., eQTL, to better understand the mechanisms behind the role of specific genes (especially if implicated by the hypothesis-free GWA approach) in intermediate or complex phenotypes [5].

The degree of expression of genes is typically inferred from the transcriptome, i.e., the messenger ribonucleic acid (mRNA) levels of the genes. One of the reasons for the

delayed translation of transcriptome-wide association studies (TWAS) into clinical practice is that a gene expression profile is tissue-specific [6]. This is a crucial factor for a correct clinical interpretation of the eQTLs identified in the TWAS. However, in clinical practice, sampling invasiveness is the most important determinant for decision making regarding tissue sampling [7]. Indeed, measuring the expression of a given gene is invasive for many tissues, including the human brain, requiring postmortem sampling. Therefore, there is an urgent need for accurate statistical methods for the non-invasive estimation of gene expression in tissues where sampling presents more risks than the expected clinical benefit. Recently, efforts have been put forward to compile large-scale concomitant transcriptomic and genomic datasets, i.e., eQTL datasets, such as the Genotype-Tissue Expression (GTEx) project [8] across tissues, and the brain-specific Brain eQTL Almanac (BrainEAC) [9] and CommonMind Consortium (CMC) [10]. Using these emerging eQTL datasets, gene expression can be used as an intermediate molecular phenotype to potentially address the functional gap in GWAS findings and take a much needed step closer to understanding the underlying mechanisms and molecular pathways of complex disorders.

Recently, a gene-based method, PrediXcan, was developed and has been widely used to predict gene expression levels from SNPs [11–14]. In particular, using eQTL data and effect sizes determined through comprehensive eQTL analyses on reference transcriptome datasets, PrediXcan can predict expression levels for the whole transcriptome in multiple tissues [12]. Gamazon and colleagues [12] compared two methods for eQTL-based gene expression prediction: one based on traditional polygenic scoring [15] with one based on a regularized regression analysis using elastic net [16]; they showed that the latter yielded higher correlation between the observed and predicted gene expressions. However, the comparison was potentially biased, since (1) linkage disequilibrium (LD) between variations was not accounted for in the polygenic approach used; and (2) PrediXcan models do not account for individual missing genotypes; i.e., these are simply replaced by zero, thus assuming that that particular SNP makes the same contribution to the predicted gene expression level as a common homozygous genotype. The elastic net method consisted of (1) variable selection (i.e., selecting only SNPs that influence gene expression); and (2) handling highly correlated SNPs (i.e., those in high LD) by balancing their contribution to the variance in gene expression [12]. The polygenic method used consisted of (1) the selection of SNPs influencing gene expression by individually testing the association of each allele with gene expression through a linear regression; and (2) predicting gene expression as a weighted sum (taken from the individual linear regression analysis) of the SNP's alleles showing an association with the observed gene expression below a significance threshold (i.e., $p <$ single top SNP, 1×10^{-4}, 0.001, 0.01, 0.05, 0.5 or 1).

We herein tackled the two above-described limitations, i.e., the inaccurate polygenic method to which PrediXcan was originally compared and the passive incorporation of missing genotypes into the gene expression prediction. We did this by using an improved polygenic method to predict gene expression levels based on genome-wide genotypes, the eGenScore. In particular, we addressed the LD between gene expression-associated SNPs by filtering them out and the missing genotype issue by incorporating an adjustment factor to the weighted sum of SNP alleles based on the expected proportion of those alleles in a standardized population. After addressing the two issues above, we then aimed to compare our improved polygenic method with the PrediXcan elastic net method. Our second aim was to assess how training these tools with the most recent versions of each of the two main transcriptomic and genomic databases available, BrainEAC and GTEx, would affect their performance. To achieve both purposes, we trained eQTL models (which yield eQTL scores as a proxy of gene expression) with both frameworks, eGenScore and PrediXcan, and with each of the two databases, BrainEAC and GTEx, using transcriptomic data (i.e., gene expression levels) from the frontal cortex. We then compared the performance of the eQTL models across different frameworks (i.e., eGenScore vs. PrediXcan) and across databases (i.e., BrainEAC vs. GTEx), using an internal cross-validation approach and an external validation approach by applying the eQTL model to a third database from the

CMC. Although the eGenScore method could theoretically be applied to any tissue type, herein we focused on the frontal cortex, as this tissue is the only one common to the three databases used in this study, i.e., BrainEAC, GTEx, and CMC.

2. Materials and Methods

An overview of the datasets and methods used in this study is represented in Figure 1. All quality control procedures, described below, were performed by the database providers.

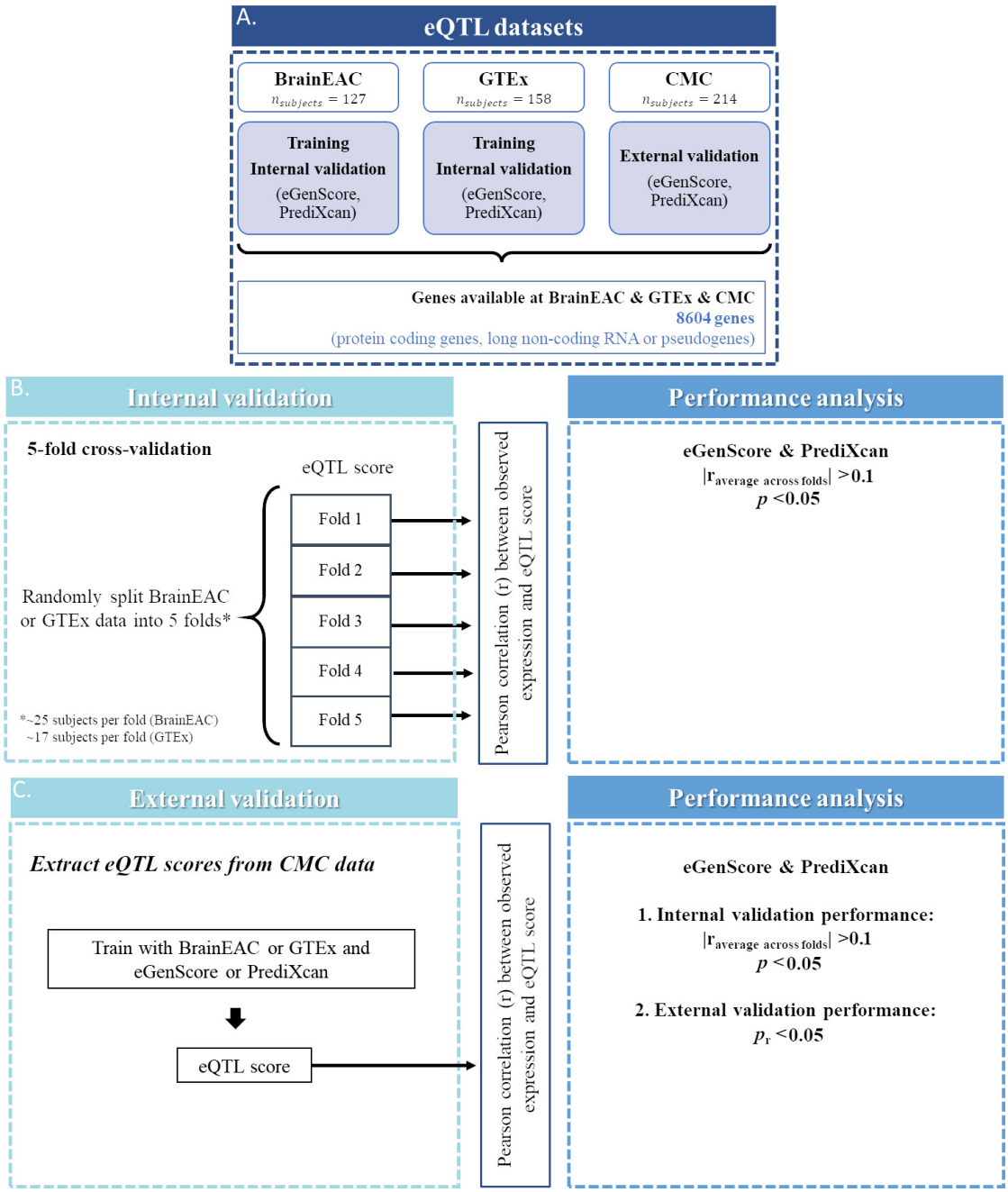


Figure 1. Representation of the steps taken for the selection of genes (**A**) for which an expression quantitative trait loci (eQTL) model was trained and validated, both internally (**B**) and externally (**C**). RNA: ribonucleic acid.

2.1. Genomic and Transcriptomic Datasets

BrainEAC. The BrainEAC dataset was used to train and internally validate eQTL models using eGenScore or PrediXcan. The dataset belongs to the UK Brain Expression

Consortium (UKBEC) [9], was downloaded from the first version of the BrainEAC website (http://www.braineac.org/ (accessed on 19 May 2020)), and is composed of genome-wide genotypes and gene expression levels in the frontal cortex of 127 individuals. All samples and 5,712,227 SNPs have passed quality control (exclusion of individuals with non-European ancestry; samples with call rate < 95%; p-value for deviation from HWE < 10^{-4}; genotyping call rate < 95%; poor post-imputation quality (R^2 < 0.50); and minor allele frequency (MAF) < 5%). From these 127 individuals, gene expression levels for 25,501 genes (normalized using robust multi-array average, $\log_2$ transformed, and corrected for batch effects, sex, and brain bank) were provided by the BrainEAC dataset. Furthermore, the genomic data were mapped onto the human genome assembly GRCh37/hg19, the transcriptomic data were annotated according to NCBI Reference Sequence build 36, and, as for the databases GTEx and CMC and the 1000 Genomes data described below, only SNPs and exon-specific transcripts from chromosomes 1 to 22 were included in this study (i.e., sex chromosomes were excluded).

GTEx. The GTEx dataset (accessed from the GTEx Portal and dbGaP accession number phs000424.v8.p2, on 1 September 2020) was used to train and internally validate eQTL models using eGenScore or PrediXcan. The dataset is part of the Genotype Tissue Expression (GTEx) project conducted by the GTEx Consortium [17], and comprises whole genome sequencing and gene expression levels in the frontal cortex (Broadman area 9) of 158 individuals. All samples and 8,113,423 SNPs have passed quality control (exclusion of individuals with non-European ancestry; samples with call rate < 85%; p-value for deviation from HWE < 10^{-8}; genotyping call rate < 85%; and minor allele frequency (MAF) < 1%). From these 158 individuals, gene expression levels for 17,354 genes in transcripts per million were provided by the GTEx dataset. Furthermore, the genomic data were mapped onto the human genome assembly GRCh38/hg38, and the transcriptomic data were mapped to GENCODE 26.

CMC. The CMC dataset (Release 1) was used to externally validate the eQTL models trained with eGenScore or PrediXcan using BrainEAC or GTEx. The dataset belongs to the CommonMind Consortium [10] and comprises genome-wide genotypes and gene expression levels in the frontal cortex (dorsolateral prefrontal cortex) of 214 individuals. All samples and 39,107,633 SNPs have passed quality control (exclusion of individuals with neuropsychiatric diseases—bipolar disorder, schizophrenia, or affective disorder—and with non-European ancestry; samples with call rate < 90%; p-value for deviation from HWE < 5×10^{-5}; genotyping call rate < 98%). From these 214 individuals, gene expression levels for 15,478 genes in counts per million (normalized by scaling each sample's read count to the total counts by gene, $\log_2$ transformed, and corrected for covariates using surrogate variables analysis) were provided by the CMC dataset. Furthermore, the genomic data were mapped onto the human genome assembly GRCh37/hg19, and the transcriptomic data were annotated to GENCODE 26.

1000 Genomes. The 1000 Genomes datasets (phase 3, October 2015, EUR panel) were used to compute LD and to adjust the weight of each SNP in the eQTL models trained with eGenScore using the BrainEAC dataset (1000 Genomes dataset 1) or the GTEx dataset (1000 Genomes dataset 2) [18]. Both datasets comprise genome-wide genotypes of individuals with European ancestry only (including Finnish). Dataset 1 comprises 78,089,780 SNPs mapped onto the human genome assembly GRCh37/hg19 of 503 individuals. Dataset 2 comprises 73,159,508 SNPs mapped onto the human genome assembly GRCh38/hg38 of 522 individuals.

2.2. Gene Overlap between Datasets

In this study we analyzed only genes that were labeled as protein coding, long non-coding RNA, or pseudogenes in GENCODE (v26, https://www.gencodegenes.org/human/release_26.html (accessed on 1 September 2020) and only if expression levels were available simultaneously in the BrainEAC, GTEx, and CMC datasets. Gene transcript IDs from BrainEAC and GTEx or CMC gene ensemble IDs were aligned using BioMart [19]

by the following criteria: (1) the transcript ID and the gene ensemble ID should be from the same strand (i.e., positive or negative); and (2) when more than one transcript ID in the BrainEAC database correspond to the same gene ensemble ID in the GTEx or CMC database, the transcript ID with the largest overlap (in nucleotide base pairs) with the gene ensemble ID is chosen. Gene expression models were trained and validated internally and externally for 8604 genes with expression levels available in the BrainEAC, GTEx, and CMC databases (Figure 1A).

2.3. eQTL Model Training

An eQTL model for each gene was trained using each combination of the eGenScore and PrediXcan frameworks with the BrainEAC and GTEx datasets. The main differences between the eGenScore and the PrediXcan methods are represented in Figure 2. The first step, common to both frameworks, was to select SNPs located 1 million base pairs upstream and downstream of the gene location in the genome. The following steps are described separately for each framework below.

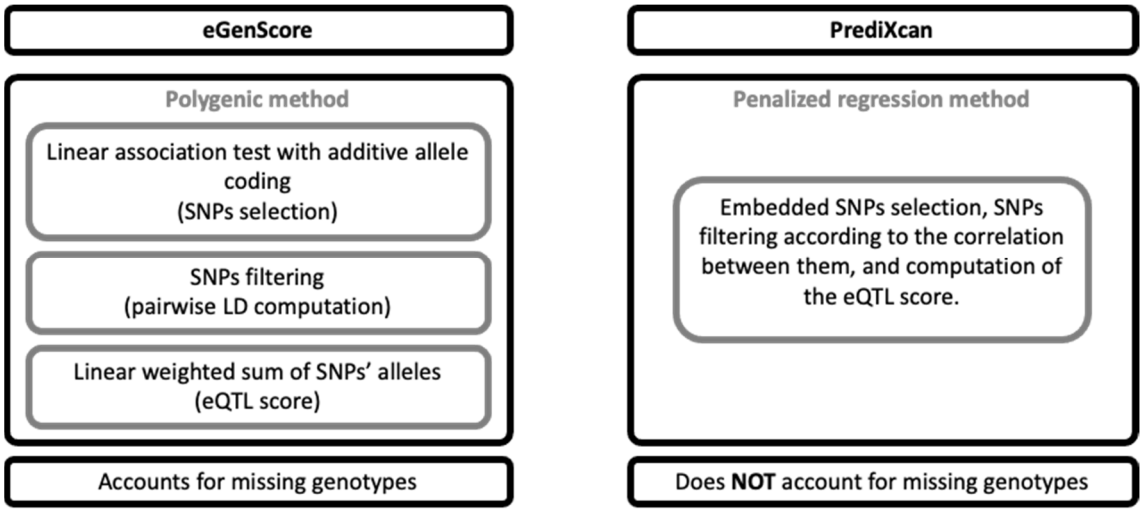


Figure 2. Comparison of the eGenScore's and PrediXcan's methodological features. LD: linkage disequilibrium; SNP: single nucleotide polymorphism.

eGenScore. The association between the SNPs and the gene expression level was tested using linear regression and an additive allele coding (i.e., 0, 1, or 2 tested alleles) for each SNP_i individually (Equation 1) and each $gene_j$ as implemented in Matrix eQTL [20].

$$expression_{gene_j} = \beta_i \times SNP_i \quad (1)$$

SNPs nominally associated ($p < 0.05$) with gene expression were clumped using LD information from the 1000 Genomes dataset. In detail, the SNPs were first ordered by statistical significance (i.e., from the lowest to the highest p-value). Secondly, for every possible unique pair of SNPs, the LD was measured using the 1000 Genomes dataset. Thirdly, for each pair of SNPs in high LD (i.e., $r^2 > 0.3$), the SNP with the lowest significance (i.e., the highest p-value) was excluded. Fourthly, the third step was iterated across all pairs of SNPs in high LD. Each SNP was weighted by the contribution of one tested allele of the SNP to the gene expression level (i.e., the β coefficients from the linear regression described above). The eQTL score, which represents the predicted gene expression, was computed for each $gene_j$ as the weighted sum of each SNP_i's tested alleles adjusted to the expected proportion of those alleles in a standardized population (i.e., the 1000 Genomes dataset) (Equations (2) and (3)). For each SNP_i, this adjustment centers the expected contribution of the SNP_i to the $eQTL\ score_{gene_j}$ at zero. If the genotype of SNP_i is missing in an individual ($CalledSNP_i = 0$), the contribution of SNP_i to $eQTL\ score_{gene_j}$ is also set to zero. In this way, the contribution of a missing SNP to the eQTL score is equal to its expected contribution in

the standardized population [21]. Furthermore, some examples of eQTL score computation are provided in Supplementary Material, Tables S1 and S2.

$$eQTL\ score_{gene_j} = \sum_i ((\beta_i \times SNP_i - adjustment\ factor_i) \times CalledSNP_i) \quad (2)$$

where $CalledSNP_i = \begin{cases} 1, \text{ if the genotype of } SNP_i \text{ is present} \\ 0, \text{ if the genotype of } SNP_i \text{ is missing} \end{cases}$.

$$adjustment\ factor_i = \beta_i \times Proportion_{1\ ref\ allele;\ i} + 2 \times \beta_i \times Proportion_{2\ ref\ allele;\ i} \quad (3)$$

PrediXcan. The PrediXcan framework combines all the SNP selection, SNP LD exclusion, and gene expression prediction steps as described in the eGenScore section into one single step by using regularized linear regression methods [12]. In detail, gene expression is predicted by applying an elastic net regression to the original set of SNPs within the gene region (i.e., 1 million base pairs up- and downstream). Elastic net uses L1 and L2 penalties from least absolute shrinkage and selection operator (LASSO) and ridge regression, respectively, which allows the selection of a set of uncorrelated (i.e., sparse) SNPs [16]. Furthermore, the elastic net regression analysis was conducted using the R package glmnet [22] with $\alpha = 0.5$. Herein, we used the gene expression models that were trained with the GTEx dataset and that are publicly available at PredictDB Data Repository (http://predictdb.org/ (accessed on 1 September 2020)). Additionally, we trained the gene expression models using the BrainEAC dataset, using the same specifications as those used with the GTEx database [12].

2.4. Internal Validation

The eQTL models were internally validated using a 5-fold cross-validation approach. In each iteration, the following measures were computed for each gene using the hold-out fold: (1) the Pearson correlation coefficient (r) between the observed gene expression and the eQTL score; and (2) the p-value corresponding to the null hypothesis of no correlation between the observed gene expression and the eQTL score. Then, as an overall performance measure of the gene expression model, the Pearson correlation coefficient was averaged across the 5 folds (r_{avg}) and squared ($r_{avg}{}^2$). Furthermore, the $r_{avg}{}^2$ is herein interpreted as the variance in the observed gene expression levels that can be explained by the eQTL score (i.e., the predicted gene expression levels). The global p-value was computed using Fisher's method [23], which was previously used by the authors of PrediXcan [12]. Furthermore, the models were considered significant if the averaged correlation between the observed gene expression and the eQTL score was statistically significant (i.e., Fisher's p-value < 0.05) and of at least small size (i.e., $|\, r_{avg} \,| > 0.1$). These performance measures were extracted for all models trained with eGenScore and the BrainEAC or GTEx dataset and PrediXcan and the BrainEAC dataset. They were already available for models trained with PrediXcan and GTEx (at the PredictDB Data Repository).

2.5. External Validation

The eQTL models which were shown to be significant at the internal validation were externally validated using the CMC dataset. The external validation performance was assessed by computing the Pearson correlation coefficient (r) between the observed gene expression and the eQTL score in the CMC dataset and considered statistically significant if the p-value corresponding to the null hypothesis of no correlation between the observed gene expression and the eQTL score was below 0.05. We additionally calculated the squared Pearson correlation coefficient (r^2). These performance measures were extracted for all models trained with eGenScore or PrediXcan and the BrainEAC or GTEx dataset and PrediXcan and the BrainEAC dataset.

2.6. Performance Comparison between eQTL Models

The squared averaged Pearson correlation coefficient (i.e., r_{avg}^2) and the squared Pearson correlation coefficient (i.e., r^2) of the internal and external validations, respectively, were compared across training datasets (BrainEAC vs. GTEx) and across frameworks (i.e., eGenScore vs. PrediXcan) for the genes whose models were significant in the tested combinations (i.e., BrainEAC vs. GTEx or eGenScore vs. PrediXcan) using a two-tailed paired *t*-test. Cohen's *d* was computed as the effect size of the difference, and its interpretation was performed using Kristoffer Magnusson's web tool (Interpreting Cohen's d effect size, https://rpsychologist.com/d3/cohend/ (accessed on 5 April 2021)).

3. Results

3.1. eQTL Models Significant in Each Method

The proportion of eQTL models (i.e., of genes) shown to be statistically significant (i.e., $|r_{avg}| > 0.1$ and Fisher's *p*-value < 0.05) was highest when they were trained with the PrediXcan framework and the GTEx database (23.5%; 2023 out of 8604 genes) and, in descending order, was followed by models trained with (1) eGenScore and GTEx (7.3%; 626 out of 8604 genes); (2) eGenScore and BrainEAC (6.9%; 594 out of 8604 genes); and (3) PrediXcan and BrainEAC (0.8%; 66 out of 8604) (Figure 3).

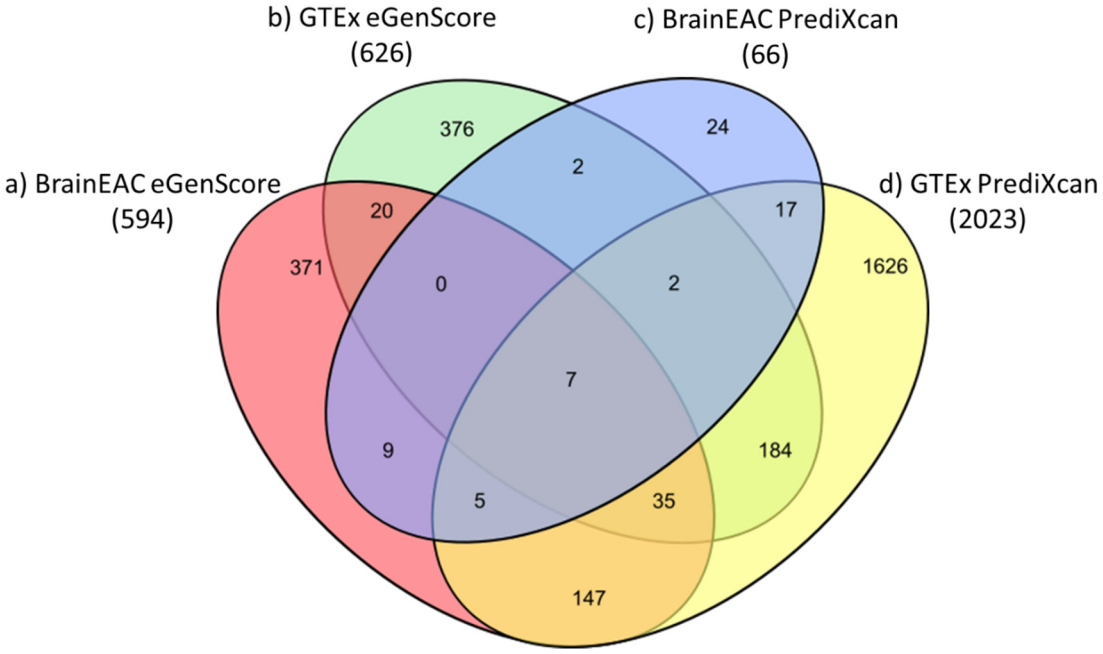


Figure 3. Venn diagram representing the number of genes for which an eQTL model trained with (a) BrainEAC and eGenScore (orange); (b) GTEx and eGenScore (green); (c) BrainEAC and PrediXcan (blue); or (d) GTEx and PrediXcan (yellow) was found to be statistically significant during the internal validation (i.e., $|r_{avg}| > 0.1$ and Fisher's *p*-value < 0.05).

3.2. eQTL Model's Internal Validation

When comparing the internal validation performance of the gene expression models across databases, the squared averaged Pearson correlation coefficient between the observed gene expression and the eQTL score was statistically different between BrainEAC and GTEx using both frameworks (eGenScore, $p = 0.003$, or PrediXcan, $p = 0.001$) (Table 1 and Figure 4). Furthermore, the performance was shown to be higher for BrainEAC when using eGenScore and for GTEx when using PrediXcan. When comparing the eQTL models' performance across frameworks, models trained with PrediXcan showed better performance than the ones trained with eGenScore ($p < 0.001$) but only when using the GTEx database (Table 1 and Figure 4).

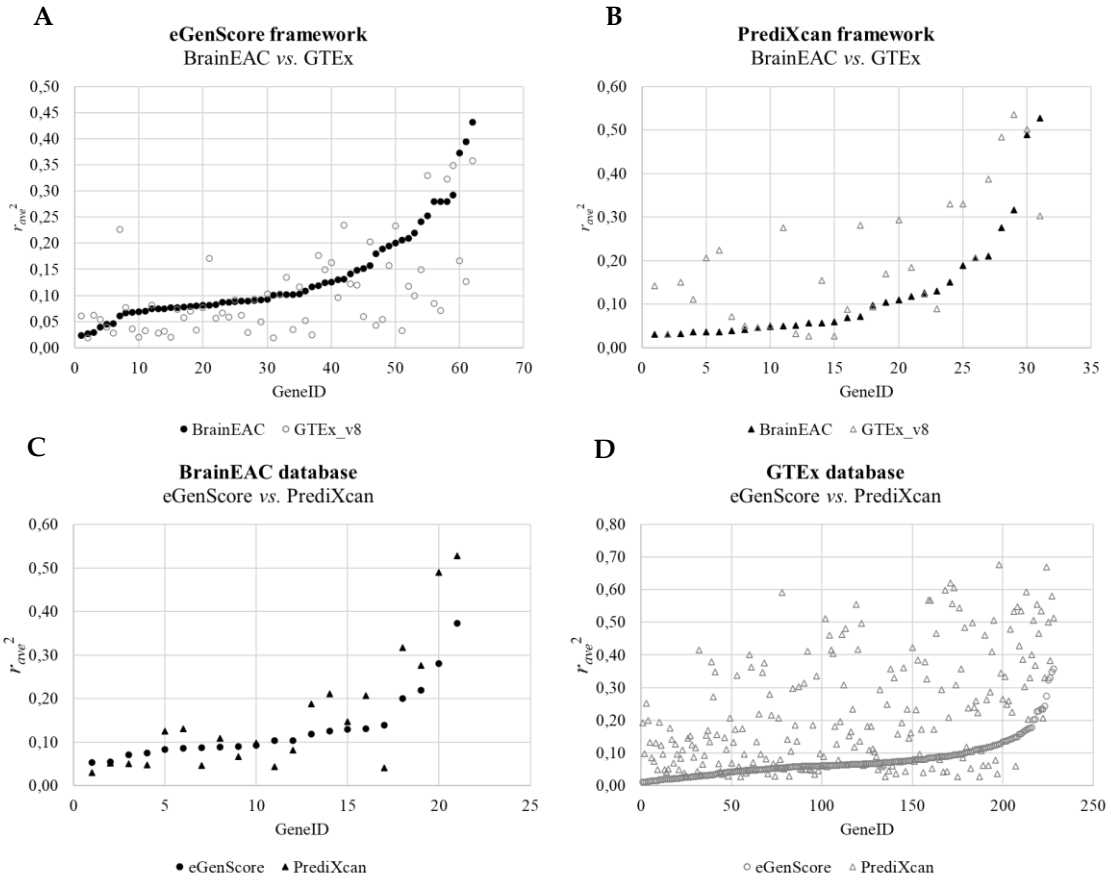


Figure 4. Comparison of the squared averaged Pearson correlation coefficient (r_{avg}^2) between the predicted and observed gene expressions during the internal cross-validation across databases (i.e., BrainEAC vs. GTEx) using the eGenScore (**A**) or the PrediXcan (**B**) framework and across frameworks (i.e., eGenScore vs. PrediXcan) using the BrainEAC (**C**) or GTEx (**D**) dataset. The model's performance is represented with a filled black or hollow gray marker if trained with the BrainEAC or GTEx database, respectively, and with a circle or triangle if trained with the eGenScore or PrediXcan framework, respectively.

Table 1. Comparison of the gene expression models' internal validation performance (i.e., the squared averaged Pearson correlation coefficient between the predicted and observed gene expressions) across datasets (i.e., BrainEAC vs. GTEx) and across frameworks (i.e., eGenScore vs. PrediXcan).

Comparison	df, *t*	*p*	Cohen's *d*
eGenScore framework (BrainEAC vs. GTEx)	61, 3.10	0.003 **	0.39
PrediXcan framework (BrainEAC vs. GTEx)	30, −3.63	0.001 ***	0.65
BrainEAC dataset (eGenScore vs. PrediXcan)	20, −1.79	0.088	0.39
GTEx dataset (eGenScore vs. PrediXcan)	227, −13.86	<0.001 ***	0.92

Only genes with a significant model (i.e., with an absolute averaged Pearson correlation coefficient between the predicted and observed gene expressions above 0.1 and a Fisher's *p*-value below 0.05) were considered for this comparison. A two-sided paired-sample *t*-test was conducted and considered statistically significant at a *p*-value < 0.05. df: degrees of freedom (i.e., number of genes for which there was a significant model minus one). **: $p < 0.01$; ***: $p < 0.001$; *t*: t-statistic.

3.3. eQTL Model's External Validation

Across databases, the external validation performance of the eQTL models (i.e., when applied to the CMC database) was shown to be higher when trained with the BrainEAC than with the GTEx database but only when using the eGenScore framework ($p = 0.015$) (Table 2). No statistically significant difference in the external validation performance of the eQTL models was found when models were trained with the BrainEAC or GTEx database

and PrediXcan framework (Table 2). Across frameworks, the performance of the gene expression models was shown to be higher for PrediXcan compared to eGenScore when using BrainEAC ($p = 0.018$) or GTEx ($p < 0.001$) datasets (Table 2).

Table 2. Comparison of the gene expression models' external validation performance (i.e., the squared Pearson correlation coefficient between the predicted and observed gene expressions in the CMC dataset) across datasets (i.e., BrainEAC vs. GTEx) and across frameworks (i.e., eGenScore vs. PrediXcan).

Comparison	df, *t*	*p*	Cohen's *d*
eGenScore framework (BrainEAC vs. GTEx)	33, 2.57	0.015 *	0.44
PrediXcan framework (BrainEAC vs. GTEx)	15, −2.04	0.060	0.51
BrainEAC dataset (eGenScore vs. PrediXcan)	8, −2.95	0.018 *	0.98
GTEx dataset (eGenScore *vs.* PrediXcan)	115, −15.76	<0.001 ***	1.46

Only genes with a statistically significant Pearson correlation coefficient between the predicted and observed gene expressions in the CMC dataset (p-value < 0.05) were considered for this comparison. A two-sided paired-sample t-test was conducted and considered statistically significant at a p-value < 0.05. df: degrees of freedom (i.e., number of genes for which there was a model whose predicted expression correlated significantly with the observed expression of that gene minus one). *: $p < 0.05$; ***: $p < 0.001$; t: t-statistic.

4. Discussion

We herein presented eGenScore, a polygenic score-based method to predict gene expression levels from genotypes. In a previous paper, the elastic net-based framework PrediXcan was shown to outperform a polygenic score-based method [12]. However, that polygenic score-based method was methodologically limited; in comparison to it, eGenScore better handles the following two issues: (1) the LD between SNPs shown to be individually associated with gene expression; and (2) missing genotypes. We compared the performance of frontal cortex-specific gene expression models trained with different frameworks, eGenScore vs. PrediXcan, as well as with different datasets, BrainEAC vs. GTEx, after both internal and external validation steps.

Overall, our results confirm that elastic net-based methods are superior to polygenic score-based methods for the prediction of gene expression based on eQTL genotypes. PrediXcan predicted gene expression levels with a higher performance than eGenScore regardless of the database (i.e., BrainEAC or GTEx) used for model training. Indeed, the observed difference in the internal validation performance between frameworks when using the GTEx database corresponded to a large effect size, with roughly 82% of the gene expression models showing higher performance when trained with PrediXcan than with eGenScore (i.e., 186 out of 228 genes; Cohen's $d = 0.92$). This effect was enlarged when the gene expression models were applied to an external database (i.e., CMC), with 93% of the gene expression models showing a higher correlation between the observed and predicted gene expressions in the CMC database when trained with PrediXcan than with eGenScore (i.e., 108 out of 116 genes; Cohen's $d = 1.46$). However, the frameworks differed in their best training dataset; models trained with BrainEAC outperformed those trained with GTEx when the eGenScore framework was used, whereas the opposite was observed when the PrediXcan framework was used. The effect of the training database on performance was shown to be higher for the PrediXcan framework, with 74% of the gene expression models showing higher internal performance when trained with GTEx (i.e., 22 out of 31 genes; Cohen's $d = 0.65$). This higher dependence on the training dataset may compromise an assumption of generalizability of PrediXcan across training sets.

Given that PrediXcan was shown to be a better framework for predicting gene expression levels than eGenScore, our results suggest that GTEx should be used as the training database for these gene expression models. When compared with BrainEAC, GTEx is a more comprehensive transcriptomic and genomic database with a slightly larger sample size for brain gene expression and uses whole-genome sequence data and gene expression data (compared to gene expression array data employed in BrainEAC). This different source

of gene expression data may explain the lack of concordance between eQTL models derived using the same framework in different datasets and the poorer performance of PrediXcan in BrainEAC compared to GTEx. The challenges of obtaining gene expression data in brain tissues may necessitate a variety of approaches to measuring brain gene expression. As such, a valuable direction for future research may be to improve the generalizability of elastic net-based frameworks such as PrediXcan to work effectively across different sources of gene expression data. Indeed, this may be a particularly important step towards a universal, non-invasive, statistical estimator of tissue-specific gene expression, a very important tool for an effective translation of TWAS into clinical practice [7].

5. Limitations

Our study had several limitations that need to be addressed. Firstly, the gene expression data of the BrainEAC and GTEx databases were annotated to different human genome assemblies, which hindered the exact correspondence between the transcriptomic data of the two databases and, therefore, narrowed the number of possible genes that we could analyze. Secondly, although we restricted the comparisons of the performance of gene expression models to genes expressed in the frontal cortex, in fact, the brain samples from which the transcriptomic data were extracted across databases were not exactly from the same brain location. It is reasonable to expect that the expression level of a given gene might slightly vary depending on which exact location in the brain it is taken from. Therefore, the comparison of model performance across datasets might be influenced by this factor. Thirdly, both eGenScore and PrediXcan methods rely on genotype–gene expression association data (such as are provided in the BrainEAC and GTEx databases) and, therefore, are only valid for the age interval of the sample used in these databases. They cannot account for the epigenetic effects on gene expression across the lifespan.

6. Conclusions

In this study, we compared the performance of eQTL models trained with: (1) different frameworks—eGenScore, a novel (introduced for the first time herein) and improved polygenic method that, compared with the original polygenic method presented along with PrediXcan, addresses high LD between SNPs and handles individual missing genotypes; and PrediXcan, a previously published and regularized linear regression method (i.e., elastic net)—; and (2) different training datasets—BrainEAC and GTEx. Taken together, our results show that: (1) PrediXcan outperforms eGenScore regardless of the training database that is used (i.e., BrainEAC or GTEx); and (2) GTEx yields eQTL models with a higher performance than BrainEAC when using PrediXcan. Therefore, we encourage the use of models trained with the GTEx database and using the PrediXcan framework when predicting gene expression from genotype data.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes12101531/s1, Table S1: β values, adjustment factor, and coding alleles for the list of SNPs to be included in the eQTL score model for the neuregulin 4 gene (*NRG4*); Table S2: Examples of how the eQTL score is computed. In detail, eQTL score was computed for four hypothetical individuals.

Author Contributions: V.T. was involved in the study design, analysis, results interpretation, and writing of the manuscript. J.M. contributed to data analysis. E.V., J.C. and D.P. were involved in the study design and critically revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: For this work, V.T. received support from a Fundação para a Ciência e a Tecnologia (FCT) PhD fellowship (PD/BD/114460/2016) and DSAIPA/DS/0065/2018 grant. D.P. received support from FCT grants FCT-IF/00787/2014, LISBOA-01–0145-FEDER-030907, and DSAIPA/DS/0065/2018; a European Commission Marie Curie Career Integration Grant (FP7-PEOPLE-2013-CIG 631952); a Breakthrough Idea Grant from the Director's Fund of iMM Lisboa (2016); and a Bial Foundation Psychophysiology Grant (Ref. 292/16).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available upon request to the authors.

Acknowledgments: We thank the United Kingdom Brain Expression Consortium (UKBEC) investigators for BrainEAC access, whose work was supported by the MRC through the MRC Sudden Death Brain Bank (G0901254), a Training Fellowship (G0802462), and by the King Faisal Specialist Hospital and Research Centre, Saudi Arabia. We also thank the investigators of the Genotype-Tissue Expression (GTEx) Project, which was supported by the Common Fund of the Office of the Director of the National Institutes of Health (NIH) and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Finally, we thank the CommonMind Consortium, supported by funding from Takeda Pharmaceuticals Company Limited, F. Hoffmann-La Roche Ltd, and NIH grants R01MH085542, R01MH093725, P50MH066392, P50MH080405, R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881, AG02219, AG05138, MH06692, R01MH110921, R01MH109677, R01MH109897, U01MH103392, and contract HHSN271201300031C through IRP NIMH. Brain tissue for the study was obtained from the following brain bank collections: the Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania Alzheimer's Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain and Tissue Repositories, and the NIMH Human Brain Collection Core. CMC Leadership: Panos Roussos, Joseph Buxbaum, Andrew Chess, Schahram Akbarian, Vahram Haroutunian (Icahn School of Medicine at Mount Sinai), Bernie Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University of Pennsylvania), Enrico Domenici (University of Trento), Mette A. Peters, Solveig Sieberts (Sage Bionetworks), Thomas Lehner, Stefano Marenco, and Barbara K. Lipska (NIMH). This paper represents independent research partly funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre in South London, the Maudsley NHS Foundation Trust, and King's College, London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

Conflicts of Interest: D.P. is a co-founder and shareholder of the neuroimaging research services company NeuroPsyAI, Ltd. J.C. is an editorial board member for Genes. All authors declare that they have no conflict of interest.

References

1. Trifu, S.; Kohn, B.; Vlasie, A.; Patrichi, B.-E. Genetics of schizophrenia (Review). *Exp. Ther. Med.* **2020**, 59–70. [CrossRef]
2. Lappalainen, T.; Sammeth, M.; Friedländer, M.R.; 'T Hoen, P.A.C.; Monlong, J.; Rivas, M.A.; Gonzàlez-Porta, M.; Kurbatova, N.; Griebel, T.; Ferreira, P.G.; et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **2013**, *501*, 506–511. [CrossRef]
3. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J.P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367. [CrossRef]
4. Watanabe, K.; Stringer, S.; Frei, O.; Mirkov, M.U.; de Leeuw, C.; Polderman, T.J.C.; van der Sluis, S.; Andreassen, O.A.; Neale, B.M.; Posthuma, D. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **2019**, *51*, 1339–1348. [CrossRef]
5. Aguet, F.; Brown, A.A.; Castel, S.E.; Davis, J.R.; He, Y.; Jo, B.; Mohammadi, P.; Park, Y.S.; Parsana, P.; Segrè, A.V.; et al. Genetic effects on gene expression across human tissues. *Nature* **2017**, *550*, 204–213. [CrossRef]
6. Cummings, B.B.; Marshall, J.L.; Tukiainen, T.; Lek, M.; Donkervoort, S.; Foley, A.R.; Bolduc, V.; Waddell, L.B.; Sandaradura, S.A.; O'Grady, G.L.; et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **2017**, *9*, eaal5209. [CrossRef]
7. Marco-Puche, G.; Lois, S.; Benítez, J.; Trivino, J.C. RNA-Seq Perspectives to Improve Clinical Diagnosis. *Front. Genet.* **2019**, *10*, 1152. [CrossRef]
8. Lonsdale, J.; Thomas, J.; Salvatore, M.; Phillips, R.; Lo, E.; Shad, S.; Hasz, R.; Walters, G.; Garcia, F.; Young, N.; et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **2013**, *45*, 580–585. [CrossRef] [PubMed]
9. Ramasamy, A.; Trabzuni, D.; Guelfi, S.; Varghese, V.; Smith, C.; Walker, R.; De, T.; Coin, L.; De Silva, R.; Cookson, M.R. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.* **2014**, *17*, 1418–1428. [CrossRef] [PubMed]
10. Hoffman, G.E.; Bendl, J.; Voloudakis, G.; Montgomery, K.S.; Sloofman, L.; Wang, Y.C.; Shah, H.R.; Hauberg, M.E.; Johnson, J.S.; Girdhar, K.; et al. CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. *Sci. Data* **2019**, *6*, 180. [CrossRef] [PubMed]
11. Barbeira, A.N.; Dickinson, S.P.; Bonazzola, R.; Zheng, J.; Wheeler, H.E.; Torres, J.M.; Torstenson, E.S.; Shah, K.P.; Garcia, T.; Edwards, T.L.; et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **2018**, *9*, 1–20. [CrossRef] [PubMed]

12. Gamazon, E.R.; Wheeler, H.E.; Shah, K.P.; Mozaffari, S.V.; Aquino-Michaels, K.; Carroll, R.J.; Eyler, A.E.; Denny, J.C.; Nicolae, D.L.; Cox, N.J.; et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **2015**, *47*, 1091–1098. [CrossRef] [PubMed]
13. Huckins, L.M.; Dobbyn, A.; Ruderfer, D.M.; Hoffman, G.; Wang, W.; Pardiñas, A.F.; Rajagopal, V.M.; Als, T.D.; TNguyen, H.; Girdhar, K.; et al. Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nat. Genet.* **2019**, *51*, 659–674. [CrossRef] [PubMed]
14. Wang, J.; Gamazon, E.R.; Pierce, B.L.; Stranger, B.E.; Im, H.K.; Gibbons, R.D.; Cox, N.J.; Nicolae, D.L.; Chen, L.S. Imputing Gene Expression in Uncollected Tissues Within and beyond GTEx. *Am. J. Hum. Genet.* **2016**, *98*, 697–708. [CrossRef] [PubMed]
15. Wray, N.R.; Lee, S.H.; Mehta, D.; Vinkhuyzen AA, E.; Dudbridge, F.; Middeldorp, C.M. Research Review: Polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* **2014**, *55*, 1068–1087. [CrossRef]
16. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]
17. Aguet, F.; Barbeira, A.N.; Bonazzola, R.; Jo, B.; Kasela, S.; Liang, Y.; Parsana, P.; Aguet, F.; Battle, A.; Brown, A.; et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **2020**, *369*, 1318–1330. [CrossRef]
18. Auton, A.; Abecasis, G.R.; Altshuler, D.M.; Durbin, R.M.; Bentley, D.R.; Chakravarti, A.; Clark, A.G.; Donnelly, P.; Eichler, E.E.; Flicek, P.; et al. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. [CrossRef]
19. Durinck, S.; Moreau, Y.; Kasprzyk, A.; Davis, S.; De Moor, B.; Brazma, A.; Huber, W. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics* **2005**, *21*, 3439–3440. [CrossRef]
20. Shabalin, A.A. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **2012**, *28*, 1353–1358. [CrossRef]
21. Vassos, E.; Sham, P.; Kempton, M.; Trotta, A.; Stilo, S.A.; Gayer-Anderson, C.; Di Forti, M.; Lewis, C.M.; Murray, R.M.; Morgan, C. The Maudsley environmental risk score for psychosis. *Psychol. Med.* **2020**, *50*, 2213–2220. [CrossRef]
22. Friedman, J.H.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef] [PubMed]
23. Fisher, R.A. Statistical Methods for Research Workers. In *Breakthroughs in Statistics*; Springer Series in Statistics (Perspectives in, Statistics); Kotz, S., Johnson, N.L., Eds.; Springer: New York, NY, USA, 1992; pp. 66–70. [CrossRef]