Cardiff University

Doctoral Thesis

Machine Learning for Genetic Prediction of Schizophrenia

Author:
Matthew Smith

Supervisor:
Professor Valentina Escott-Price

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

MRC Centre for Neuropsychiatric Genetics and Genomics
School of Medicine

August, 2021

CARDIFF UNIVERSITY

Abstract

The complexity of schizophrenia raises a formidable challenge. Its diverse genetic architecture, influence from environmental factors from the prenatal period through to adolescence, and the absence of a laboratory-based diagnostic test complicate efforts to "carve nature at its joints". Twinned with attempts to disentangle schizophrenia's origins are those aiming to predict it.

Prediction is essential to precision psychiatry and attempts to improve patient outcomes. Genetic prediction only became feasible relatively recently, following the discovery of robust risk loci in association studies. Polygenic risk scoring (PRS) is a popular method which relies on univariable tests of association and typically assumes additivity within and between loci, but explains only a small fraction of liability to schizophrenia. Machine learning (ML) methods have evolved out of the artificial intelligence and statistics communities which learn predictive patterns from labelled data. They are an enticing option in genetics, as they allow for multivariable predictive modelling, complex predictor relationships including interactions and can learn from datasets where the number of predictors exceeds observations. However, their predictive performance in schizophrenia is largely unknown.

The ability of penalised logistic regression, support vector machines, random forests (RFs), gradient boosting machines (GBMs) and neural networks to predict schizophrenia from genetic data was investigated. A review systematically assessed predictive performance and methodology in machine learning on psychiatric disorders, finding poor reporting, widespread inadequate modelling approaches and high risk of bias. Simulations assessed performance in the presence of additive or interaction effects. Flexible ML approaches including RFs and GBMs performed best under interactions, but worse than PRS and sparse linear models for additive effects. Evaluation in real data assessed modelling procedures including calibration and deconfounding. Prediction was maximised when combining genetic and non-genetic factors; no evidence was found to support choosing machine learning approaches over logistic regression or PRS.

Acknowledgements

I would like to acknowledge my supervisor Valentina for all her help, guidance and patience over the last 3 and a half years. I could not have asked for a more supportive supervisor. I would also like to thank my other supervisors, George Kirov, Elliott Rees and Andrey Pepelyshev for their help, particularly for giving up so much time to read drafts of my thesis.

I have been fortunate to have been funded for both a PhD and 3 rotations, which has been a hugely enriching experience. I would like to thank the MRC for funding and Cardiff University for their support.

I would also like to thank my family. Particularly my wife, Han, for her love and support; you made two children and a PhD seem achievable (and even enjoyable at times!). Lastly, I have to thank my daughter, Alba, and my son, Emri, for always reminding me what is really important.

Contents

List of Figures

List of Tables

List of Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
ASD	Autism Spectrum Disorder
AUC	Area Under the receiver operating characteristic Curve
CARTs	Classification And Regression Trees
CNVs	Copy Number Variants
DEV	Development
DSM	Diagnostic and Statistical Manual
DT	Decision Tree
EPV	Events Per Variable
ePRS	external Polygenic Risk Score
FDR	False Discovery Rate
GBM	Gradient Boosting Machine
GWAS	Genome-Wide Association Studies
GWS	Genome-Wide Significant
HP	Hyperparameter
ICD	International statistical Classification of Diseases
iPRS	internal Polygenic Risk Score
***k*-NN**	*k*-Nearest Neighbours
LASSO	Least Absolute Shrinkage and Selection Operator
LD	Linkage Disequilibrium
LR	Logistic Regression
MAF	Minor Allele Frequency
ML	Machine Learning
MLP	Multi-Layer Perception
NN	Neural Network
OR	Odds Ratio
PCA	Principal Component Analysis
PGC	Psychiatric Genomics Consortium
PROBAST	Prediction model Risk Of Bias ASsessment Tool
PRS	Polygenic Risk Score
RBF SVM	Radial Basis Function Support Vector Machine
RF	Random Forest
ROB	Risk of Bias
ROC	Receiver Operator Characteristic

SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SVM	Support Vector Machine
VAL	Validation
XGBoost	eXtreme Gradient Boosting
XOR	Exclusive-Or

Chapter 1

Introduction

1.1 Introduction

Schizophrenia is "the archetypal form of madness" (Frith and Johnstone, 2003) and referred to as the "heartland" of psychiatry (Goodwin and Geddes, 2007). Despite its prominence in the field, the research focus this has engendered and the commonly-held observation that so-called "madness" runs in families, it was just over a century since the term schizophrenia was first coined that common genetic variation was reliably associated with the disorder (Purcell et al., 2009). Family, twin and adoption studies laid the groundwork for later linkage, candidate gene and genome-wide association studies which firmly established the polygenic nature of schizophrenia. This has been supported by studies of rare variation, confirming the role of common polymorphisms, rare variants, insertions, deletions or large structural changes in an intricate genetic landscape (Legge et al., 2021). These relatively recent successes join well-established epidemiological evidence of environmental risk factors (Stilo and Murray, 2019). In an era of precision medicine, it is perhaps unsurprising that prediction has been suggested over association as a means of achieving precision psychiatry (Bzdok, Varoquaux, and Steyerberg, 2020), and that supervised machine learning has been highlighted as a group of flexible statistical methods which are adept at pattern matching and may potentially hasten its arrival (Manchia et al., 2020).

1.2 Schizophrenia

1.2.1 Overview

Schizophrenia is a complex and highly heritable psychiatric disorder that can be caused by genetics, environmental factors or a combination. It makes an out-sized contribution to the global burden of disease at around 21 million individuals in 2016, and 13.4 million years of life lived with disability (Charlson et al., 2018). The overall societal cost of schizophrenia was estimated at £6.7 billion in 2004-2005 alone (Mangalore and Knapp, 2007). Life expectancy reduces by 10-20 years in individuals with schizophrenia (Owen, Sawa, and Mortensen, 2016), with elevated risk of all-cause mortality assessed by standardised mortality ratio (SMR) at 2.5 (2.2-2.4 95% CI), reaching 2.4 and 3 in females and males respectively (Chesney,

Goodwin, and Fazel, 2014). Individuals with schizophrenia are particularly at increased risk for suicide, with an SMR of 12.9 (0.7-174 95% CI) (Chesney, Goodwin, and Fazel, 2014); accident, cardiovascular disease, cancer, infection, substance use and chronic obstructive pulmonary disease also contribute to elevated mortality (Olfson et al., 2015). Individuals with schizophrenia account for 10% of the homeless population (Ayano, Tesfaw, and Shumet, 2019) and have rates of 80-90% unemployment (Marwaha and Johnson, 2004). Though some cases remit and do not re-occur, around 87% of individuals have been estimated to have a chronic illness course (Jääskeläinen et al., 2013). Overall, the impact on the individual and society is severe.

1.2.2 Symptoms

Symptoms are of paramount importance in schizophrenia as they provide both a description of the disorder and the basis for its diagnosis. Schizophrenia is characterised by positive, negative and cognitive symptoms (Owen, Sawa, and Mortensen, 2016). Positive symptoms refers to the addition of characteristics which are not normally present, namely psychosis (delusions and hallucinations), while negative symptoms include aspects normally present in the population but which are absent in the individual. These include apathy, anhedonia, alogia, asociality, anergia, avolition and flattened affect. Cognitive symptoms, also known as disorganised symptoms or thought disorder, describe disorganisation of speech and behaviour, as well as reduced performance in cognitive tests (Burton, 2016).

1.2.2.1 Positive symptoms

A delusion is a fixed belief which is maintained despite evidence against it, and which is not explained by religion or culture. Hallucinations describe perceptions without corresponding stimuli which have not been consciously manipulated (Burton, 2016). Though symptoms of psychosis may be found outside schizophrenia, such as in response to amphetamines, infection or central nervous system (CNS) pathology, psychosis in schizophrenia typically includes fully-formed auditory hallucinations, with delusions and hallucinations often framed around a loss of control. Of particular relevance in schizophrenia are first rank symptoms (FRS), proposed by Kurt Schneider in 1959, with the aim of identifying aspects of psychosis which are pathognomic of schizophrenia (Schneider, 1959). Despite widespread uptake and a large impact on diagnostic criteria, these have been shown not to be uniquely descriptive of schizophrenia (Peralta and Cuesta, 1999); diminished aspects of psychosis such as paranoia and auditory hallucinations occur in over 5% of healthy individuals (Van Os et al., 2009). FRS still provide an informative summary of the more florid psychotic symptoms found in schizophrenia.

First rank symptoms are divided into four main areas: auditory hallucinations, delusions of thought control, other delusions of control, and delusional perception. Auditory hallucinations may be in the form of the third person, where multiple voices converse about the patient, a running commentary on the individual's movements and thoughts, or thought

echo, also known as gedankenlautwerden or écho de la pensée. Delusions of thought control centre around the belief that the individual's thoughts are being manipulated through insertion into their head by another, withdrawal from their head by a third party, or that their thoughts are being broadcast, overheard or accessed by others. Other delusions of control describe passivity phenomena, whereby an individual believes their volition, affect or impulses are being controlled by another, or delusions of somatic passivity where the body of the person is under control. Finally, delusional perception is the delusional interpretation of normal phenomena. Positive symptoms are often prominent in attempts to classify schizophrenia (Burton, 2016).

1.2.2.2 Negative symptoms

Negative symptoms were central to Bleuler's definition of "the schizophrenias"; he held that aspects such as ambivalence and asociality take primacy over hallucinations and delusions (Frith and Johnstone, 2003). Negative symptoms are more difficult to identify than psychosis, as they are often misinterpreted as part of adolescence, other psychiatric disorders or even side affects from antipsychotics (Burton, 2016). The extent to which positive or negative symptoms predominate differs between individuals, with negative symptoms apparently absent in some (Owen, Sawa, and Mortensen, 2016). Though the course of schizophrenia varies greatly, a typical experience is for more florid positive symptoms to occur in acute phases early on in the course of the disorder, undergoing episodes of relapse and remission, followed by greater prominence for less obvious negative symptoms, which may persist through a prolonged "chronic" phase, including remission, or for the rest of the individual's life.

1.2.2.3 Cognitive symptoms

In contrast to Bleuler, Kraeplin's earlier definition of "dementia praecox" highlighted the cognitive decline he observed in patients (Frith and Johnstone, 2003), and which has been suggested as fundamental to schizophrenia (Elvevag and Goldberg, 2000; Kahn and Keefe, 2013). For simplicity here, reductions in intelligence quotient (IQ) and measures assessed in cognitive batteries are combined under the umbrella of cognitive deficits. Epidemiological evidence for cognitive impairment prior to onset in schizophrenia is available from the Israeli draft board register, military records which provide a psychiatric assessment undertaken by all individuals upon being drafted at 17 years old. Those with below-expected IQ were at higher risk for later hospitalisation with schizophrenia (Reichenberg et al., 2005). Meta-analyses of premorbid IQ have supported this (Khandaker et al., 2011; Dickson et al., 2012). Poor premorbid school performance has also been associated with risk for schizophrenia (MacCabe et al., 2008), while particularly good school performance has been associated with risk for bipolar disorder (MacCabe et al., 2010) in studies of all individuals in a Swedish national cohort.

In addition to the period preceding illness, cognitive impairment is also associated with course of the illness itself. Analysis of individuals from the Israeli draft board data demonstrate worsening IQ scores as assessment approached the timing of the first episode of psychosis, with the large difference in scores between affected and unaffected individuals observed in those who had already experienced their first episode (Rabinowitz et al., 2000). Individuals with schizophrenia are typically 1-1.5 standard deviations below the mean on standard cognitive measures (Keefe and Fenton, 2007). The association of cognitive deficits with genetic risk for schizophrenia has been demonstrated in carriers of copy number variants (CNVs) in the UK Biobank, where cognitive measures are again shifted below the sample mean by up to 1 standard deviation for participants with schizophrenia and up to 0.5 standard deviations for unaffected carriers (Kendall et al., 2017). Analysis of common variation has also shown negative genetic correlation of -0.2 between schizophrenia and intelligence (Anttila et al., 2018), and cognitive deficits are present in unaffected relatives of probands (Sitskoorn et al., 2004), indicating cognitive deficits may be partly explained by genetic liability to schizophrenia.

Cognitive deficits in schizophrenia are broad (Heinrichs and Zakzanis, 1998), but are particularly pronounced for working memory, attention and executive function (Elvevag and Goldberg, 2000). Affected individuals frequently perform poorly on tests of verbal fluency, such as listing items beginning with a particular letter, which is interpreted as impairment in execution function through inability to plan a search strategy (Elvevag and Goldberg, 2000). Individuals may also show deficits in speech, such as difficulty in synthesising or expressing thoughts to others. Poor performance in cognitive tests is associated with understanding rather than motivation (Green et al., 1992; Hellman et al., 1998). It is present independent of institutionalisation (Johnstone et al., 1981) and medication (Mohamed et al., 1999), in addition to having been described before chlorpromazine, the first antipsychotic, was introduced in the 1950s.

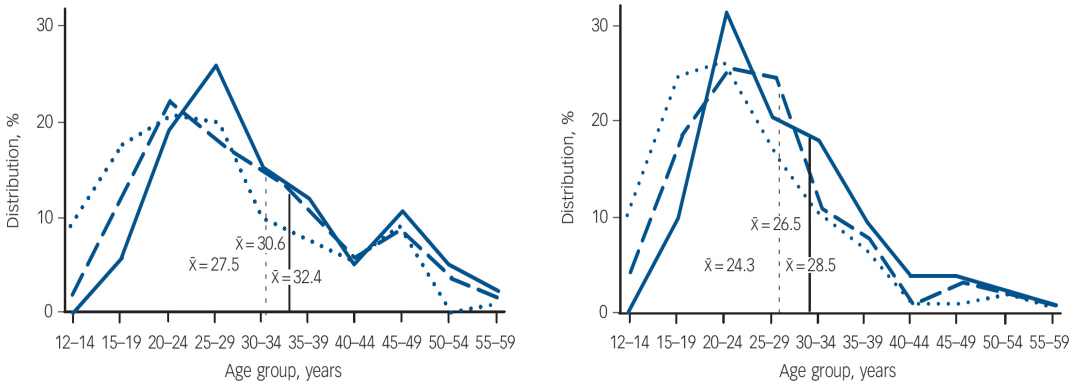
The emphasis on cognitive deficits in schizophrenia marks a separation from bipolar disorder; though genetic risk for both may explain a small fraction of the variance in creativity (Power et al., 2015), bipolar disorder itself has been associated with creativity and genius (Andreasen, 1987; Ludwig, 1992; Jamison, 1996), perhaps due to periods of euthymia and partial normality present in bipolar disorder but absent in schizophrenia, and does not appear to be associated with premorbid cognitive decline (Reichenberg et al., 2002) or poor premorbid school performance (MacCabe et al., 2010).

1.2.3 Epidemiology

Incidence and prevalence measure what proportion of a population have an outcome. Incidence measures new cases in a set timeframe, while prevalence includes all cases that are counted for a single timepoint (point prevalence) or a period of time (period prevalence) (Rothman, 2012). Median incidence of schizophrenia is estimated at 15.2 (7.7-43.0, 10th-90th percentile) per 100,000 (McGrath et al., 2008). Lifetime prevalence of schizophrenia

is often labelled as around 1%. However, estimates of prevalence differ by how the disorder is defined and whether point prevalence, period prevalence (typically 12 months), lifetime prevalence (proportion of individuals experiencing an outcome up to time of assessment) or lifetime morbid risk (proportion that may experience outcome at some point) are reported. In addition, prevalence can be affected by changing diagnostic criteria over time, ascertainment bias due to study design, cultural and healthcare differences between countries and regions and sample size, and is likely to be higher in specific populations, such as those in the homeless population or in hospitals (Simeone et al., 2015).

Estimates of the prevalence of schizophrenia have been the subject of several systematic reviews (Saha et al., 2005; McGrath et al., 2008; Simeone et al., 2015). The most frequently cited are those of McGrath et al., 2008, which estimates the median (10th percentile, 90th percentile) for point prevalence, period prevalence, lifetime prevalence and lifetime morbid risk for schizophrenia to be 0.46% (0.19% - 1%), 0.33% (0.13% - 0.82%), 0.40% (0.16%, 1.21%) and 0.72% (0.31% - 2.71%). All systematic reviews note large heterogeneity between estimates in studies. The global burden of disease study estimates point prevalence as lower, at 0.28% (0.24 – 0.31 95% CI), though the authors note the difficulty in comparison to Saha et al., 2005, due to differences in inclusion criteria and methodology (Charlson et al., 2018).



(A) Females (B) Males

FIGURE 1.1: Age of onset in schizophrenia for females (A) and males (B). Ages are shown for first signs of disturbance (dotted line), first symptom of psychosis (dashed line) and point of admission (solid line). Mean age of onset is given for each line. Adapted from Jones, 2013; data from Hafner et al., 1994.

Prevalence is generally similar between geographical populations (Charlson et al., 2018). Though estimates of 12-month period prevalence and lifetime prevalence show some differences across countries (Simeone et al., 2015), particularly with latitude (Saha et al., 2006), many of the factors listed above may contribute to these, with estimates for psychosis shown to be affected by prevalence type, population studied, disorder definition and study quality (Moreno-Küstner, Martin, and Pastor, 2018). Within countries, schizophrenia prevalence estimates vary between urban and rural areas, as outlined in section 1.2.6.

The preponderance of males in schizophrenia (Figure 1.1) was noted early on by Kraeplin (Kraepelin, 1919). Two systematic reviews have reported higher incidence in males; both estimated incidence risk ratio of men to women at around 1.4 (Aleman, Kahn, and Selten, 2003; McGrath et al., 2004). The first of these undertook a random effects meta-analysis, finding that the mean male to female incidence ratio of 1.42 (1.30-1.56 95% CI) was similar in studies showing the lowest risk of bias, at 1.39 (1.15-1.68), and individuals over 64 years old, at 1.32 (1.13-1.55). Incidence was lower in studies sampled before 1980 and not significantly different from 1 in developing countries (Aleman, Kahn, and Selten, 2003). The second found a similar median ratio of 1.40 (0.9–2.4, 10th-90th percentile) (McGrath et al., 2004). In addition to incidence, sexes also diverge in regard to presentation, course and treatment response, with men typically more severely affected and more likely to show negative symptoms, with a more favourable prognosis generally ascribed to women (Abel, Drake, and Goldstein, 2010). Despite clear, robust sex differences in incidence, differences in prevalence have not been observed (Saha et al., 2005; McGrath et al., 2008; Charlson et al., 2018).

Age of onset for schizophrenia is reasonably broad, typically occurring around late adolescence and early adulthood with onset less frequently commencing after the age of 45, though it can and does occur outside these values. Onset also differs by sex (Figure 1.1). On average, men receive a diagnosis at a younger age. Typically reported values for average age of onset of around 28 in men and 32 in women are the mean (Burton, 2016); however, right-tailed distributions are observed for both sexes. Age of onset for males has a strong modal peak around 21-25 (Jones, 2013). The distribution of onset in women is bimodal and less concentrated in late adolescence and early adulthood, with the first mode 3-5 years later than in men at around 25-29, and the second at around 45-49 years which is suggested to be related to menopause (Jones, 2013).

1.2.4 Diagnosis

As with other psychiatric disorders, diagnosis of schizophrenia relies on symptoms; there is no clinical diagnostic test. Assessment is done through psychiatric history and a mental state examination (MSE). The psychiatric history includes not just that of the individual and family history of mental disorders, but also medical history and past drug and substance use. The MSE is an informal assessment principally undertaken through questioning and observation, in addition to formal cognitive tests where required. Areas of assessment fall under appearance, behaviour, mood, speech, thoughts, perception, cognition and insight (Burton, 2016).

Diagnostic procedures can be assessed for their reliability and validity. The former describes consistency through agreement of independent assessments of the same individuals, while the latter refers to the ability of a tool to assess its target outcome. The issue of reliability was studied by several groups prior to standardisation of diagnostic procedures (Spitzer, Fleiss, et al., 1974). Most of all, reliability was brought to the fore by the US-UK Diagnostic

Project, in which a key study found individuals were more likely to be diagnosed with schizophrenia in New York than London, and that use of standardised procedures - the World Health Organisation (WHO) international statistical classification of diseases (ICD) - greatly reduced these differences (Cooper et al., 1972). The two current diagnostic tools are ICD-10 (Organization, 2004) and the diagnostic and statistical manual of mental disorders 5 (DSM-5) (Association et al., 2013). Earlier forms of these were already in existence at the time of the US-UK Diagnostic Project, DSM-I having been created in 1952 and psychiatric disorders first incorporated into ICD-6 in 1949; however, it was not until the overhaul of DSM-III, led by Robert Spitzer, that diagnostic manuals found greater focus on standardisation and reliability. The pooled estimate for test-retest reliability for DSM-5 diagnoses of schizophrenia is considered good ($\kappa = 0.46$, 0.34-0.59 95% CI) (Regier et al., 2013).

ICD-10 combines symptoms into 9 groups, labelled a to i, where items a to e refer to delusions and hallucinations centred around FRS, f to disorganised speech, g to catatonic behaviours, and h and i to negative symptoms and behavioural changes. Diagnosis by ICD-10 criteria requires one or more clear symptom from a-d or two or more symptoms from e to h. The resulting diagnosis may be one of 9 subtypes: paranoid, hebephrenic, catatonic or undifferentiated schizophrenia, post-schizophrenic depression, residual, simple or other schizophrenia, and schizophrenia unspecified. DSM-5 breaks symptoms down into 5 options: psychosis (delusions, hallucinations), cognitive symptoms (as disorganised speech or behaviour) and negative symptoms. Two symptoms must be present, with at least one being delusions, hallucinations or disorganised speech. Other possible diagnoses, such as affective disorders, must be ruled-out, the symptoms present for over a month, and work or social life must be disrupted for a diagnosis to be given.

DSM-5 and ICD-10 can both be used to frame a diagnosis of schizophrenia. They contrast principally in their use of subtypes of schizophrenia, which are present in ICD-10 but absent in DSM-5, and preference for Schneider's first rank symptoms, which are less prominent in DSM-5 (Burton, 2016). Many studies have also used the DSM-IV or DSM-IV text revision (TR); these are similar to DSM-5, but have more emphasis on first rank symptoms, presence of schizophrenia subtypes, and do not require one of the two symptoms to be delusions, hallucinations or disorganised speech (Tandon, 2013). ICD-10 and DSM-IV show good reliability for schizophrenia, with diagnoses more common in ICD-10 (Cheniaux, Landeira-Fernandez, and Versiani, 2009). ICD-11, not due to be adopted until 2022 at the earliest, reformulates the coding of schizophrenia to remove subtypes, replacing them with first episode, multiple episode and continuous, in addition to "other specified episode" and unspecified schizophrenia. Coding for ICD-11 relies on a dimensional approach, recording individuals as being mild, moderate or severe for positive, negative, cognitive, psychomotor, depressive or manic symptoms.

Diagnosis in psychiatric genetics uses DSM-IV, DSM-IV text revision (TR), DSM-5 or ICD-10 criteria, though may use older revised editions of DSM or ICD. Diagnosis in research is often achieved partly through a structured or semi-structured interview, such as the diagnostic interview for genetics studies (DIGS) (Nurnberger et al., 1994), schedules for clinical assessment in neuropsychiatry (SCAN) (Wing et al., 1990), schedule for affective disorders and schizophrenia (SADS) (Endicott and Spitzer, 1987) or the structured clinical interview for DSM (SCID) (First, 2014). The gold standard is a consensus best estimate diagnosis from two independent psychiatrists using results from a structured clinical interview, family psychiatric history and medical records, which has higher sensitivity than interview alone (Kosten and Rounsaville, 1992).

Schizophrenia shares symptoms with many other psychiatric disorders; the border between schizophrenia and other disorders, or even the absence of a disorder, can be ambiguous. Many features overlap between schizophrenia and the affective psychoses from which it was detached by Kraeplin. In particular, schizoaffective disorder and bipolar affective disorder both consist of psychosis with an affective component. Diagnosis of schizoaffective disorder requires equal prominence of affective and schizophrenia symptoms, while bipolar disorder gives primacy to mood symptoms. Schizotypal disorder, induced delusional disorder, persistent delusional disorder, acute or brief psychotic disorder and schizophreniform disorder (DSM-5 only) are also part of the psychiatric differential for schizophrenia (Burton, 2016). Schizoaffective disorder is often combined with schizophrenia to form cases in case-control studies.

1.2.5 Genetic risk

Schizophrenia shows evidence of familial aggregation. That schizophrenia runs in families is not a recent discovery. The phenomenon had long been observed informally for mental illness. Bethlem Hospital, for instance, included an assessment of whether illness was inherited as early as 1820 (Plomin et al., 2013); it is difficult to trace knowledge of inheritance in schizophrenia back further, as clearly identifiable descriptions of schizophrenia do not appear until the 1800s (Gottesman, 1991). However, with clinical descriptions of schizophrenia and the growth of Victorian asylums came careful collection and study of data on heredity and psychiatry on a scale which attracted statisticians such as Karl Pearson, and led to the foundations of the quantification of the genetic component in psychiatric disorders, in addition to fodder for ideas on eugenics and "feeblemindedness" (Porter, 2020). Genetic risk for schizophrenia is now understood to be highly polygenic. It is commonly interpreted under a liability-threshold model, whereby common and rare variants confer liability to the disorder and only those surpassing a threshold of risk meet diagnostic criteria for schizophrenia (Falconer, 1965; Gottesman and Shields, 1972) (Figure 1.2).

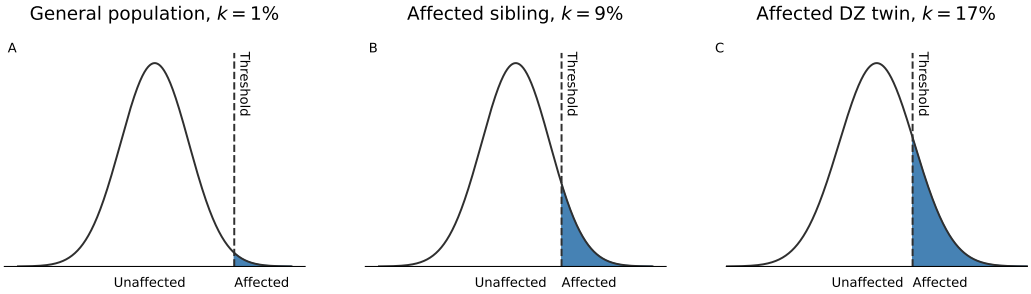


FIGURE 1.2: Liability-threshold models assuming normally distributed polygenic liability to schizophrenia. The prevalence, k, is used to derive the threshold at which individuals become affected. DZ: dizygotic (fraternal) twins. Varying k are given using risks in relatives from Gottesman, 1991.

1.2.5.1 Family studies

Ernst Rüdin, working alongside Kraeplin, is considered to have carried out the first systematic family studies (Gejman, Sanders, and Kendler, 2011), while more comprehensive assessments of families were later undertaken by Franz J. Kallmann (Gottesman, 1991). Family studies which track the pedigree of affected individuals (probands) indicate that risk of schizophrenia increases with genetic relatedness. Pooling of 40 studies spanning a large part of the 20th century illustrates these effects (Figure 1.3) (Gottesman, 1991). Risk is highest for monozygotic (MZ) twins (48%), then progressively decreases for first, second and third degree relatives, suggesting patterns of inheritance follow genetic relatedness. Risk differs by group within first and second degree relatives. In particular, parents show lowest risk (6%), followed by siblings (9%), children (13%) and dizygotic (DZ) twins (17%). The low risk in parents is likely due to effects of selection against presence of schizophrenia in marriage and parenthood. Values of siblings may be lower than children as siblings may show deviations from the expected 50% genetic overlap due to crossing over in meiosis, while DZ twins likely show the highest risk due to shared prenatal environment and identical ages (Gottesman, 1991). Though presence of familiarity is essential for showing genetic influence, it does not distinguish between genetic and environmental factors. To exclude the chance of shared environment causing elevated risk in families, twin and adoption studies are required.

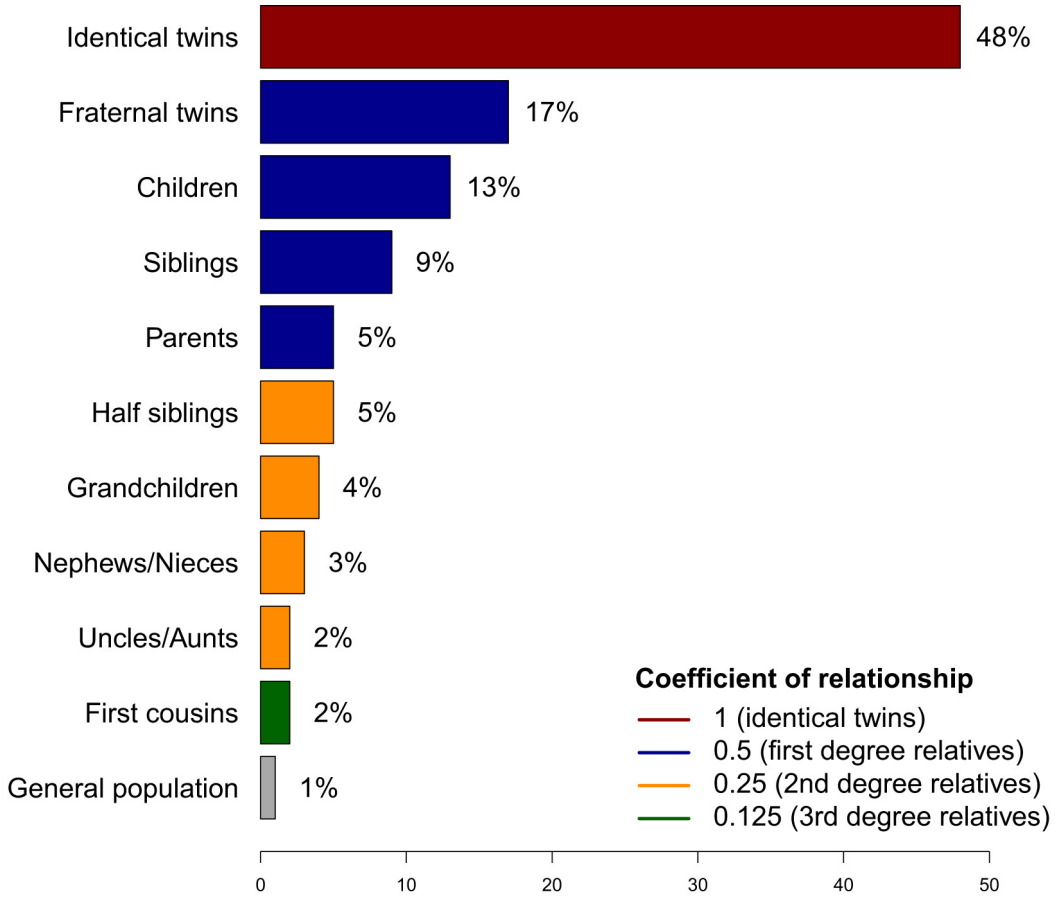


FIGURE 1.3: Risk in relatives. Risk is shown for decreasing relatedness and compared to population lifetime risk of 1%. Reproduced from Baselmans et al., 2020; data from Gottesman, 1991.

1.2.5.2 Twin studies

Twin studies support findings from family studies. Twinning occurs naturally and results in a pair of individuals which derive from the same zygote which splits (monozygotic or MZ twins, also called identical twins) or from separately fertilised zygotes (dizygotic or DZ twins, also called fraternal twins). Around a third of twins are MZ, a third same-sex DZ and the remaining opposite sex DZ. The majority of MZ twins share the same chorion and so share a more similar prenatal environment to the third of MZ twins with separate chorions (Phillips, 1993). Their study is made easier through now commonplace twin registers, such as the Bethlem and Maudsley Twin Register established in 1948. Monozygotic (MZ) twins have 48% concordance for schizophrenia and DZ twins have 17%, implying a strong genetic component as MZ twins are genetically identical, excepting *de novo* mutations and epigenetic variation, while DZ twins are 1st degree relatives sharing 50% of their DNA on average.

Twins allow for more than comparison of concordances; studies can be used to provide a measure of broad-sense heritability, H^2, the proportion of variance in a trait explained by the combined influence of all genetic factors. This improves on family studies by separating genetic and environmental factors and in its simplest form quantifies the relationship by

taking twice the difference of DZ and MZ twin concordances. If MZ and DZ concordances are equal, $H^2 = 0$, and when MZ concordance is 1 and DZ 0.5, $H^2 = 1$. In schizophrenia, concordances are converted to correlations to assess heritability of liability to the disorder. This produces a H^2 of around 0.6 for schizophrenia using the concordances of MZ and DZ twins from Gottesman, 1991, but these pooled estimates of concordance include studies from the first half of the 20th century, before reliability became central to diagnostic criteria. Studies employing standardised assessments have found heritability for schizophrenia be around 0.80-0.85 (Cardno and Gottesman, 2000). Subsequent meta-analysis of 12 twin studies produced a revised estimate of heritability for liability to schizophrenia of 0.81 (0.73-0.90 95% CI) (Sullivan, Kendler, and Neale, 2003). This indicates the majority of variance in liability can be explained by genetic differences. Caution is required in interpretation of heritability as measures are specific to the studied population during the period they were assessed for the observed outcome; the relative influence of genetics and environment can change as environment becomes more uniform, known as equalising environments, such as through the same standard of healthcare and education becoming available to all children, causing the relative impact of genetic factors to be higher. The reverse may also occur where inequality drives large differences in environment and so decreases measures of heritability. In addition, heritability accounts only for differences between individuals in a group, and cannot indicate the proportion of risk explained by genetics in a single person (Plomin et al., 2013).

1.2.5.3 Adoption studies

Results from adoption studies concur with those of twin studies. They provide a method of examining the effects of genetics and environment by taking advantage of another commonly occurring process, and were first used to investigate IQ (Senden Theis, 1924). Offspring of individuals with schizophrenia may be adopted by parents without schizophrenia, and vice versa. This can be used to compare the influence of genetics and the prenatal environment against post-natal environmental factors. Adoption studies allow for comparison of the offspring with "genetic" parents which are biologically related but did not raise the offspring, and with "environmental" parents which adopted the child but are unrelated. The similarity of the biological or environmental parent with the child informs of the genetic and environmental contribution. Where environmental and genetic parents also have biological children, the effects of the family environment can be assessed similarly using "genetic" siblings which have been adopted apart and "environmental" siblings which have been brought together through adoption. These contrast with family members in normal family studies, where parents and siblings are both "genetics plus environment". Design of adoption studies may follow the "adoptees' study method" by looking at the children of affected individuals which have been adopted by unaffected parents, the "adoptees' family method" which takes affected offspring and compares presence of the outcome in their biological and adoptive families, or more complicated designs.

Despite evidence of heritability from twin studies, the early influence of parents and family environment were believed to be the primary cause of schizophrenia prior to adoption studies (Plomin et al., 2013). Adoption studies were first used in schizophrenia by Leonard Heston using the adoptees' study method. He compared 47 offspring adopted-away from a mother with schizophrenia and matched offspring of unaffected parents, finding no events in biological offspring of unaffected parents, and 11% incidence in biological offspring of affected parents (Heston, 1966). This is only slightly lower than the 13% risk for children of affected parents given previously; these findings do not agree with any notion of schizophrenia as induced by the home environment created by parents with schizophrenia. Subsequent studies have supported this. In one such study, researchers also used the adoptees' study method to investigate individuals in a Danish sample (Rosenthal et al., 1968). Unlike Heston, they utilised affected mothers and fathers. They followed-up those that had been hospitalised and had offspring adopted-away, and compared to adopted-away offspring of unaffected biological parents; however, unaffected parents were classified by hospital records, with later work indicating a substantial proportion showed psychiatric symptoms. Lower incidence of 7% was found in offspring of affected biological parents than in Heston, 1966, and no events were found in adopted-away offspring of unaffected parents, replicating Heston's work. Further work in a Finnish sample also found around 10% of adopted-away offspring of affected biological parents had schizophrenia (Tienari et al., 1985), while a study using the adoptees' family method found 5% of proband and 0% of control adoptees' 1st degree relatives had schizophrenia (Kety et al., 1994). Adoption studies find no risk from being adopted by a parent with schizophrenia. Family, twin and adoption studies take different approaches to tackle the question of whether genetics influences liability to schizophrenia, yet all 3 arrive at the conclusion that a substantial portion of risk can be attributed to genetic factors.

1.2.5.4 Issues in family, twin and adoption studies

Limitations of twin, adoption and family study methods must also be acknowledged. Family studies are useful and informative, but cannot alone provide evidence of genetic influence on an outcome as they only incorporate genetic-plus-environment parents and siblings. While adoption studies can distinguish between genetics and environment, they too have drawbacks. Designs may be limited where adopted children, or the biological and environmental parents, are not representative of the broader population, though some evidence indicates similarity between studied parents and those which have not adopted children (Plomin et al., 2013). In addition, adoption studies do not directly compare genetics and environment, as prenatal risk is still contributed by biological mothers. This is of some concern for schizophrenia where, as section 1.2.7 makes clear, several risk factors occur during pregnancy; there is some evidence that this effect may be minimal as paternal and maternal half-siblings share similar risk (Kety, 1987). Finally, findings from adoption studies may be muddied by introduction of correlation between biological and adoptive environments, such as through selective placement or "open" adoptions where contact can occur between

biological and adoptive families (Plomin et al., 2013). Twin studies face similar obstacles in that twins may not be representative of the general population as they are typically born premature with lower birth weight than singletons (Phillips, 1993), have differences in brain structure (Knickmeyer et al., 2011), show lower IQ (Deary et al., 2005; Ronalds, De Stavola, and Leon, 2005; Eriksen, Sundet, and Tambs, 2012), and may be distinctly different from each other in that MZ twins show bigger birth weight differences from each other than DZ twins, with differences most pronounced between monochorionic and dichorionic MZ twins (Corey et al., 1979). However, twinning does not appear to be associated with schizophrenia (Rosenthal, 1960; Kendler et al., 1996). Studies have used physical characteristics to distinguish between MZ and DZ twins before genetic techniques were available. In addition, the increase in concordance from DZ to MZ could be partly due to environmental aspects such as more similar treatment due to similar appearance, though the equal environments assumption holds that environment is similar between DZ and MZ twins and tests of this have found it justified (Plomin et al., 2013).

1.2.5.5 Linkage studies

Family, adoption and twin studies indicate not just that schizophrenia is partly genetic, but that it is likely to be polygenic. Risk patterns in family studies do not agree with those expected under monogenic dominant or recessive patterns of inheritance. For example, risk in offspring where both parents are affected is just below 50% (Gottesman, 1991), while such a cross should produce 100% or 75% of offspring with schizophrenia for recessive and dominant Mendelian inheritance respectively. Under a highly polygenic model, multiple loci of minor effects are expected, in contrast to loci of major effects expected under monogenic model, or where there is low polygenicity (Falconer and Mackay, 1996).

Following cues that schizophrenia has a large genetic component and may be polygenic, the search for genes began. Despite much focused effort, schizophrenia would refuse to yield until larger association studies. Early efforts focused on linkage studies, which take a family-based approach to search for regions of chromosomes which segregate with the disorder in pedigrees. In the early days of the linkage era (1985-2005) (Burmeister, McInnis, and Zöllner, 2008), several results were reported from linkage studies (for example, Sherrington et al., 1988). However, later meta-analysis did not replicate findings (McGuffin et al., 1990). Overall, reports from linkage studies did not indicate loci of major effects, and the poor replicability of findings was negatively impacted by the low power of linkage studies to detect loci with small effect sizes. A meta-analysis which incorporated genome-wide linkage scans from 20 studies did indicate linkage in 2q, 6p and 8p (Lewis et al., 2003), with more regions reported in a larger and more recent meta-analysis of 32 studies (Ng et al., 2009).

1.2.5.6 Candidate gene association studies

While linkage studies employ a family-based design to identify segregating sites, association studies rely on unrelated participants. Initial association studies used a candidate gene

approach, whereby variants in genes for which there is some theorised role in the disease or disorder (biological candidate genes), or for which there is prior evidence from linkage studies (positional candidate genes), are tested for allelic or genotypic differences between cases and controls. Several promising findings emerged for schizophrenia from candidate gene studies. Many associations were identified, including *neuregulin-1 (Nrgn1)* (Stefansson et al., 2002), *d-amino acid oxidase (DAAO)* (Chumakov et al., 2002) and *DTNBP1* (Straub et al., 2002), with over 1,000 genes tested in total and over 1,700 studies undertaken by 2011 (Gejman, Sanders, and Kendler, 2011). However, meta-analysis of 14 promising candidate genes failed to replicate the findings in a European sample of around 1800 cases and 2000 controls (Sanders et al., 2008). Failed replication was likely due to low sample size and inadequate procedures for multiple testing correction (Gejman, Sanders, and Kendler, 2011).

1.2.5.7 Genome-wide association studies

Since 2007, association studies began spanning the genome (Burmeister, McInnis, and Zöllner, 2008). As with candidate gene association studies, genome-wide association studies (GWAS) rely on linkage disequilibrium (LD) between tagged variants which are incorporated into a genotyping array and causal variants, and on the common-disease common-variant hypothesis (Reich and Lander, 2001). GWAS test association for many common variants (minor allele frequency $> 1\%$) called single nucleotide polymorphisms (SNPs), which are typically bi-allelic. As they include many more SNPs than candidate gene studies, hundreds of thousands to millions, they can more span regions more densely and find associations where it is unclear which genes may have been suitable for candidate gene studies. To account for the large number of tests, GWAS are corrected for multiple testing at $5x10^{-8}$ (Dudbridge and Gusnanto, 2008).

Early studies found few loci. Two studies reported regions which fell short of genome-wide significance and had large odds ratios. Lencz et al., 2007 reported hits for *CSF2RA/SHOX* with an OR of 3.23 at $p = 3.7x10^{-7}$, while a subsequent GWAS reported the *AGBLI* gene at OR of 6.01 and $p = 1.71x10^{-6}$. (Sullivan et al., 2008). Smaller effects were reported for *ZNF804A* (O'donovan et al., 2008), at 1.12, and *ADAMTSL3* at 0.68 (Need et al., 2009). The International Schizophrenia Consortium (ISC) performed the first large meta-analysis of over 3,000 cases and around 3,500 controls, finding associated SNPs in the major histocompatibility complex (MHC) region at 6p.22.1 which passed the genome-wide significance threshold (Purcell et al., 2009). Associated SNPs were largely intergenic but causal loci were not clear due to long-range complex LD in the MHC region.

The ISC paper also examined the polygenic nature of schizophrenia by generating polygenic risk scores (PRS). These use estimates of effect sizes for common variants from a training set, derived by univariable tests of association with the outcome such as χ^2 or logistic regression. Effect sizes are used as weights in a linear combination of variants in an independent target set to create a single weighted score of additive risk. Testing PRS in the ISC data under different significance thresholds found that average PRS were higher in individuals with

schizophrenia than controls, providing strong support for a polygenic theory of schizophrenia (Purcell et al., 2009).

Several large studies followed. Findings in the MHC region were mirrored in GWAS using the molecular genetics of schizophrenia (MGS) data (Shi et al., 2009), and a GWAS using the SGENE-plus data (Stefansson et al., 2009), while the latter also highlighted *TCF4* and *NRGN* as genes of interest. Most notable are the meta-analyses using data from the psychiatric genomics consortium (PGC). While 2009 had seen the first potential loci for schizophrenia identified, 5 more were added with the first PGC study, labelled PGC1 hereon, which also implicated *microRNA 137* (*MIR137*) to give 7 loci showing genome-wide significant association with schizophrenia (Ripke et al., 2011), which was subsequently extended to 22 loci (Ripke et al., 2013).

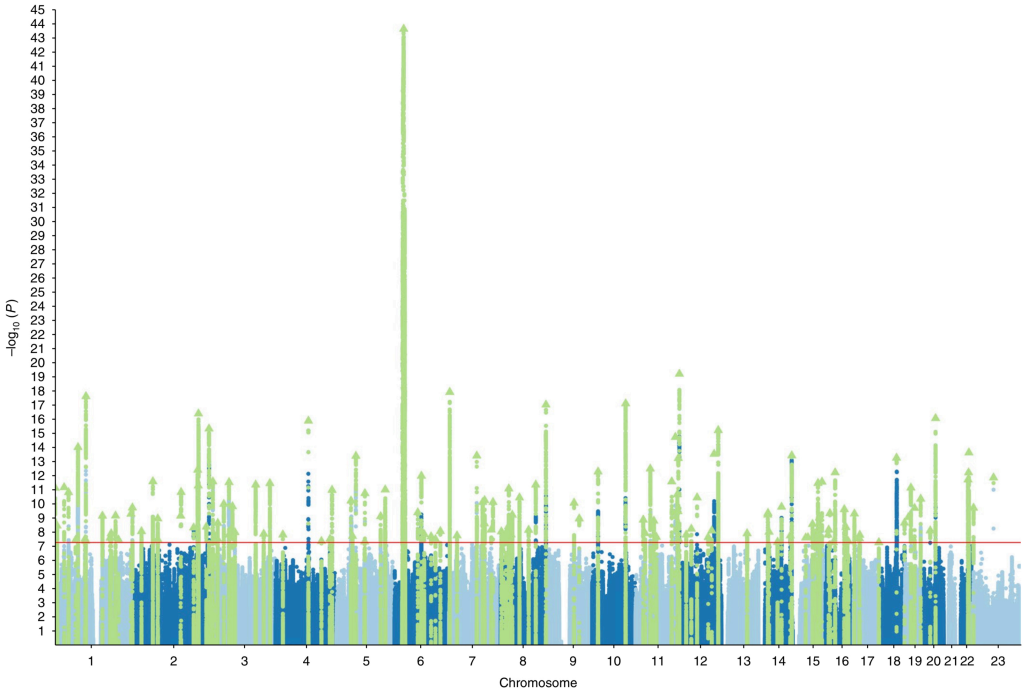


FIGURE 1.4: Manhattan plot of log(*p*-values), showing 125 genome-wide significant loci. Reproduced from Pardiñas et al., 2018.

In 2014 the number of loci reaching genome-wide significance was greatly increased and the evidence extended for a polygenic architecture in schizophrenia (PGC2; Ripke et al., 2014). Ripke et al. identified 125 genome-wide associated SNPs, mapped to 108 loci. In addition to finding a strong signal in the MHC region, supporting previous findings (Purcell et al., 2009), loci were mapped to genes including *DRD2*, *CACNA1C* and *GRM3*, supporting a role for the synapse and neurotransmitters in schizphrenia. Many were brain-expressed loci (Ripke et al., 2014), with enrichment in the CNS for excitatory and glutamatergic neurons (Finucane et al., 2018). Number of loci was extended again in the largest peer-reviewed meta-analysis of European samples (Figure 1.4). 179 SNPs passed genome-wide significance, which were mapped to 145 loci (Pardiñas et al., 2018). Using techniques including fine-mapping, these

were reduced to 33 loci potentially explaining associations. Signal from common variants was particularly enriched in loss-of-function (LoF) intolerant genes, accounting for around 30% of narrow-sense heritability.

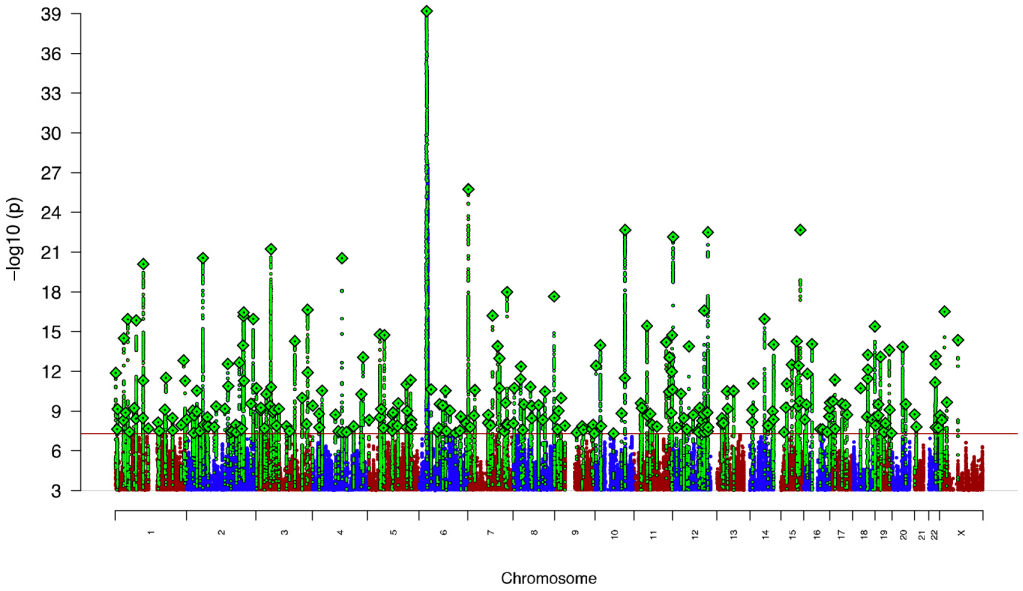


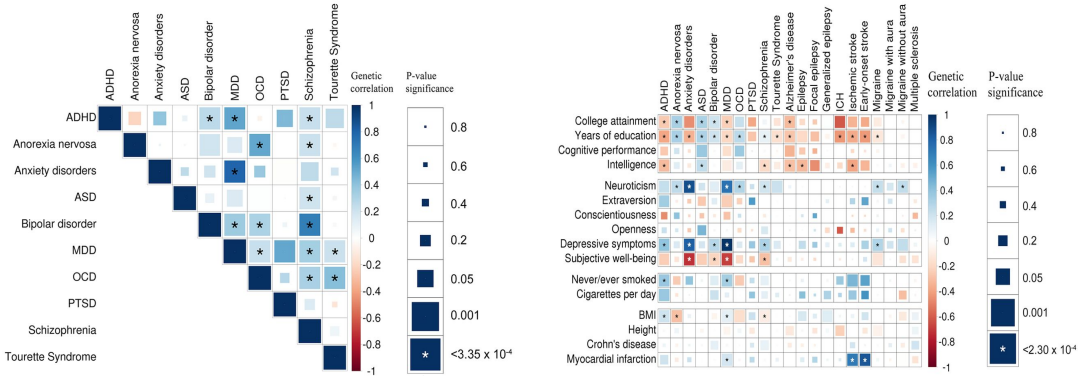
FIGURE 1.5: Manhattan plot of log(p-values), showing 329 genome-wide significant loci. Reproduced from Ripke et al., 2020.

Recent analysis from the PGC has applied GWAS of schizophrenia to East Asian samples of over 22,000 cases and 35,000 controls to identify 21 associations reaching genome-wide significance (Lam et al., 2019). The study also found similar genetic architecture between European and East Asian populations, with 208 associations identified in a meta-analysis. The overwhelming majority of association studies have been performed in European samples, though more recently GWAS have been performed in other ancestries, such as Bigdeli et al., 2020, and Lam et al., 2019, albeit with smaller sample sizes than European meta-analyses. The latest meta-analysis of schizophrenia (PGC3), available as a pre-print (Ripke et al., 2020), has identified 329 associations mapped to 270 loci using almost 70,000 cases and around 236,000 controls (Figure 1.5). Loci were further reduced to 130 using fine-mapping. Results corroborate previous findings in highlighting inhibitory and excitatory neurons and aspects of synaptic function including organisation and synaptic transmission.

A polygenic theory for the architecture of schizophrenia was suggested decades before genotyping data made its testing possible (Gottesman and Shields, 1972), and has consistently been supported by genome-wide association studies. This has been achieved partly by estimating the variance in liability (narrow-sense heritability, h^2) explained by common variants. This is under a third, with values reported at 23% (Lee et al., 2012a), 32% (Ripke et al., 2013) and 26% (Anttila et al., 2018); the current estimate from the largest European meta-analyses is 24% (Pardiñas et al., 2018; Ripke et al., 2020), assuming a prevalence of 1%. Polygenic scores also provide an assessment of polygenicity; they explain differing amounts

of variance depending on the p-value thresholds used for inclusion of SNPs. PRS derived by Ripke et al., 2014, explain a maximum of around 7% of variation in liability to schizophrenia, with just above 3% explained by genome-wide significant loci alone. Subsequent PRS analyses have supported this, reporting just below 6% R^2 on the liability scale (Pardiñas et al., 2018). The largest study to date found 7.7% of variance was explained by PRS, with 2.6% explained by GWS loci (Ripke et al., 2020).

In tandem with accruing evidence for a polygenic architecture of schizophrenia, and the involvement of dopamine, glutamate, GABA and LoF genes, there has been increasing evidence for associations between PRS and aspects of schizophrenia (Legge et al., 2021) and support for genetic correlation between psychiatric disorders (Grotzinger, 2021). Schizophrenia shows strong positive genetic correlation, r_g, with bipolar disorder ($r_g = 0.7$) (Stahl et al., 2019), weaker positive correlation with major depressive disorder ($r_g = 0.34$) and obsessive compulsive disorder ($r_g = 0.33$) (Anttila et al., 2018), and weak positive correlation of around 0.22 for autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD) and anorexia nervosa (Anttila et al., 2018). High genetic correlation between schizophrenia and bipolar disorder is well-replicated, supported by evidence from family data which show significant increased risk for either disorder in relatives of probands (Lichtenstein et al., 2009), and challenges the Kraeplinean split between the two disorders (Craddock and Owen, 2010). Schizophrenia also shows genetic correlation with cognitive function, as previously noted, BMI, cardiovascular disease, cannabis use (expanded on in section 1.2.6), volume of bain structures and personality traits such as neuroticism (Smeland et al., 2020).



(A) Psychiatic disorders

(B) Physical, behavioural and cognitive outcomes

FIGURE 1.6: Genetic correlations, r_g, between psychiatric disorders (A) and between psychiatric disorders and physical, behavioural and cognitive outcomes (B). Reproduced from (Anttila et al., 2018).

1.2.5.8 Limitations of GWAS

Linkage and candidate gene studies were both underpowered, hampered by small sample sizes and insufficiently stringent procedures for multiple comparisons, resulting in failed replications and an period of uncertainty about what genes may be involved in schizophrenia. Furthermore, the candidate gene approach relied on prior knowledge of which genes to test

for association. GWAS which combined many samples together largely ameliorated these concerns by increasing sample size, improving marker coverage and applying appropriate correction to the multiple testing problem. However, GWAS require large sample sizes to detect modest effects, replication of findings, and the reliance of GWAS on LD means the route from association to causal variant is indirect (McCarthy et al., 2008). As reported for *FTO* in obesity, connections between associated variants and risk genes are not always simple (Smemo et al., 2014). Furthermore, because of the reliance on LD, GWAS are only well-powered to consider common variants, and therefore usually do not include some of the strongest risk factors for schizophrenia.

1.2.5.9 Rare variants

Rare variation (minor allele frequency $< 1\%$) has been associated with schizophrenia as copy number variants (CNVs), single nucleotide variants (SNVs) and indels (small insertions or deletions) (Legge et al., 2021).

CNVs are large duplications or deletions in the genome that are greater than 1 kilobase (kb) (Kirov, 2015). Techniques such as array comparative genomic hybridisation (aCGH) have been utilised for their detection, as CNVs are not visible by karyotyping; they are now routinely assessed using genotyping arrays. Syndromes that result from CNVs have been labelled genomic disorders (Lupski, 1998). CNVs associated with schizophrenia show high penetrance relative to common variants (Kirov et al., 2014). The first locus to be robustly associated with schizophrenia was a CNV, 22q11.2 deletion (Murphy, Jones, and Owen, 1999), which causes velocardiofacial syndrome (VCFS), also known as DiGeorge Syndrome or 22q11.2 deletion syndrome. VCFS is a rare 1.5Mb or 3Mb deletion, occurring in less than 0.1% of foetuses, and has characteristic physical effects including hypoparathyroidism, immunodeficiency and congenital heart disease (McDonald-McGinn et al., 2015). Follow-up of adults with VCFS demonstrated an excess of psychiatric disorders, with around a quarter having schizophrenia (Murphy, Jones, and Owen, 1999). In addition to high rates of schizophrenia in individuals with VCFS, investigation of schizophrenia samples found some participants with previously unidentified deletions on 22q (Karayiorgou et al., 1995), with around 0.3% of individuals with schizophrenia suggested to have a 22q11.2 deletion (Rees et al., 2014).

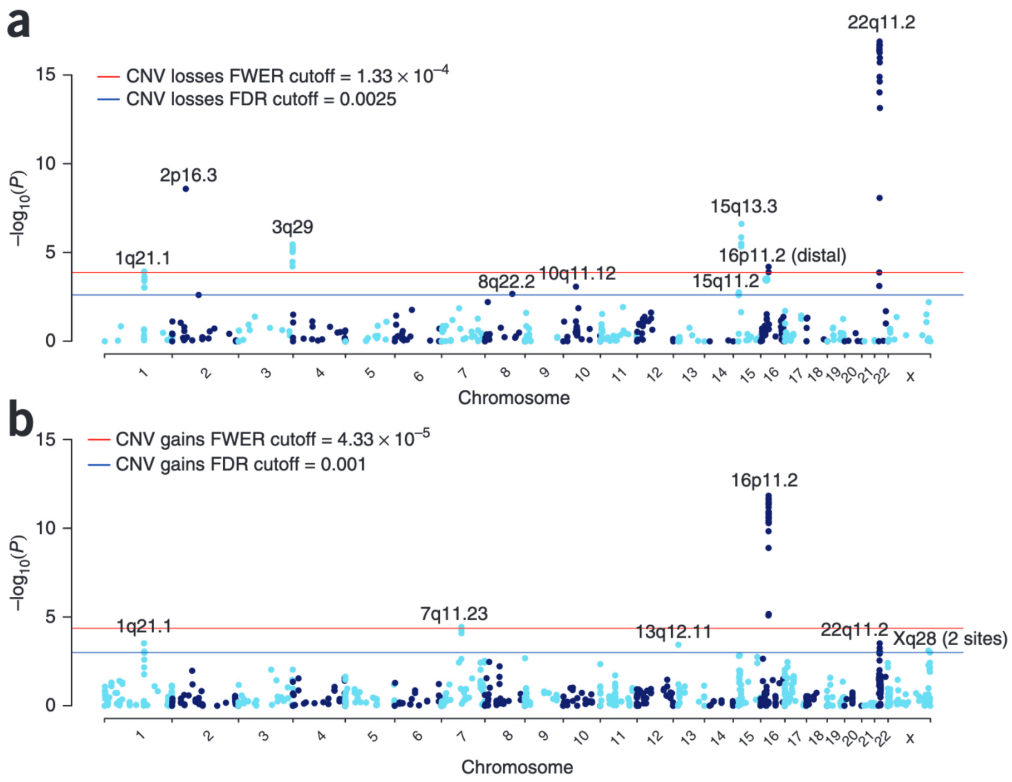


FIGURE 1.7: Manhattan plot of gene-based associations from CNV losses (a) and gains (b). Red and blue lines indicate family-wise and false discovery rate correction at 5% respectively. Reproduced from Marshall et al., 2017.

At least 12 CNVs have now been identified for association with schizophrenia, including 1q21.1 deletion and duplication, 2p16.3 (NRXN) deletion, 3q29 deletion, 7q11.23 (Williams-Beuren syndrome; WBS) duplication, 15q11.2 deletion, 15q11-q13 (Angelman/Prader–Willi Syndrome; AS/PWS) duplication, 15q13.3 deletion, 16p11.2 duplication, 16p12.1 deletion, 16p13.11 duplication and 22q11.2 deletion (Rees et al., 2014; Rees et al., 2016). Together they are present in 2.4% of individuals with schizophrenia and also result in higher risk for neurodevelopmental disorders including ASD and developmental delay (DD) (Kirov, 2015). Recent analysis of CNVs by Marshall et al., 2017, found 8 regions passing genome-wide significance: 1q21.1, 2p16.3, 3q29, 7q11.2, 15q13.3, distal 16p11.2, proximal 16p11.2 and 22q11.2 (Figure 1.7). While some odds ratios for schizophrenia in more common CNVs are low, at around 2, others are over 50 (22q11.2 deletion) or frequently cannot be calculated (3q29 deletion) when carriers are not observed in controls. Additional CNVs also show association and may become significant as sample sizes increase (Kirov et al., 2014; Marshall et al., 2017). Risk for schizophrenia in CNV carriers is also affected by common variants Tansey et al., 2016.

Rare SNVs and indels have been investigated by exome studies, typically binning variants together into genes or gene sets to improve power and reduce multiple testing. Cost and sample size remain issues for sequencing studies, with whole genome sequencing studies still limited in size. Early exome studies also suffered from small sample sizes, with number of

cases less than 100 (Avramopoulos, 2018). Drawing on a larger sample of around 5000, half of which were cases, Purcell et al., 2014, identified that distributed rare variants collectively impact risk for schizophrenia in a polygenic manner. These results are supported by later work employing around 10,000 cases and over 13,000 controls from exome genotyping arrays, with both studies highlighting targets of the FMRP protein (Leonenko et al., 2017). Ultra-rare protein-disrupting variants have also been highlighted in a Swedish sample including close to 5,000 cases (Genovese et al., 2016). Analysis of *de novo* variants, which are absent in the parents of probands, has found enrichment in LoF intolerant genes (Rees et al., 2020). Most recent analysis by the Schizophrenia Exome Sequencing Meta-Analysis (SCHEMA) Consortium, available as a pre-print, identified 10 genes containing risk variants for schizophrenia, with a range of odds ratios similar to those observed for CNVs (Singh et al., 2020). Collectively, studies outline a case for polygenic impact of rare and ultra-rare disruptive mutations, particularly in LoF intolerant genes and those involved in FMRP targets. Results of CNV and exome studies both identify risk factors with large effect sizes which confer risk for schizophrenia and developmental disorders including ASD and developmental delay (DD) (Kirov, 2015; Singh et al., 2020). Analysis from SCHEMA and PGC3 also indicate convergence of rare and common risk of schizophrenia (Ripke et al., 2020; Singh et al., 2020).

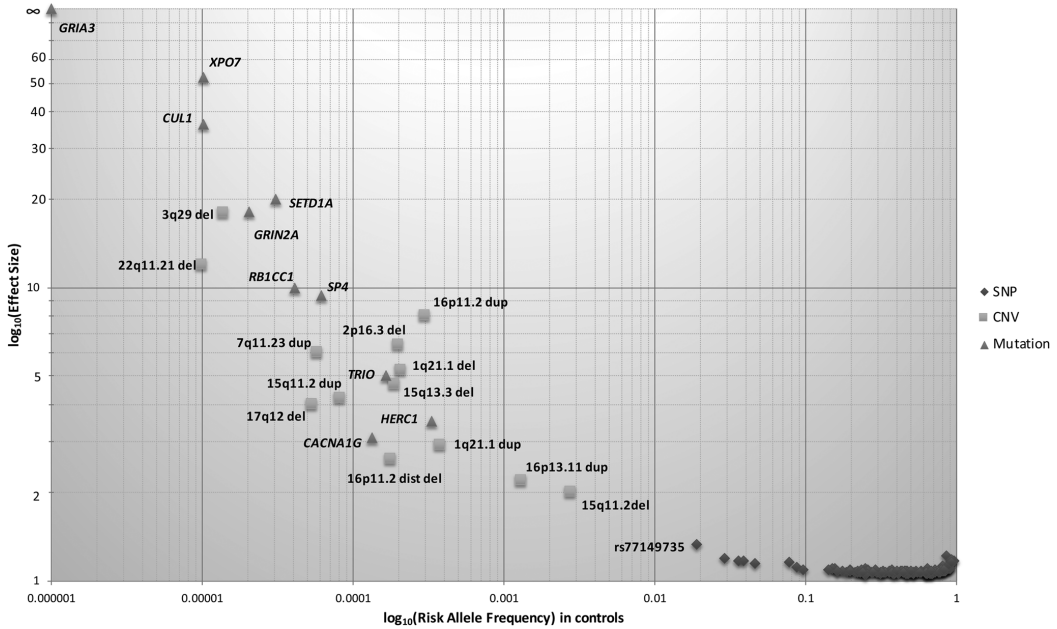


FIGURE 1.8: Distribution of effect sizes for genetic risk factors for schizophrenia. Log(effect size) is shown against log(allele frequency). Reproduced from Legge et al., 2021.

Overall, studies indicate that genetic risk for schizophrenia is complex. Family, adoption and twin studies support a large role for genetic factors in schizophrenia risk, with incidence of below 1% for schizophrenia increasing with genetic relatedness to around 10% for 1st-degree relatives and close to 50% for MZ twins. Though many early efforts in linkage and candidate gene studies found difficulty with replication, CNV studies, modern GWAS

and recent exome studies have found robust associations and have strongly reinforced the notion of schizophrenia as a polygenic disorder. Common and rare variants both contribute to the genetic architecture of schizophrenia and confer increased liability for the disorder (Figure 1.8). Genetic studies have also demonstrated widespread pleiotropy, with genetic overlap occurring between many psychiatric disorders; schizophrenia shows particularly high symptomatic and genetic correlation with bipolar disorder.

1.2.6 Environmental risk

Though heritability is high for schizophrenia, it is important to reiterate that monozygotic twin concordance is estimated at around 50%, and that affected individuals do not typically have an affected relative (Gejman, Sanders, and Kendler, 2011), leaving substantial room for non-genetic factors. Offspring of discordant MZ twins also show the same incidence of schizophrenia (Kringlen and Cramer, 1989), confirming both inherited risk in the unaffected twin despite no manifest illness, and the role of environmental factors in schizophrenia risk through discordance despite almost identical genetic risk. Non-genetic factors may be epigenetic or involve shared or non-shared environment. While genetic risk confers some liability to schizophrenia, environmental factors may further predispose or precipitate onset (Murray and Fearon, 1999). Non-genetic factors are labelled as environmental factors here, but may still be partly heritable. They are loosely grouped into prenatal and postnatal events.

1.2.7 Prenatal factors

Risk factors for schizophrenia have also centred around insults *in utero* or in delivery. These include infection and famine during pregnancy, obstetric complications and season of birth.

Maternal infection in pregnancy is associated with several neurodevelopmental conditions. Ecological studies implicated the 2nd and 3rd trimester as risk periods for influenza infection on the likelihood of schizophrenia in the foetus, mainly through investigations of the influenza pandemic of 1957 (for example, Mednick et al., 1988); however, methodological limitations necessitated the use of birth cohort studies, as ecological studies only captured mothers that were pregnant during the pandemic, and could not confirm exposure to influenza. Birth cohort studies have supported a role for prenatal infections in schizophrenia risk, including influenza, *toxoplasma gondii* and genital or reproductive infections, potentially via increases in maternal cytokines in response to infection (Brown and Derkits, 2010).

Evidence for the role of prenatal malnutrition comes primarily from the Dutch hunger winter, 1944-45, wherein extensive military records and a German blockage in part of the Netherlands combined to produce a natural experiment wherein a severe shortage of food occurred for only a limited period of time (Hoek, Brown, and Susser, 1998). Studies of this phenomenon have investigated disorders in adulthood for individuals who were exposed to the famine *in utero*. These have found associations with schizophrenia and schizophrenia spectrum

disorders, in addition to congenital CNS abnormalities (Susser and Lin, 1992; Susser et al., 1996). Risk for these was defined as exposure during the 1st trimester, though exposure to famine during the 2nd and 3rd trimesters has also been indicated for increased risk of affective psychosis in adulthood (Brown et al., 1995). A similar natural experiment occurred in China, 1959-61, which has replicated the results of increased schizophrenia risk in individuals who experienced prenatal famine (St Clair et al., 2005).

The season of birth (SOB) effect is hypothesised to be related to maternal infection. As with other epidemiological evidence for environmental risks, the association is not new; it was reported first by Tramer in 1929 with additional evidence accumulating over the 20th century (Bradbury and Miller, 1985). Strong evidence has been found for weak effects of winter birth on increased risk of a later diagnosis of schizophrenia, with odds ratios generally lower than those or other known factors such as urbanicity (Boyd, Pulver, and Stewart, 1986; Baron and Gruen, 1988; Mortensen et al., 1999). These findings have replicated in many studies in the Northern hemisphere (Davies et al., 2003), with weaker but still elevated effects in the Southern hemisphere (McGrath and Welham, 1999). The causal factors underlying the SOB effect are not clear, though maternal nutrition and infection are among those suggested (Torrey, Torrey, and Peterson, 1977).

Neurodevelopmental insults are also implicated through obstetric complications, which have been shown to occur more frequently in the births of children who go on to receive a diagnosis of schizophrenia (Clarke, Harley, and Cannon, 2006). Confidence has mounted for three areas of risk: prenatal complications such as preeclampsia and prenatal diabetes, delivery complications, and abnormal development or growth in pregnancy, evidenced by diminished head circumference, low birth weight or congenital malformations (Cannon, Jones, and Murray, 2002). Pooled effects are generally reported; identifying specific causal factors has proved difficult, though hypoxia has been suggested as the likely common event in prenatal and delivery complications (Cannon, Jones, and Murray, 2002), and findings in some studies may be skewed by factors such as recall bias, where mothers have been asked to recall obstetric complications (Geddes and Lawrie, 1995). Prenatal factors, including malnutrition, obstetric complications and maternal infection have been noted for their inconsistent replication in regards to effects and which trimester is affected (Van Os, Kenis, and Rutten, 2010). Furthermore, as mentioned in section 1.2.5, findings from adoption studies show similar risk for maternal and paternal half-siblings of adoptees with schizophrenia, with both showing just above 15% risk for half siblings of affected adoptees and less than 5% for half-siblings of unaffected adoptees, suggesting a lower impact from the prenatal environment (Kety, 1987).

1.2.7.1 Postnatal factors

Attempts to apply psychodynamic theories to explain schizophrenia have met with some success. The idea of stressful life events, which grew out of such theories, found that good or bad events which typically cause stress on individuals, such as marriage or death of a

loved one, are more common in those which go on to experience schizophrenia (Brown and Birley, 1968; Day et al., 1987). Not all associations have been replicated (Norman and Malla, 1993a) and assessment of life events is made challenging by defining stressful events, as patients often show cognitive disturbances and worsening of function in the prelude to diagnosis, and so events such as job loss may be consequence, rather than cause (Norman and Malla, 1993b). Attempts to explain schizophrenia through family risk, also a consequence of psychodynamic theories, have found more mixed results. Higher expressed emotion (EE) in households has been linked to higher likelihood of relapse (Brown, Birley, and Wing, 1972), though not as an aetiological factor. This finding has been well replicated; EE is now known to be one of the most robust predictors of relapse in schizophrenia (Bebbington and Kuipers, 1994; Butzlaff and Hooley, 1998). However, the suggestion that abnormal speech in schizophrenia is a learned behaviour through unclear and conflicting messages from parents during childhood (Wynne and Singer, 1963) was popular and popularised in the ideas of R.D. Laing, but was not replicated in later work (Hirsch and Leff, 1975). Psychodynamic ideas of family risk also led to mother-blaming that is typical for neurodevelopmental disorders such as ASD, with the damaging and now roundly disproved term "schizophrenogenic mother", coined by Frieda Fromm-Reichmann in 1948, becoming widely known.

Some of the most long-standing and robust evidence for environmental risk in schizophrenia has come from epidemiological studies. Two of these, already mentioned in discussions of incidence, are sex and geography, with risk greater in males and associated with country, especially latitude. Geography is also implicated through urbanicity, as incidence of schizophrenia is higher in more urban areas (McGrath et al., 2004). Early evidence of this association has largely held up through subsequent studies (Vassos et al., 2012). Explanations revolve around the so-called "breeder" and "drifter" or "social drift" hypotheses. The former suggests the urban environment plays a causal role in schizophrenia, while the latter implicates prior risk of schizophrenia as causing individuals to move into urban areas; a similar argument, that unaffected individuals are more likely to move out from urban to rural areas is known as the "social residue" hypothesis. The cause of the urbanicity-schizophrenia association is unknown. Social factors associated with urban environments such as greater population density, overcrowding, and deprivation may be involved, while other potential causal factors in development such as infection, pollution and obstetric complications have also been suggested (Murray et al., 2002).

Migration has been associated with increased risk of schizophrenia. Early evidence for this came from studies of migrants to the United States (Ødegaard, 1932). In the UK, the most well-investigated findings come from second generation descendants of Afro-Caribbean migrants to the UK, for which an increased risk of schizophrenia is observed (Harrison et al., 1988). Though studies have been less successful for some other migrant groups to the UK (Harrison, 1990), the effect of migration status is seen in other countries for other migrant groups, implicating migration rather than a specific population of individuals. Causal factors in migration may be somewhat linked to urbancity, as migrants are more likely to

reside in cities, however, the association of migration is only partly reduced by accounting for urbanicity (Murray et al., 2002). It has been suggested that individuals who migrate may be more likely to have less family ties or support or more deprivation before uprooting, and so may already be at increased risk for schizophrenia prior to migration, known as Ødegaard's selective migration hypothesis (Ødegaard, 1932). This has not received support from studies (Selten et al., 2002; Ven et al., 2015). Instead, the explanation of increased risk from social exclusion, isolation and racism is favoured (Henssler et al., 2019). Such a suggestion ties-in with the higher incidence of schizophrenia observed in ethnic minorities, which is elevated when minorities constitute a smaller proportion of the population in a region (Boydell et al., 2001). Increased likelihood of Afro-Carribean migrants to the UK to encounter psychiatric services or be misdiagnosed may contribute to the effect (Sharpley et al., 2001). Findings of urbanicity and migration may also be related to deprivation. Indices of social deprivation, which calculate a weighted score to summarise the degree of deprivation in a region, have been related to higher incidence of schizophrenia both at birth and presentation (O'Donoghue, Roche, and Lane, 2016). Together these findings implicate the social environment in elevated risk for schizophrenia.

Increased cannabis use has often been noted in individuals with schizophrenia. This association has received plausible arguments for a causal role of liability to schizophrenia in increasing the likelihood that an individual would seek out cannabis, or of cannabis consumption increasing schizophrenia risk in itself, where the former may be due to common causal factors shared between schizophrenia and cannabis use, or that cannabis may be sought to alleviate symptoms of schizophrenia. Evidence from early studies obtaining information on substance use prior to onset suggested a causal relationship. Some of the strongest evidence to this effect has been found in studies of Swedish conscripts (Andréasson et al., 1987; Zammit et al., 2002) and longitudinal studies (Moore et al., 2007). More recent analysis of genetic data indicate schizophrenia increases risk of cannabis use. A Mendelian randomisation study found evidence for schizophrenia increasing risk of cannabis use, but also weaker evidence of increased schizophrenia risk as a result of cannabis use (Gage et al., 2017), while another Mendelian randomisation study found schizophrenia increases risk of cannabis use (Vaucher et al., 2018). The most well-powered study to detect effects found genetic overlap between cannabis use and schizophrenia, and evidence for a causal role of schizophrenia in risk of cannabis use (Pasman et al., 2018). Subsequent analysis is consistent with the interpretation that genetic liability to schizophrenia increases risk of cannabis use (Jones et al., 2020).

An additional factor in early childhood associated with increased risk for schizophrenia is childhood trauma. Such studies posit that trauma experienced in childhood, typically including sexual or verbal abuse, parental neglect or bullying, can increase risk for later diagnosis of schizophrenia. Though early work found inconsistent results (Van Os, Kenis, and Rutten, 2010), subsequent longitudinal studies have shown that childhood events inflicted by parents or peers (Arseneault et al., 2011), or peers alone (Schreier et al., 2009) increase the

likelihood of reporting symptoms of psychosis in early adolescence, including after correcting for environmental factors and genetic risk (Arseneault et al., 2011). A meta-analysis of childhood adverse events found these associations are similar across study designs (Varese et al., 2012). Childhood trauma is also associated with increasing severity of hallucinations and delusions (Bailey et al., 2018).

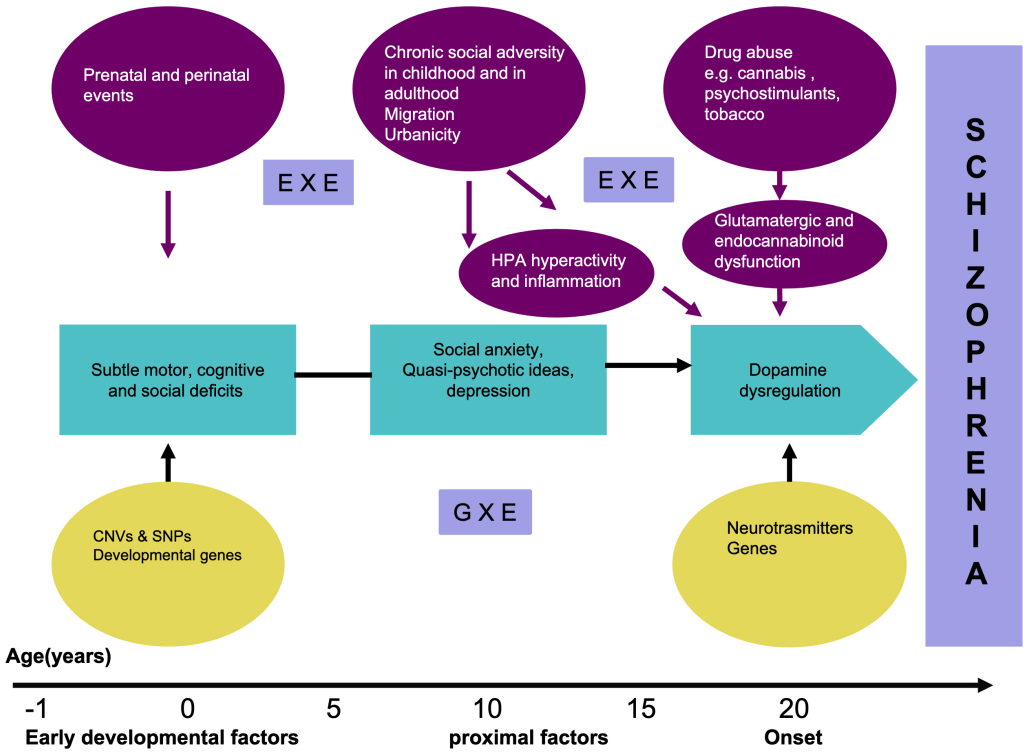


FIGURE 1.9: The interplay of genetic and environmental risk factors. Potential environmental interactions (ExE) and gene-environment interactions (GxE) are annotated. Reproduced from Stilo and Murray, 2019.

Collectively, environmental risk factors contribute significantly to liability to schizophrenia, with risk converging on insult in the intrauterine environment and adverse social experiences during childhood and early adolescence. These findings support the notion of schizophrenia as a complex disorder influenced by both genetic and non-genetic risk (Figure 1.9). They also tie-in with theories of increasing predisposition to schizophrenia from both genetic and environmental factors, suggested to occur via dysregulation of the dopamine system (Howes and Kapur, 2009), such that stressful events have a far greater impact and precipitate psychosis (Howes et al., 2017).

1.3 Prediction

1.3.1 Does psychiatry need prediction?

Prediction modelling and explanatory modelling are distinct endeavours. As elaborated in chapter 2, the latter is often the focus of statistical methods which aim to draw inference

about a population by fitting a model to a sample; in a frequentist setting, these may be used to test hypotheses about the process under study (Bzdok, Altman, and Krzywinski, 2018). By contrast, prediction aims to assign a value to unobserved data, and so aligns more clearly with the aims of precision medicine (Bzdok, Varoquaux, and Steyerberg, 2020). The distinction is important as increasingly large datasets allow for reporting of p-values which fall below significance thresholds despite small effect sizes; such associations may be genuine yet not predictive (Bzdok, Engemann, and Thirion, 2020).

Leo Breiman, who developed random forests (Breiman, 2001), bagging (Breiman, 1996) and introduced a form of decision tree alongside Friedman and Stone (Breiman et al., 1984), wrote of the distinction between what he called "the two cultures" of statistical modelling, prompting much discussion (Breiman et al., 2001). Breiman, whose experience is coloured by time spent on both abstract mathematics and applied statistics, in academia and industry, argued against excessive focus on "data models", such as regression models, which assume a stochastic data generating process and use this to estimate coefficients for use in explanation and possibly prediction. Instead, he advocated for a shift to include more "algorithmic models" which regard this process as unknown, and suggested a wider statistical toolset focused on practical problem solving would prove fruitful.

Prediction is an important part of achieving improved outcomes in psychiatry, and machine learning approaches are one, but by no means the only, method of achieving this.

1.3.2 Machine learning for precision psychiatry

The picture described in section 1.2 is one of complexity. Many factors have been associated with liability to schizophrenia, though none alone are sufficient to cause the disorder. As it stands, univariable tests of association are commonplace, despite longstanding critiques of such practices from statisticians (Sun, Shook, and Kay, 1996; Harrell Jr, 2015; Steyerberg et al., 2019). A drive for both prediction and multivariable models is therefore warranted. Faced with a multifactorial outcome, supervised machine learning is a promising tool for several reasons.

First, machine learning puts firm focus on prediction. Machine learning methods trade-off between bias and variance to improve prediction; this approach can partially account for the "winner's curse" that sees effect sizes from association studies decrease upon replication, as estimation in the training data can be constrained in order to shrink coefficient estimates. Approaches also employ training procedures to optimise for prediction, including resampling approaches which allowing the tuning of models to maximise generalisation, not association.

Second, machine learning approaches can handle datasets where predictors, p, outnumber observations, n, known as the $p > n$ scenario, allowing for use of multivariable prediction models without resorting to univariable tests of association. In psychiatry, a large increase in the volume of available data has been observed and is expected to continue (Monteith et al., 2015); machine learning is a potential method for dealing with this (Iniesta, Stahl,

and McGuffin, 2016). The $p >> n$ case, a situation where the number of predictors is much greater than the number of observations, is a common situation in genomics and other omics areas where a huge number of predictors are assayed. In particular, models which use embedded predictor selection, discussed in chapter 2, may perform well when only a portion of predictors are associated in high dimensions.

Third, machine learning allows for flexibility. Though some methods are linear additive models, many are able to learn complex patterns including non-linear predictor-response relationships and interactions between predictors. While additive models are common in schizophrenia, interactions may explain aspects of "missing heritability" (Eichler et al., 2010; Woo et al., 2017), a term arising from the gap between twin study estimates of heritability and variance explained by significant GWAS loci. Gene-gene and gene-environment interactions have been reported in schizophrenia (Nicodemus et al., 2014; Bernardo et al., 2017); gene-environment interactions may even be expected given suggested roles of genetic liability and environmental stresses in the aetiology of schizophrenia, as discussed in section 1.2 (Van Os, Kenis, and Rutten, 2010). Notably, these features of machine learning contrast with PRS, which typically apply no shrinkage to coefficient estimates, use univariable tests of association and do not consider deviations from additivity within or between loci.

1.3.3 Challenges in machine learning

There are several challenges remaining for applying machine learning in psychiatry. Great successes in machine learning have typically come off the back of a triad of methodological improvements in modelling (for instance, Krizhevsky, Sutskever, and Hinton, 2012), increased computational power, and the greater availability of large well-labelled datasets (Halevy, Norvig, and Pereira, 2009). While strides have been made in adapting machine learning approaches to high performance computing (HPC) clusters (Herrera et al., 2019; Bayat et al., 2020), algorithms are typically optimised to handle a large number of observations, not the extremely large number of predictors seen in genomics (Genuer et al., 2017). An area where deep learning has excelled is image recognition, where it now exceeds human performance. ImageNet, a standard image dataset often used for pre-training deep learning models known as convolutional neural networks, contains over 14 million images and 20,000 labels, which have been carefully manually curated (Deng et al., 2009). By contrast, the most recent analyses in schizophrenia have obtained impressive but much smaller sizes of 69,369 and 24,248 cases for GWAS and exome data respectively (Ripke et al., 2020; Singh et al., 2020), and the strategy for increasing dataset sizes in genetics is often partly to relax outcome definitions. Compounded with the potential for misclassification in psychiatry, and the absence of any objective laboratory test to confirm diagnoses, it is unclear whether machine learning models can repeat previous successes in psychiatric genetics.

Methodological considerations must also be raised. As examined in chapter 3, application of machine learning to psychiatric genetics is relatively recent and has not been widely adopted. Teething problems are expected when any methodology is introduced to a field. Early issues

in GWAS of schizophrenia, for example, have now been ironed-out (Flint and Munafo, 2014). At the moment it is not only unclear what the predictive performance is of machine learning methods in psychiatric genetics, but also how and how well they are being applied.

1.4 Aims

The aim of the thesis is to systematically evaluate the predictive performance of a range of machine learning models for prediction of schizophrenia. This will be achieved through an assessment of predictive performance and potential bias in the literature, examination of how approaches perform under simulations of main and interaction effects, and application to genetic and non-genetic predictors in real data.

1.5 Outline of thesis

Chapter 2 gives an overview of machine learning methods used in the thesis, including techniques for the training and evaluation of models. After considering the general principles of machine learning which are germane to the thesis, aspects of model development and validation are discussed, including cross-validation procedures and scalar measures of model performance. The remainder of the chapter expounds on the techniques available. With reference to logistic regression, methods including penalised regression, support vector machines, random forests, gradient boosting and deep learning are introduced.

A systematic review of machine learning models for genetic prediction in psychiatry is undertaken in chapter 3 to establish how well methods perform. Discrimination is compiled for 77 models across 13 studies, including 4 psychiatric disorders. Crucially, the review also assesses within-study risk-of-bias, a procedure which is often ignored in reviews of machine learning, but is essential for interpreting results. Methodological practices and pitfalls are also documented for all models.

The question of whether machine learning models can approximate interactions between loci and use them to improve prediction is taken-up in chapter 4. This chapter further considers additive models, the effects of introducing LD, how well models perform under both $p > n$ and $n > p$ scenarios, and whether variations of simulation parameters affect predictive performance.

Chapter 5 considers the application of machine learning approaches to genetic and non-genetic factors in real data for schizophrenia. Taking advantage of the deeply phenotyped UK Biobank dataset, models and modelling procedures are assessed for discrimination and calibration. Focus is paid to the importance of multivariable models which include genetic and environmental factors, adjustment for potential confounding from population structure, and methodological procedures for training and calibrating models in a nested case-control sample.

Much optimism surrounds machine learning's potential role in precision psychiatry. A realistic assessment of performance is necessary to guide whether its application may be useful. Previous studies have often either introduced a novel machine learning method or compared a selection of existing methods without reference to a baseline such as logistic regression or polygenic risk scores. Furthermore studies often use inadequate procedures for model development and validation. In this thesis, a wide range of supervised machine learning methods are systematically compared against polygenic risk scores and logistic regression to assess their predictive performance in schizophrenia.

Chapter 2

Methods

2.1 Introduction to machine learning

2.1.1 Learning from data

The two canonical definitions of machine learning are separated by around 40 years. In 1959, Arthur Samuel defined machine learning as the "field of study that gives computers the ability to learn without being explicitly programmed". In 1997, Tom Mitchell gave the definition that "a computer is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E". Both definitions are still relevant today, and point to a methodology that eschews hard-coding of pre-defined rules in favour of learning from data.

In this thesis, supervised machine learning is taken to be a collection of methods for estimating the function that correctly maps inputs (often genotypes) of individuals to outputs (schizophrenia status). Under this scenario, the learning problem can be constructed as shown in Figure 2.1, wherein the true target function f is unknown and a dataset $\mathcal{D}$ is available, consisting of observation-outcome combinations, $(x_1, y_1), \ldots, (x_N, y_N)$, and x_i is a scalar or vector for all $i = 1, \ldots, N$. Each observation x and outcome y is drawn from the space of possible inputs $\mathcal{X}$ and outputs $\mathcal{Y}$.

As f is unknown, the aim is to approximate f using $\mathcal{D}$ by a function g that follows the formula $g : \mathcal{X} \rightarrow \mathcal{Y}$. Obtaining g is not trivial. To do so, a learning algorithm $\mathcal{A}$ is applied which draws potential functions h from the set of all hypotheses obtainable by the learner, $\mathcal{H}$. $\mathcal{A}$ is our supervised machine learning method. The process of approximating f is learning from data.

In practical terms, the parameters of g are estimated automatically from the data. Hyperparameters are also learned empirically following input from the user. The values can be used to adjust properties of the model, and may constrain or enable flexibility of g. The options for training and evaluating the resulting models can be complex and are the subject of later sections.

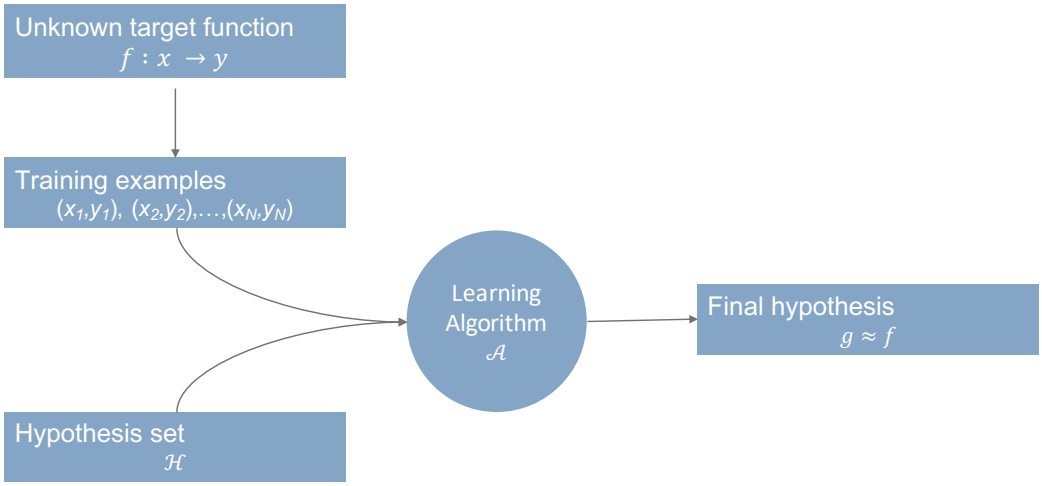


FIGURE 2.1: Supervised machine learning takes labelled training examples $(x_1, y_1), \ldots, (x_N, y_N)$ to approximate an unknown function f using a learning algorithm $\mathcal{A}$ with hypothesis space $\mathcal{H}$. The output, g, is the supervised machine learning model. Adapted from Abu-Mostafa, Magdon-Ismail, and Lin, 2012

2.1.2 Types of learning

There are several ways to divide up the array of machine learning methods into categories. Problems are labelled as either classification or regression depending on the outcome. Problems with a numerical output, such as age of onset for a disease or quantitative traits like cognitive scores are regression problems, even though the method used to solve them may not be a form of regression in the statistical sense. Problems with categorical outcomes, like disorder-status, medication response or disease staging, are classification problems.

Machine learning can also be partitioned into supervised and unsupervised learning. Supervised learning, as previously described, uses input data which come labelled with the outcome. The sample is then often split into training and test data, with the labels removed from the test data. The association with features and labels are learned from the training set and evaluated on the test set. By contrast, unsupervised learning involves unlabelled data for which we wish to detect patterns or structure in the data. Additional groups include semi-supervised learning, where some of the data are labelled, and reinforcement learning, which aims to train an agent in an environment to learn an optimal strategy, or policy, through rewards and penalties.

Methods may be further categorised as parametric and non-parametric. In a dataset with many predictors, and therefore lots of parameters to learn, parametric models can be useful. They simplify the estimation problem to a restricted number of parameters; this imposed structure makes for easier computation, a simpler and more explainable model, and may possibly mean better generalisation through reduced variance. However, it also increases bias; consequently, h is likely to be a poorer approximation of the true form of our unknown function f. Non-parametric methods can fit a wider variety of the potential forms of f as

they make fewer assumptions about the relationships in the data, but run the risk of having high variance through overfitting.

2.1.3 Machine learning, statistics and statistical learning

The distinction between machine learning and statistics can be nebulous. The definition given in section 2.1.1 could be seen to include methods such as linear and logistic regression; many of the same problems and methods can be categorised under both fields.

Generally, supervised machine learning involves steps such as hyperparameter tuning and resampling to find the predictors and model which maximise performance empirically. By contrast, statistics may be more concerned with choosing the most appropriate modelling procedures up-front in order to draw inference from a sample. The two fields also differ in language used to describe the same phenomenon. Terms such as "parameters" or "variables" are used more commonly in statistics, while "features" or "weights" are more frequently used in machine learning. Perhaps the most reasonable distinguishing features of machine learning then are both the type of method used and the way in which it is applied. A more recent suggestion is that this categorisation runs the risk of dichotomising a continuous characteristic. A potential solution to this is the machine-learning spectrum, where the degree to which something falls more under machine learning than statistics roughly correlates with a move away from interpretability. (Beam and Kohane, 2018).

The term statistical learning was coined by Robert Tibshirani and Trevor Hastie and is similar in definition to machine learning (Hastie, Tibshirani, and Friedman, 2009). While machine learning grew out of computer science and the artificial intelligence community, statistical learning grew out of statistics. Traditionally, machine learning has been mainly associated with maximising prediction, while statistical learning has had a focus on human-interpretable results and offering some insight into the problem. However, the difference in these terms has become difficult to distinguish as the machine learning community has begun to focus more effort on explainability. Partly for this reason, this thesis will only use the term machine learning, which is taken to be a sub-field of artificial intelligence (AI). The following sections further explore the trade-offs and key principles behind machine learning.

2.1.4 The bias-variance trade-off

One of the most central problems in machine learning is the bias-variance trade-off (Figure 2.2). The term applies to the change in predicted values from a model when it is repeatedly re-trained on a new data set. A highly flexible function that touches every data point will give a perfect prediction in the training set. This flexibility means there is no systematic skew in the predicted values, but that they follow the idiosyncrasies of each training set; the model has learned both the signal and the noise of the data and has overfit. When predictions vary according to both real associated changes and random fluctuations in predictors, the model has high variance.

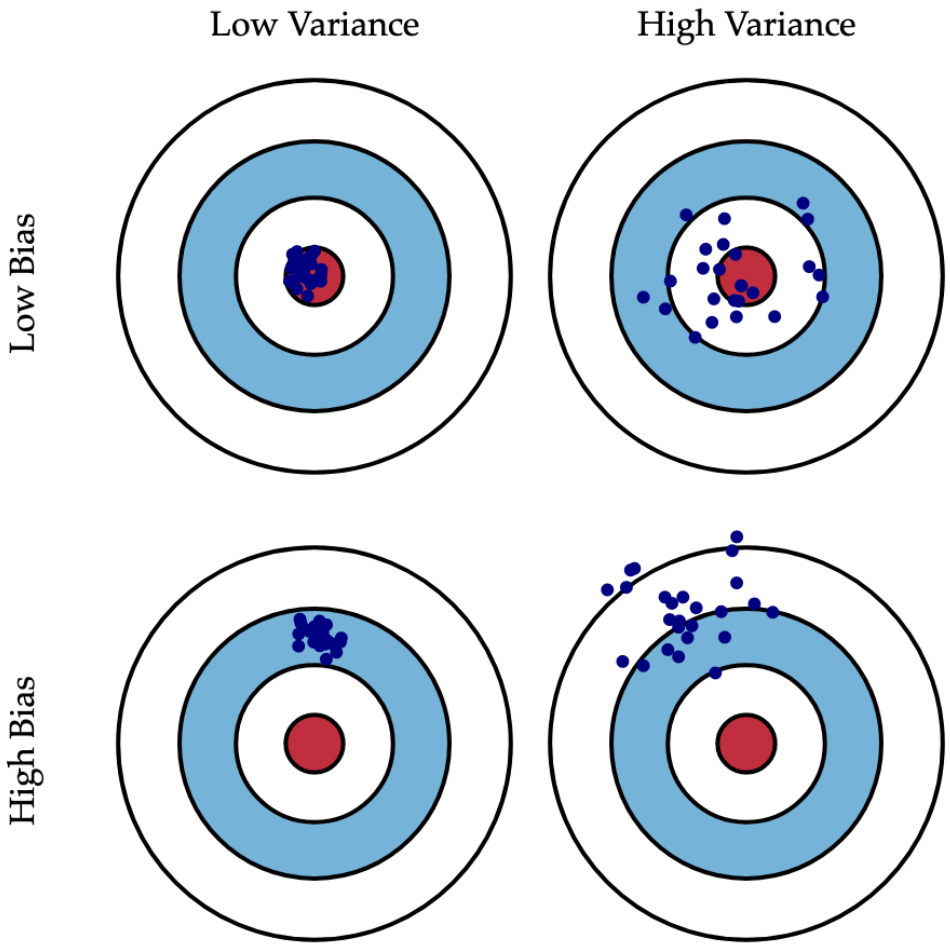


Fig. 1 Graphical illustration of bias and variance.

FIGURE 2.2: The bias-variance trade-off. Targets illustrate the concept of predictions being systematically skewed (high bias), or widely dispersed (high variance). In the case prediction may have high bias and variance, as in the bottom left target. Ideally predictions have both low bias and low variance, as shown in the top left. Reproduced from Formann-Roe, 2012.

By contrast, a restricted model may limit the number of parameters and be highly inflexible. In the extreme, the function $y = 1$ is just a horizontal line across the data. The values predicted do not vary at all, regardless of what the training set looks like. When predictions systematically deviate from the true predicted value they are biased, and the model has underfit. It is typically the case that as bias increases, variance decreases, and vice versa. Figure (2.2) shows the prediction of the same point after retraining on a new training set each time.

The challenge of choosing the best method for a task comes down to balancing this trade-off (Figure 2.3). Often the aim is for both bias and variance to be low. Predictions should not systematically deviate from the true value, but also should not be so heavily based on the

training set that they vary greatly when new data are introduced. Methods for controlling the flexibility of a model aid in balancing this trade-off.

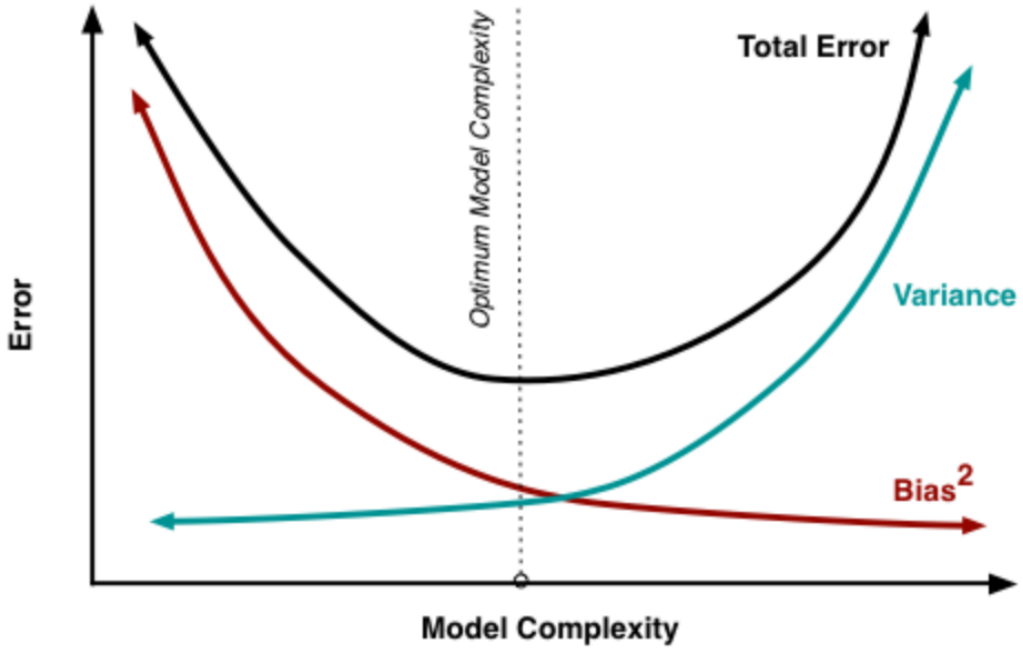


FIGURE 2.3: The balance between bias and variance. As model flexibility increases, bias decreases but variance increases. This is accompanied by progressively decreasing error on the training set; error in the test decreases initially but ultimately increases. Reproduced from Formann-Roe, 2012.

2.1.5 The accuracy-interpretability trade-off

There is often a trade-off between accuracy and interpretability. A logistic regression may perform reasonably well at predicting an outcome and provides odds ratios for predictors; it is easy to explain and some inference can be drawn using the sample. By contrast, support vector machines may provide an improvement on prediction, but it can be difficult to explain how a single predictor is associated with the outcome, particularly for non-linear kernels such as radial basis function. This can be viewed as a trade-off between model flexibility and interpretability (Figure 2.4). The use of complex supervised machine learning methods for a prediction problem implies that some sacrifice of interpretability is acceptable to the user.

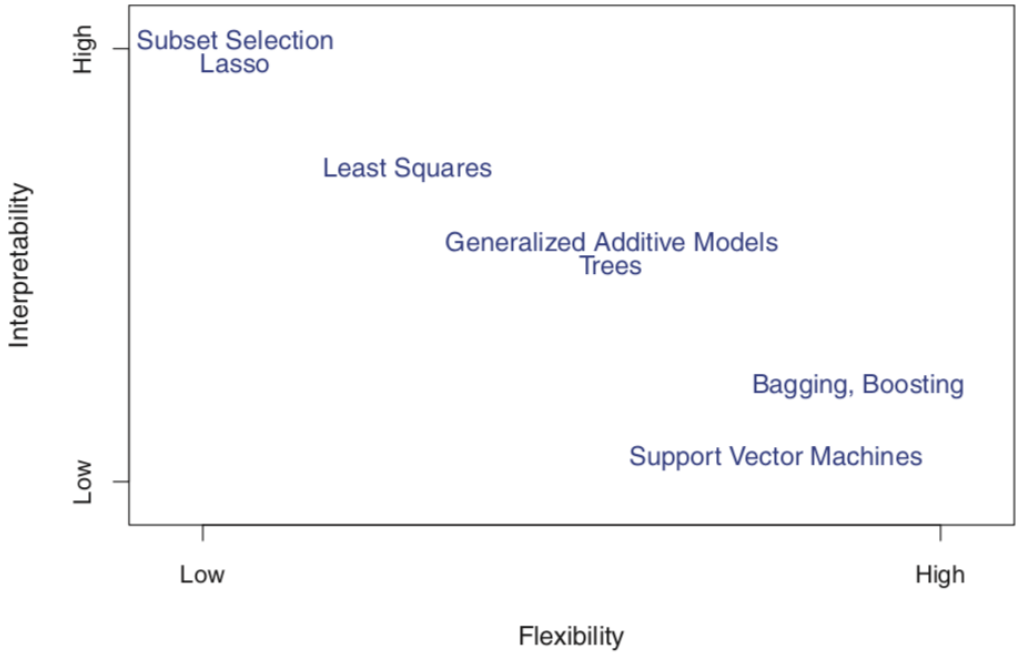


FIGURE 2.4: Flexibility-interpretability trade-off. Penalised linear models are perhaps the most interpretable, as they have fewer coefficients. They exhibit low flexibility. Ensembles and support vector machines can show extreme flexibility but also low interpretability. From James et al., 2013.

A result of this reasoning may be to suggest the use of an inherently explainable parametric approach such as logistic regression. However, this would require the fulfilment of all the ensuing assumptions, such as linearity of the predictor-response relationship and homoschedasticity of the errors, and these may not be met. Further, applying a parametric model, and following explanatory modelling procedures such as removing correlated predictors, may result in an unacceptable loss of prediction performance. The burgeoning field of explainable AI and sharp rise in tools for improving interpretability of machine learning methods has dampened the argument against complex models. Consideration of Occam's razor enables a more practical approach.

2.1.6 Occam's razor

Also known as the law of parsimony, Occam's razor suggests that the simplest answer is usually the correct one. A naive interpretation of this is that more complex models, which involve more parameters should be eschewed in favour of simpler models with fewer parameters. Such a heuristic is a useful one to keep in mind when building models on small datasets with noisy predictors, and it can be seen in common practices such as Breiman's rule for choosing hyperparameters, in which he suggests picking the most conservative value for a hyperparameter for which the prediction accuracy is within one standard deviation of the best-performing option (Breiman et al., 1984). However, the heuristic cannot be taken

as a rule, particularly given the success of ensembling and deep learning which can build large complex models that out-compete simpler ones.

The tendency for a more parsimonious model has been further broken down into two razors (Domingos, 1998):

- prefer the simplest model when two or more have the same test-set performance
- prefer the simplest model when two or more have the same training-set performance.

Domingos suggests that while the first of these is correct, the latter is often disproved. This more pragmatic approach, which seeks not to elect for parsimony above-all but instead prefer it when models are equally generalisable, is often employed and more in-keeping with the success of complex models in some fields.

2.1.7 The curse of dimensionality

High dimensional space influences both our capacity to reason about a problem and the ability of algorithms to learn from data. Observations which are close together in low-dimensional space become significantly further apart in high dimensions. As the predictor space grows with added dimensions, the training observations take up a greatly reduced portion of the space. This leads to an issue in machine learning known as the curse of dimensionality.

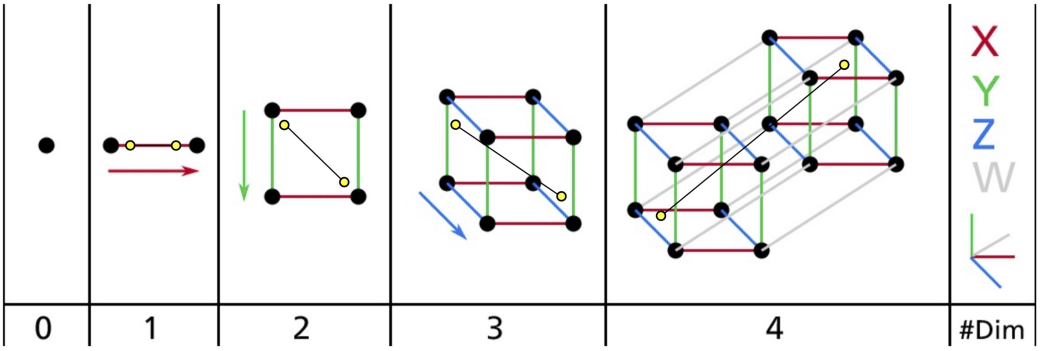


FIGURE 2.5: The curse of dimensionality. Progressing from a single point to a unit line, square, cube and then hypercube increases the area which observations can occupy, and so increases the average distance between randomly chosen points, illustrated by Euclidean distance (black lines) between random points (yellow circles). Nodes are shown as black circles. Red, green, blue and grey lines show the x, y, z, and w axes to represent 1, 2, 3 and 4-dimensions respectively. Adapted from Géron, 2019.

For training data with a set number of observations, increasing the number of predictors leads many learning approaches to overfit. As observations move further apart in higher dimensions, the space of possible decision boundaries for separating classes increases, and the selected option is more likely to be influenced by noise. As the training data are farther apart, so too are the testing data, resulting in poor generalisation. In algorithms such as k-nearest neighbours, there are no longer any observations that can be considered nearby.

In the simplest case, two randomly placed observations in a 2-dimensional unit square are, on average, closer together than two randomly placed points in a 3-dimensional unit cube (Figure 2.5). The difference is greater again between a square and a 4-dimensional unit hypercube, and only increases with the number of dimensions. Spreading the same number of observations uniformly across squares, cubes, and high dimensional hypercubes leads them to be increasingly located on the periphery (Domingos, 2012). More generally, our intuitions in 2 or 3-dimensional space for how a problem may be solved, or how a decision boundary may be formed, do not necessarily generalise to higher-dimensional space.

Models such as k-nearest neighbours can be much more affected by the curse of dimensionality, as it relies solely on distance between observations, and obtaining a larger sample size in an attempt to cover the predictor space is intractable. However, in a countering force labelled "the blessing of non-uniformity" (Domingos, 2012), most observations are not uniformly distributed across the predictor space, and methods do not spend time exhaustively searching it; they can often find a lower-dimensional representation of the problem by making small assumptions.

2.1.8 The bet-on-sparsity principle

A relationship may be complex and represented by many factors, some with very small effect sizes. Assuming population coefficients are drawn from a Gaussian distribution, their true values are all non-zero; this is called "dense". By contrast, if only a fraction of all predictors have non-zero population coefficients, the problem is termed "sparse". The bet-on-sparsity principle states that a sparse solution is preferred, as no method can effectively capture the relationships in a dense dataset (Hastie, Tibshirani, and Friedman, 2009). If a problem is dense, including all predictors in the model may seem appropriate. However, estimating small effects and generalising with them to a new dataset, under the constraint of how large a dataset can be collected for training, often means that a better or as-good model can be obtained by setting the weights for predictors with very small population coefficients to zero. In practice, the bet-on-sparsity principle means that it may be more sensible to choose an L_1 penalty over an L_2 penalty, discussed in section 2.3.2.2, in situations with many predictors of small effect sizes.

2.2 Design

This section gives an overview of the main considerations in designing a machine learning study. First, approaches for developing and validating a machine learning model are considered, including split-sample and cross-validation. Following this, methods for evaluating prediction performance in a classification setting are expanded on, principally discrimination and calibration. The section culminates in a discussion of techniques for comparing the performance classifiers.

2.2.1 Validation

Evaluating a model for predictive performance is referred to as model validation. The purpose of prediction models is to generalise to previously unseen data. As such, models are trained and evaluated on separate observations. This is important to ensure that unbiased estimates of predictive performance are obtained, as evaluating predictions in the same observations in which the model was trained will give an overly-optimistic estimate of prediction error. In classification problems, evaluation of predictive performance should include evaluation for both discrimination and calibration.

Model development involves the training of each model and any internal validation, where the latter includes all model evaluation done in any subset of the original sample. By contrast, external validation involves evaluating on a separate sample. Such a sample may be partly or fully external in the extent to which it diverges from the original training data. This may give a better estimate of the true error of the model. However, if the new sample is drawn from a region with contrasting exposures that may affect prevalence or predictors, it may not be representative of the target group, and so may give an overly-pessimistic view of the model.

In machine learning literature, a 'hold-out' partition from the training data may be used in training neural networks or gradient boosting, for example, and can be referred to as a validation set; to avoid confusion, this convention is not used here.

FIGURE 2.6: Internal validation. Data may be used for both training and testing (apparent validation), but is preferably split-up to keep training and testing observations independent.

2.2.1.1 Apparent validation

The weakest form of internal validation is apparent validation. This simply involves training and evaluating on the same observations. Discrimination in such a sample will be overly optimistic, and calibration on the sample in which the model was trained will always be perfect. See Figure 2.6. Discrimination may only be representative if the sample size is so large that sample coefficient estimates do no deviate from population parameters.

2.2.1.2 Split-sample validation

A simple method to constrain this optimism to split the data into two partitions, where one portion is used for training and another is used for testing. This can provide a more realistic estimate of predictive performance; however, the degree to which this works depends on sample size. As a single train-test split takes only one sample of the possible estimates of discrimination and calibration that can be obtained in internal validation, it may not be representative, and has been shown to be overly pessimistic on average (Steyerberg et al., 2001). As sample size grows, the validation error from split-sample validation approaches the true error rate. However, the point at which prediction performance in a split-sample approach becomes a good estimate is often unfeasible in medicine and means that the model is so unlikely to overfit that split-sample validation may no longer necessary with well-regularised models. Altering the size of the split is a compromise between gaining a sufficient sample size to learn accurate estimates for model coefficients and having a large enough test set for predictions to generalise well; choice is therefore controlled by the bias-variance trade-off. In practice, for models trained on tabular data in medicine, often the training set comprises between 60% and 80% of the sample. However, datasets with particularly large number of observations may use above 90%, such as when training large neural networks on imaging data with millions of observations.

2.2.1.3 k-fold cross-validation

Instead of a single split, an alternative is to use multiple train-test splits, where the training sets of each of these overlap (with the exception of 2-fold cross-validation). The model is refit in each training split and evaluated in its corresponding test split. This is an example of a resampling approach. It endeavours to gain a more realistic estimate of predictive performance by evaluating multiple correlated models and taking the average of their distribution as the estimate of model performance. In this sense, the average from resampling represents the entire dataset, whereas estimates from a split-sample approach only characterise performance on a fraction of the data.

k-fold cross-validation is an approach to resampling which maintains independence between each of the test sets (Figure 2.6). The dataset is divided into k chunks. These are used to build k training sets consisting of $k - 1$ chunks. Each training set is paired with the remaining chunk, which is the test set. Train and test splits within cross-validation are

referred to as the train fold and test fold. The predictions for each test fold are used to calculate performance metrics separately; the average and standard error across folds is then used to understand prediction. As models are not independent, it is not a true standard error, but has been shown to behave similarly and so remains useful.

The value chosen for k is typically 5 or 10. The choice is a trade-off between bias and variance, with leave-one-out cross-validation (LOOCV), where $k = n$, at one extreme, and 2-fold CV at the other. LOOCV is computationally expensive as many models are built, and give low bias in estimating prediction performance as each training fold includes $n - 1$ observations. By contrast, 2-fold CV has higher bias because it makes use of smaller chunks of the dataset, but lower variance because it averages a smaller number of less-correlated models (James et al., 2013). 5 or 10-fold CV are convenient choices which keep bias and variance reasonably low.

2.2.2 Tuning hyperparameters

The average across test folds from k-fold CV can be used for both model selection and model evaluation. Model selection is the practice of selecting between versions of the same model which have been trained with different hyperparameters. It navigates the trade-off between bias and variance by tuning model flexibility. Unlike parameters which control the weight given to predictors or nodes, hyperparameters are a meta-aspect of machine learning models which control the overall fit of the model.

Evaluation of different hyperparameter choices is most often done through cross-validation. As many different hyperparameters may need to be tuned for a single learning approach, each possible combination of the different hyperparameter values to be tried must be evaluated in a separate cross-validation. For example, given 3 important hyperparameters, trying 10 values for each and assessing using 5-fold CV means 30 different models will be evaluated, each performed 5 times, resulting in a total of 150 models being trained.

The exact function of hyperparameters varies between learning approaches. Though not all are equally useful, some are essential and their optimal choice ultimately controls how well the model generalises to new observations. The practice of choosing their value is therefore of great interest. However, the task has not typically been approached as a formal optimisation problem. Instead, a set of reasonably simple search methods have been most popular.

By far the simplest is a manual exploration of hyperparameters. This involves keeping plots of the breadth of search space covered and the prediction performance achieved for each combination of hyperparameters with the aim of ensuring that the region with minimum error or maximum discrimination has been well-covered. Such an approach may be used after systematic searches have returned an area of the hyperparameter space for further exploration. Alternatively, it can be used to update the learning rate in neural networks on-the-fly when a single large model is being trained and computation is too expensive to

afford repeated rounds of fitting. The model's performance is evaluated in both the training set and a hold-out sample, with the learning rate altered over time to achieve convergence.

A more systematic approach is grid search. As the name suggests, this involves enumerating a grid of all values of interest for each hyperparameter and then exhaustively searching over every combination. The combination which performs the best, as given by the average test-fold performance over cross-validation, is taken forward. Where only one or two hyperparameters need to be tuned, such as in linear SVMs, this is a reasonable approach and may not be computationally burdensome. It can also be particularly useful where an idea of the optimal region of the search space is already known, and a researcher wishes to thoroughly search for the global optimum. However, as previously stated, trying only 10 values for 3 hyperparameters in 5-fold cross-validation requires 150 models. This can quickly escalate; trying 20 values of 4 hyperparameters through 10-fold CV requires training 800 models. If there is a broad search space or more than two hyperparameters to tune, it is often beneficial to turn to alternatives.

Random search provides a possible solution to this by limiting the number of combinations to try up-front. Instead of enumerating all values, the number of combinations is chosen and a distribution (in Monte Carlo random search) or selection from which to draw values is defined. For each iteration, a value is drawn from the distribution for each hyperparameter. This is useful as adding more hyperparameters does not increase the number of combinations to search. Furthermore, evaluating all combinations by grid search leaves many searches with similar values. With random search, each hyperparameter varies every time, meaning a bigger proportion of the search space can be covered. It has been shown empirically to be useful for this, and can reach a similar accuracy to grid-search using a fraction of the computation (Bergstra and Bengio, 2012).

2.2.3 Nested cross-validation

When evaluating models and tuning hyperparameters, a common pitfall is that researchers use the same rounds of cross-validation for both model selection and model evaluation. The practice of running many cross-validations with different combinations of hyperparameters means that the prediction performance for the selected values will be optimistic, in a similar manner to how *p*-values from repeatedly tweaked analyses will be optimistic.

Nested cross-validation separates model selection and evaluation (Figure 2.7). An 'outer' loop of cross-validation evaluates the model. Within the training fold of this, an 'inner' loop of cross-validation is used for model selection, and is therefore run as many times as there are combinations to select from. Such an approach sufficiently constrains the optimism, so that it is a good approximation to the true error for a model (Varma and Simon, 2006; Vabalas et al., 2019). While there are extremely large datasets in which a split-sample approach may be sufficient, practically most datasets in medicine are limited in size by the cost of sample acquisition and so require nested cross-validation.

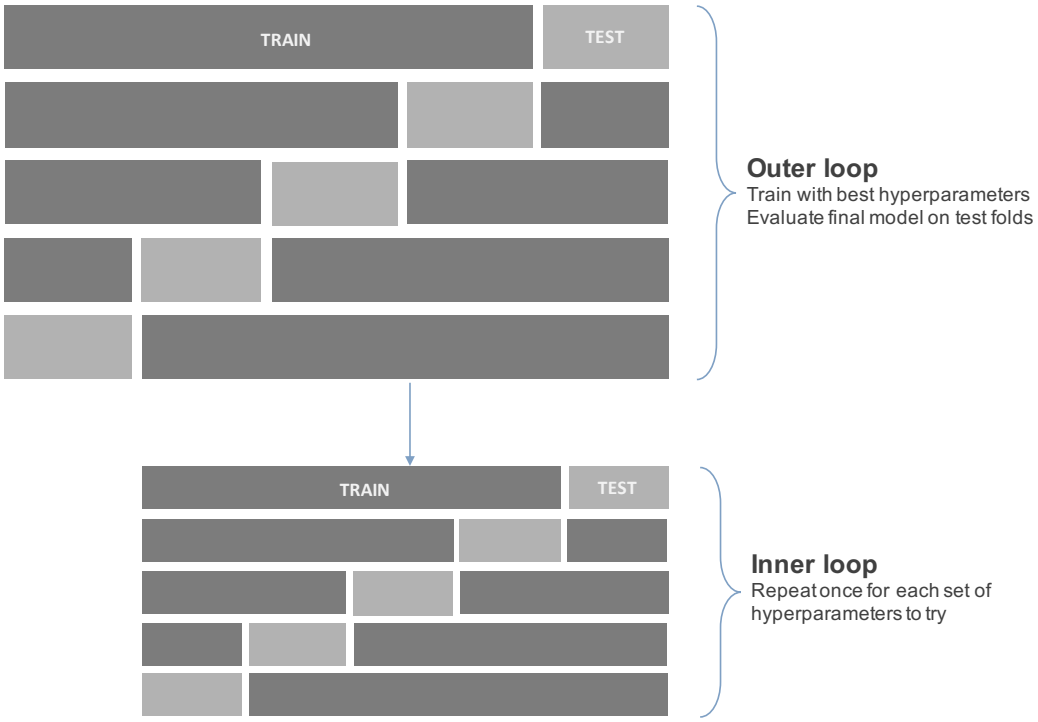


FIGURE 2.7: Nested cross-validation. An outer loop of k-fold cross-validation is used to perform model evaluation. Within each training fold of this there is an inner cross-validation, which is performed once for every combination of hyperparameters. The best model is refit to the outer loop's training fold and evaluated on the outer loop's test fold.

2.2.4 Predictor selection

Predictor selection, also called feature selection, is the practice of reducing the number of candidate predictors. Often the preferred method of doing this is to use prior knowledge, particularly if the current dataset is small. However, the use of increasingly 'wide' datasets with limited observations means that further reducing the number of models in a data-driven manner can help reduce the likelihood of overfitting, though it does not reduce events per variable (EPV), as this includes candidate predictors. The three types of predictor selection are embedded, filter and wrapper.

Embedded predictor selection uses the modelling procedure itself to drive-down the number of predictors. This is most commonly implemented using the L_1 penalty, which may be applied in support vector machines, gradient boosting and neural networks but is perhaps most commonly associated with least absolute shrinkage and selection operator (LASSO) regression. In addition to this, tree-based methods are able to reduce the number of predictors. When partitioning the data, decision trees select a predictor at each 'split' in a tree, which may mean that one or more predictors are never incorporated in the trees if they are always out-done by other variables.

Filter-based methods employ an up-front selection process which reduces the number of predictors before any modelling of the data occurs. For genetic prediction models using common variants, the most common method is to use univariable association tests on each

predictor with the outcome and filter by the strength of the evidence for association, known a p-value thresholding. Use of principal component analysis (PCA) to obtain a set of orthogonal predictors before modelling is sometimes also considered to be a filter-based selection as it is applied before modelling, though it projects to a subspace rather than selecting predictors in the input space. Alternatively, correlation between predictors or measures for association or information may be used to filter predictors.

Finally, wrapper based methods take a more algorithmic approach to predictor selection by encapsulating a modelling approach with a series of steps for dropping predictors. A popular technique in gene expression studies is to use recursive feature elimination to iteratively drop predictors (Guyon et al., 2002). Alternative methods, are to use forward or backward selection, or a mixture of the two. Backwards selection should always be preferred over forward selection, but is not possible when $p > n$ (Harrell Jr, 2015).

2.2.5 Leakage and pipelines

To successfully evaluate different modelling approaches which involve steps such as predictor selection and transformation, pipelines should be used. These exist to combat the effect of information leakage, one of the most common sources of inflation of estimates of prediction performance. This refers to the accidental incorporation of information from the test set into the predictors or model, as information has 'leaked' from the test set to the training set. For predictor transformation, anything which uses summary measures such as mean and standard deviations, are at risk. Z-transformation of predictors is a common source of this.

A pipeline requires a sequence of steps, such as z-transformation, filter-based predictor selection and machine learning model, to be defined beforehand. This pipeline is then applied in each fold of cross-validation, and is set up so that measures learned in the training fold, such as the mean and standard deviation of a predictor or its association with the outcome, are saved and then applied in the test set. The pipeline is therefore used in every fold of cross-validation, and no up-front transformation or selection of predictors should be done before cross-validation unless they avoid leakage of information from the test set. The erroneous practice of applying predictor selection outside of cross-validation is so frequent in parts of the bioinformatics community that it is highlighted in introductory machine learning textbooks (for example, James et al., 2013).

2.2.6 Model evaluation

All methods of model validation require evaluation of predictive performance. To do this requires quantifying model predictions. For automatic model selection, such as in hyperparameter tuning, this also inclines us toward scalar summaries of prediction models. The choice of method depends on modelling context, and what type of output is produced by the model.

2.2.6.1 Classification metrics

When a class is predicted, classification metrics can be calculated from a confusion matrix (Figure 2.8). If a model outputs predicted probabilities, a threshold must first be applied in order to discretise the output; a value of 0.5 is typically used by default. Classification metrics are therefore discontinuous scoring rules.

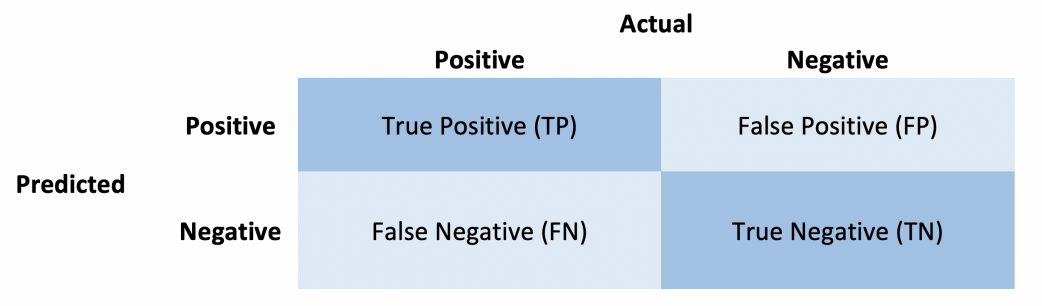


FIGURE 2.8: The confusion matrix. The true class labels are cross-tabulated with the predicted class labels to get true positive, false positive, false negative and true negative counts. From these, all classification metrics follow.

The confusion matrix can be used to derive classification metrics (Table 2.1). In medicine, concern is often placed on sensitivity and specificity, which measure the ability to detect true positives and to reject true negatives, respectively. There is a trade-off between these two measures. A greater sensitivity means catching more true positives, but the ability to detect additional labels often comes with inadvertent selection of more false positives too. For a predicted probability, moving the threshold at which class labels are assigned to be lower may increase sensitivity, but will also decrease specificity. In information retrieval and much of machine learning, the focus lies instead on the relationship between precision (positive predictive value; PPV) and recall (sensitivity). These are often summarised using the F1-score (Table 2.1) which, unlike relying on sensitivity and specificity, takes no account of the number of true negatives.

Measures	Synonyms	Calculation	Definition
Accuracy		$\frac{TP+TN}{TP+TN+FP+FN}$	The proportion of correct classified results. Equivalent to $1 -$ error.
Sensitivity	Recall, hit rate, true positive rate (TPR)	$\frac{TP}{TP+FN}$	The proportion of positive outcomes that were correctly classified.
Specificity	True negative rate (TNR)	$\frac{TN}{TN+FP}$	The proportion of negative outcomes that were correctly classified.
Positive predictive value (PPV)	Precision	$\frac{TP}{TP+FP}$	The proportion of positive predictions that are truly positive.

Measures	Synonyms	Calculation	Definition
Negative predictive value (NPV)		$\frac{TN}{TN+FN}$	The proportion of negative predictions that are truly negative.
F1-score		$\frac{2TP}{2TP+FP+FN}$	The harmonic mean of PPV and sensitivity.

TABLE 2.1: Classification metrics. Measures are derived from the confusion matrix, which gives true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

While the terms classification and prediction are used interchangeably here, the distinction is important to be aware of. For instance, support vector machines are classification methods, as they work to assign classes to observations purely via linear algebra, while logistic regression makes use of probability theory. In situations where only assignation of classes is required, and the threshold for choosing these can be decided beforehand, attacking the classification problem directly may be optimal. However, in biological and medical sciences, often an estimate of the likelihood that an observation belongs to a class is desired. The optimal choice of threshold may need to be amended based on some clinical need, or the degree of risk communicated to a patient.

2.2.6.2 Discrimination

The use of discontinuous scoring rules, such as classification metrics, is not recommended (Moons et al., 2015). Though choice of a discrete class can be necessary, classification metrics are often not useful in model development as assignment of a threshold is premature, and throws away useful information. It should ideally be delayed till clinical need demands it and directs its location. Measures of discrimination are continuous scoring rules. These provide a measure of how well the distributions of predicted probabilities of belonging to the positive class discriminate between the true classes. By far the most common measure of this is the area under the curve (AUC) for the receiver operator characteristic (ROC), also known as the c-statistic.

First pioneered in World War II for the correct detection of radar signal by a receiver operator, ROC curves (Figure 2.9) were taken-up in machine learning as early as 1989, and were subsequently embraced a decade later to meet the deficiencies of discontinuous measures (Fawcett, 2006). The AUC is now one of the most frequently used measures in both clinical prediction modelling and machine learning for binary classification. By contrast, in genetic epidemiology, studies of polygenic risk scores typically adopt an explanatory modelling paradigm, reporting measures of goodness-of-fit, such as Nagelkerke's pseudo-R^2, or variance explained on the liability scale (Lee et al., 2012b). However, AUC is also reported for large studies of polygenic risk scores, for instance (Ripke et al., 2014), has been investigated for the maximum AUC achievable given the heritability and prevalence of an outcome

(Wray et al., 2010), and related to other measures including reclassification statistics and risk of disease for different percentiles (So and Sham, 2010).

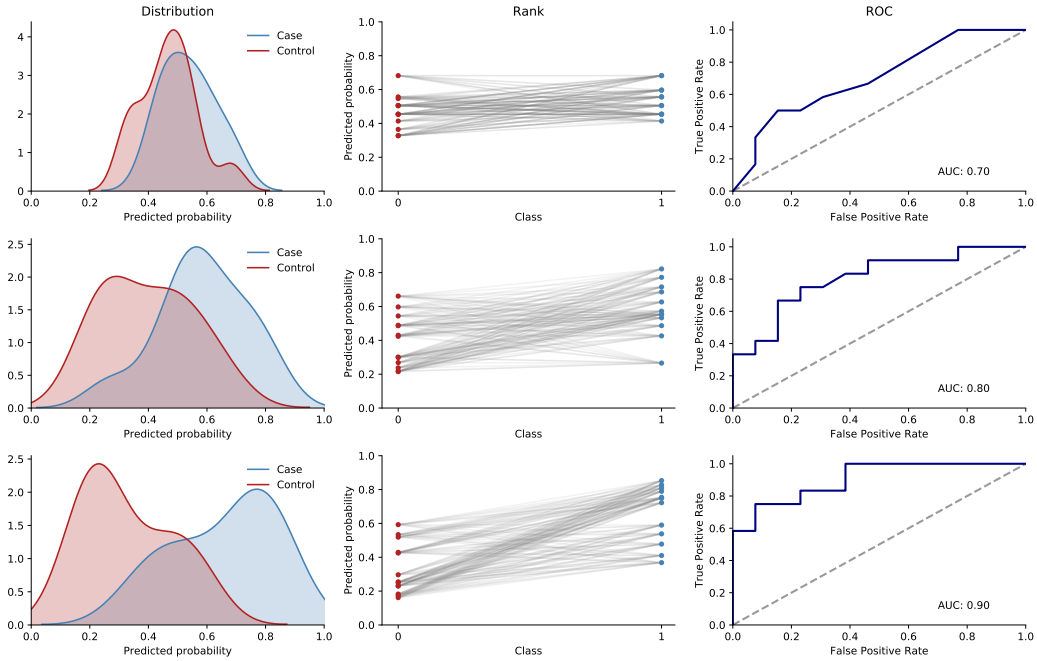


FIGURE 2.9: Assessment of discrimination via the ROC curve. Rows show predictions with increasing AUC from top to bottom. The predicted probability of class membership coloured by actual class (left), can be transformed to the ROC plot (right), which follows the diagonal at pure chance and reaches the upper left corner with perfect prediction. AUC can be calculated from the area under the ROC curve. The central column shows the predictions for all case-control pairs connected by a line. AUC is equivalent to the proportion of lines with a positive slope.

In the decision-threshold interpretation of the AUC, a plot is generated by calculating sensitivity (true positive rate) and 1-specificity (false positive rate) for each threshold for the predicted probability from a model. This presentation of the ROC plot is simply a transformation of predicted risk distributions (Janssens and Martens, 2020). The ROC plot itself may be used to choose an optimal threshold for a given task. More frequently, a single number is desired to summarise the ROC curve, which is done using the AUC (Bradley, 1997).

An alternative and more intuitive probabilistic interpretation of the AUC is achieved by taking all possible case-control pairs of observations for which a prediction has been given. Counting all the times a case was correctly assigned a higher probability than a control and dividing this sum by the number of pairs creates the AUC. For this definition, the AUC may be referred to as the concordance index. This has the same value as calculating the area under the ROC curve and aids in its interpretation. It then follows logically from this that a value of 0.5 is pure chance, 1 is perfectly correct and 0 is perfectly incorrect. Further, it also follows that the AUC is insensitive to distributional assumptions and class imbalance.

The distributional insensitivity of the AUC is also evident in the last of its "three-way equivalences" as a rank-based measure (Hanley and McNeil, 1982). The AUC is simultaneously the area under the ROC curve, the probability of correctly ranking a case higher in any randomly chosen case-control pair, and is equivalent to the non-parametric Wilcoxon/Mann-Whitney U statistic.

Criticisms levied at the AUC often revolve around insensitivity. These typically assert that the AUC changes little when discrimination is high, or that changes in the actual predicted probabilities are not reflected by a change in AUC. However, these two features follow naturally from the descriptions of the AUC given above. For a high AUC, distributions of predicted risk for cases and controls must be well-separated, and so further improvements in the model only affect the ranking of the small number of observations in the tails of the distributions, and hence contribute little to the total calculation of AUC. Similarly, the AUC is unaffected by the actual value assigned by a model, as long as the pairwise rankings of cases and controls remain the same. Consequently, a model may be clinically useful in moving a subgroup of interest into the highest risk group, or many participants could shift above or below an important decision threshold, but a concomitant increase or decrease in AUC may not be observed. The latter may easily occur where a threshold is placed toward the centre of a risk distribution. Alternative measures such as reclassification metrics may be employed to address these concerns.

The AUC has also been criticised as a scalar summary of a two-dimensional problem (Drummond and Holte, 2006), a critique which extends to all single-value summaries of prediction performance. Others have noted that it is unlikely the entire range of thresholds are potentially useful, and that AUC may be misleading where ROC curves for two classifiers cross each other (Adams and Hand, 1999).

2.2.6.3 Calibration

Discrimination is complemented by calibration (Steyerberg et al., 2010; Van Calster et al., 2019), which can be assessed where a method provides predicted probabilities of class membership. A model has achieved calibration when the observed risk of having an outcome is the same as the predicted risk assigned by the model. For instance, in a model of having or not having schizophrenia, taking all individuals assigned a 0.6 probability of having the disorder, 60% should have the true label of schizophrenia. The extent to which calibration has been achieved can be broken down into mean, weak, moderate and strong (Steyerberg et al., 2019).

Mean calibration compares the average prediction with the average outcome. It can be summarised by an odds ratio, by taking odds of the mean $\hat{y}$ (model predictions) over odds of mean y. However, mean calibration is uninformative for internal validation measures, including apparent validation and cross-validation, as it is perfect in the dataset from which

a model was developed, but is informative for external validation. The assessment of average calibration is also referred to as calibration-in-the-large.

Weak calibration is often evaluated by a calibration slope (Cox, 1958). For a logistic regression, this is assessed by fitting a model on the dataset using the linear predictor (LP), which is the linear combination of coefficients and predictors in the test set. This takes the form $\text{logit}(y) = a + b \cdot \text{LP}$, where the calibration slope b shows the direction of miscalibration, and the intercept a the overall miscalibration. A b greater than 1 indicates over-estimation in individuals with lower risk and underestimation in individuals with highest risk. b less than 1 indicates the opposite of this effect, while a b of 1 is perfect (Huang et al., 2020). During internal validation, b less than 1 therefore requires contracting predictions toward 0.5, which can be achieved through recalibration or through applying penalisation of coefficients in model development. Alternatively, shrinkage can be applied directly on predictions using the calibration slope. During external calibration, b is affected by both overfitting and differing predictor-response relationships in the development and validation sets (Steyerberg et al., 2019). Though a b of 1 is a perfect slope, predictions may still be underestimated overall if a is greater than 0, or overestimated if a is less than 0. Ideal predictions therefore have b of 1 and a of 0.

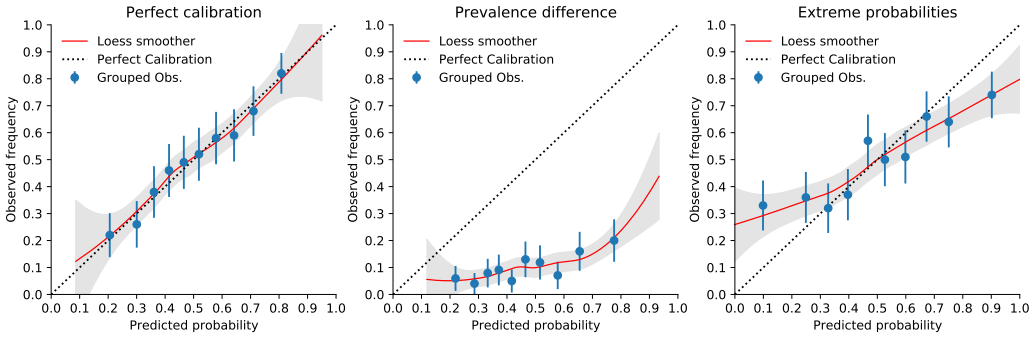


FIGURE 2.10: Graphical assessment of calibration. A loess smoother with 95% confidence interval is used to show the general relationship between predicted and observed outcomes. Observation are also grouped by decile to give a visual alternative to the Hosmer-Lemeshow goodness-of-fit test. Plots show perfect calibration (left); consistently over-predicted risk due to training in a sample with a higher prevalence than the population used for prediction (centre), as may happen if training on a case-control sample before predicting in the general population; and predictions which are too extreme (right), as coefficients have been overestimated in the training sample, leading to unlikely events assigned a probability too low and likely events assigned one too high.

Moderate calibration is best assessed graphically (Steyerberg et al., 2019) (Figure 2.10). This takes the predicted probabilities from a model on the x axis and the observed proportion of individuals in which the outcome occurred on the y axis (Austin and Steyerberg, 2014). A diagonal line here defines perfect calibration. A loess smoother is used to aggregate-together observations with similar probabilities assigned by the model to allow plotting the observed proportion on the y-axis. Following this, confidence intervals can be used, and typically the mean observed proportions and predicted probabilities in groups of observations is also

annotated, often binned into quintiles or deciles. Graphical assessment of calibration in this manner is particularly useful as it shows both mean calibration and weak calibration; it is a visual representation of the Hosmer-Lemeshow goodness-of-fit test. Interpretation of calibration plots can be influenced by the number of groupings chosen, the flexibility of the loess smoother, prevalence of the outcome and discrimination of the model. Finally, strong calibration extends this to visualise calibration for all categorical predictors. For each level of a category, such as sex, separate calibration plots are generated. These can aid in visualising poor model specification, such as omitted interaction or polynomial terms.

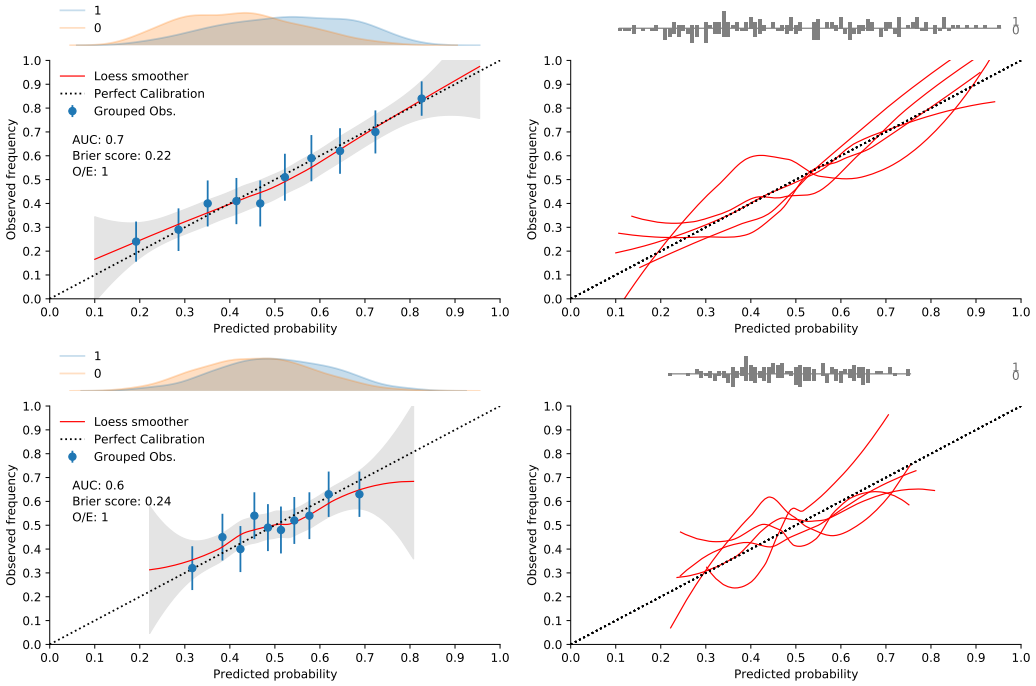


FIGURE 2.11: Visualising discrimination and calibration by validation plots. Calibration is combination with a plot of the distribution of predictions for cases and controls. Examples are given for traditional kernel density estimation (KDE) plots (left) and the mirrored histogram-style plots (right) preferred in Steyerberg et al., 2019. Calibration is also separated into a full graphical calibration, including grouped observations and confidence intervals (left), or only the loess curve, which can be used to overlay results from multiple models or rounds of cross-validation; the latter is shown here. Rows, showing 0.7 AUC on the top and 0.6 AUC on the bottom, demonstrate how calibration and summary measures can vary depending on the strength of discrimination. AUC, Brier's score and mean calibration (as observed/expected; O/E) may also be annotated (left).

2.2.6.4 General prediction measures

Overall performance measures seek to measure the difference between predictions and the true outcome (Steyerberg et al., 2019). Measures of these residuals involve elements of both discrimination and calibration.

The most frequently used are Nagelkerke's R^2 and Brier's score. R^2 gives the proportion of variation in the outcome explained by the model. It is often estimated in assessing predictions of a continuous outcome by machine learning models, in the form

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}, \tag{2.1}$$

where y_i is the continuous outcome and $\hat{y}_i$ is the predicted value for the ith observation. As R^2 cannot be computed for a binary outcome, alternative pseudo-R^2 can be calculated. The Cox & Snell R^2 is

$$R^2_{CoxSnell} = 1 - \left(\frac{\ell_0}{\ell_M}\right)^{\frac{2}{N}}, \tag{2.2}$$

where ℓ_0 is the likelihood (described in section 2.3.1) of the null model, ℓ_M is the likelihood of the fitted model and N is the number of observations. Nagelkerke's R^2, a scaling of the Cox & Snell R^2, is computed by

$$R^2_{Nagelkerke} = \frac{R^2_{CoxSnell}}{\max(R^2_{CoxSnell})}, \tag{2.3}$$

where $\max(R^2_{CoxSnell}) = 1 - \ell_0^{\frac{2}{N}}$. Though $R^2_{Nagelkerke}$ spans the same interval as R^2, it does not have the same intuitive interpretation. Brier's score summarises the squared difference between observed outcomes and predictions,

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{2.4}$$

where y_i and $\hat{y}_i$ are the observed and predicted values of the outcome, and N is the sample size. A lower Brier score means closer agreement between the outcome and predictions. However, unlike Nagelkerke's R^2, Brier score is not standardised to any range. It is consequently more difficult to interpret. In addition, its value is dependent on the proportion of cases in the sample. It is therefore mainly useful for relative comparison between classifiers. Despite these caveats, it is an attractive measure for prediction models because it is a proper scoring rule, as it is optimised when the true probabilities equal the predicted probabilities. By contrast, optimising for accuracy may simply lead to predicting the most frequent outcome.

As noted in section 2.2.6.3, graphical calibration is influenced by discrimination. To be explicit, discrimination can be incorporated directly to create a "validation plot" which provides a general assessment of predictive performance (Figure 2.11). Furthermore, scalar summaries of discrimination and calibration can be annotated to provide a more detailed picture for model evaluation.

2.2.6.5 Comparing classifiers

Where two or more competing models have been developed, quantification of the difference between them is desirable. A range of measures are used where classifier output is binary, notably McNemar's chi-squared test and a binomial exact test for pairwise comparisons, which may be used in addition to a combined F-test for comparing multiple classifiers (Raschka, 2018). Where model output is predicted probabilities, discrimination for a single split-sample approach can be evaluated in the test set using DeLong's test for two correlated ROC curves (DeLong, DeLong, and Clarke-Pearson, 1988). However, cross-validation produces a set of models, for which the average AUC is reported. Furthermore, if $k > 2$ then models are correlated, as their training sets overlap by at least one fold, making AUCs from k-fold CV non-independent. For 10-fold CV all models share 80% of their training data, and so taking all test-fold predictions together for a single statistical test, as if they had been produced by a single model, may cause the variance to be underestimated. How then should comparison of classifiers proceed?

Two general solutions are available to employ statistical tests for cross-validation results. The first is to apply a modified validation procedure. The most well-used of these is a 5x2-fold cross-validation (5x2 CV), wherein each 2-fold split is independent (Dietterich, 1998). 5x2 CV was originally developed for use in classification tasks, using proportion misclassified (error rate) as the metric. Models are trained and evaluated on each of the 2 folds, which is repeated 5 times; mean and variance of difference in proportions from the 2 folds are calculated for each of the 5 repeats, which is used to create a t-statistic. Though well-used in machine learning in general (Aggarwal, 2015), 5x2 CV is less commonly applied in biomedical fields.

The second is to use an existing test with the best empirical results. The two most prominent options are a parametric approach, the paired t-test, and the non-parametric equivalent, the Wilcoxon signed-rank test. Comparisons by Dietterich, 1998, again using error rate, showed a slightly elevated type I error for the paired t-test performed after k-fold cross-validation, and a highly elevated type I error for the "resampled paired t-test", which is performed by taking repeated random samples from the dataset. Given increased power for parametric tests, the k-fold cross-validated t-test is only recommended where type II error takes precedence over type I (Dietterich, 1998). Furthermore, distributional assumptions of parametric approaches may not be correct, and are difficult to assess when k is low.

As a result, the Wilcoxon signed-rank test is recommended for pairwise comparison of machine learning models (Demšar, 2006). This ranks the difference in AUCs by their absolute value and calculates a test statistic, W, from the product of the ranks and the sign of the differences. However, it has lower power to detect differences than the paired t-test when assumptions for the t-test have been met, so may be at greater risk of type II error.

To avoid inflation of Type I error through multiple comparisons, multiple classifiers may be first compared using the repeated measures analysis of variance (ANOVA) or the Friedman

test depending on whether a parametric or non-parametric approach is indicated. Following this, pairwise tests between classifiers may be performed. The family-wise error rate (FWER) of pairwise comparisons must also be controlled; however, Bonferroni correction is too severe for many comparisons (Aggarwal, 2015), particularly as all models produced by k-fold cross-validation are correlated, with correlation between any two folds increasing with k. Many alternatives are used in the machine learning literature to control FWER. Alternatively, controlling the false discovery rate (FDR) has been suggested to be less suitable in evaluating machine learning approaches as it requires pre-specifying the FDR when there may be no clear guidance for this (Demšar, 2006). However, it is frequently used in statistical genetics and biomedical applications, and consequently more-easily interpreted, and a threshold of 0.05 is commonly accepted as sufficiently stringent.

2.3 Learners

An enormous variety of supervised machine learning methods exist. The comparison of several approaches on a dataset is necessary and commonplace. As the true function which maps inputs to outcome is unknown, and each classifier makes different small assumptions, there is rarely a strong theoretical basis to prefer a single method over others when using tabular data. The framework of evaluating multiple approaches is also part of the more empirical approach taken in machine learning; the model which maximises predictive performance in a dataset should be chosen. This section introduces the methods used in the thesis: penalised regression, support vector machines, random forests, gradient boosting and neural networks. These are the most widely taught approaches in machine learning, for example (James et al., 2013), and as chapter 3 will make clear, these are also the most commonly used methods in psychiatric genetics. Their description is best understood by first reviewing logistic regression.

The terms "learner" and "classifier" have slightly different uses in the literature, but are here used interchangeably to refer to a supervised machine learning algorithm or its implementation for a classification problem, regardless of whether its output is a class or probability of class membership. For simplicity, learners are considered only for a binary outcome.

2.3.1 Logistic regression

Logistic regression models the conditional probability of class membership given the training data, $P(Y = 1|X)$. This contrasts with ordinary least squares (OLS) linear regression, which models the outcome directly. A multivariable linear regression contains more than one predictor and can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{2.5}$$

for p predictors. Applying this to a classification problem would be simple and even intuitive, as the average outcome, Y, is the proportion of cases in the sample. However, linear regression makes no assumptions on the limits of Y, which can only be 0 or 1 in binary classification. Logistic regression accounts for this using the logistic function. For simplicity, substituting $P(X)$ for the conditional probability of Y given X, $P(Y = 1|X = x)$, the logistic function is

$$P(X) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p}}{1 + e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p}}. \tag{2.6}$$

The output from the logistic function is sigmoidal and always within the 0-1 range. Equation 2.6 can be rewritten using the logit transformation,

$$log\left(\frac{P(X)}{1 - P(X)}\right), \tag{2.7}$$

to give

$$log\left(\frac{P(X)}{1 - P(X)}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_pX_p, \tag{2.8}$$

where the left hand side gives the log odds (logit) of the outcome, and the right hand side is known as the linear predictor.

Coefficients are obtained using maximum likelihood, which produces estimates that maximise the probability of the observed data. Taking $P(X)$ from equation 2.6, and assuming independent observations, the likelihood function is

$$\ell(\beta) = \prod_{i=1}^{N} P(X)_i^{y_i}(1 - P(X)_i)^{(1-y_i)}, \tag{2.9}$$

where $\prod$ is the product operator and N is the number of observations. In practice, the log is used for ease of computation. The log-likelihood,

$$\ell(\beta) = \sum_{i=1}^{N}\{y_i log(P(X)_i) + (1 - y_i)log(1 - P(X)_i)\}, \tag{2.10}$$

can be maximised using iterative weighted least squares via the Newton-Raphson algorithm. The resulting coefficients, $\hat{\beta}$, are the maximum likelihood estimates.

Unlike in linear regression, it cannot be concluded that a coefficient is the average change in Y per 1 unit change in X, all other effects held constant. For logistic regression, a coefficient is interpreted as the change in log odds per one unit change in X. While the relationship between log odds and the linear predictor (equation 2.8) is linear, the relationship

between $P(X)$ and the logistic function (equation 2.6) is not. As a result, the coefficient expressed in terms of $P(X)$ will vary depending on the value of X. Furthermore, unlike the normally distributed errors in linear regression, errors in logistic regression follow the binomial distribution.

Predictions on new data are made with equation 2.6 using the estimated coefficients from the training set and the predictors from the testing set.

As a broad definition of supervised machine learning includes all methods that learn from data, linear and logistic regression may be considered by some to be a machine learning approach, especially if used in prediction modelling context when using typical ML designs such as cross-validation. In this and subsequent chapters it is regarded as a contrasting method, however, due to its frequent use in statistical analysis.

2.3.2 Penalised regression

2.3.2.1 Overfitting and optimism

Linear and logistic regression models obtain maximum likelihood estimates of model coefficients. This procedure seeks an unbiased estimate in the data on which the model is fit. A possible consequence of this is that overfitting occurs. Particularly under small sample sizes, effect sizes derived in the training set are likely to overestimate true effect sizes, due to regression-to-the-mean upon generalisation to a new dataset. Penalisation addresses this issue by introducing a penalty or regularisation term to the loss function, forcing effect sizes to shrink toward zero. As sample size for the training data increases, coefficients estimated by maximum likelihood become more accurate and are less likely to require penalisation to ensure generalisation. This section introduces norms and two of the natural extensions to logistic regression they enable.

2.3.2.2 Norms

Penalisation, or regularisation of model coefficients can be applied to logistic regression models. These methods shrink coefficients to be closer zero than the maximum likelihood estimates. Such shrinkage can improve generalisation by increasing bias and decreasing variance.

The most common regularisation terms in machine learning are the L_1 and L_2 penalties. In linear algebra, the L_1 and L_2 penalties are norms, a method for measuring the size of a vector. Norms also can also be seen as giving a measure of the distance from the origin to the point assigned by the vector. The L_1 (Manhattan) norm takes the absolute values of elements in the vector, and is given by

$$||x||_1 = \sum_{j=1}^{p} |x_j|, \tag{2.11}$$

for each element x_j in a vector $x \in \mathbb{R}^p$.It is often used when discrimination between zero and near-zero is essential, as its use of absolute values means its rate of change is unaffected by the position of the vector (Goodfellow et al., 2016). By contrast the L_2 (Euclidean) norm takes the square of each element using

$$||x||_2 = \left(\sum_{j=1}^{p} x_j^2 \right)^{\frac{1}{2}} = \sqrt{x^T x}, \tag{2.12}$$

for a vector x. In contrast to L_1, the L_2 norm changes more slowly when values approach zero (Goodfellow et al., 2016). The Euclidean norm is often simply denoted $||x||$. The squared L_2 norm, $||x||_2^2$, is typically used for computational simplicity and is henceforth referred to as just the L_2 norm. Both L_1 and L_2 are cases of the generic L_q norm,

$$||x||_q = \left(\sum_{j=1}^{p} x_j^q \right)^{\frac{1}{q}} \tag{2.13}$$

for $q \in \mathbb{R}$ and $q \geq 1$. q is set to 1 and 2 for the L_1 and L_2 norms respectively. Norms are applied across machine learning, including in support vector machines, gradient boosting, neural networks and regression models, by applying them to a vector of coefficients or weights from a model. The inclusion of the L_1 or L_2 penalties in the loss function for a model therefore includes a measure of the size of the vector of weights, such that a larger size decreases the measured fit to the data. This encourages a smaller size in order to achieve a better fit, consequently forcing coefficients or weights to be closer to zero.

2.3.2.3 Ridge

Adding either of these regularisation terms creates two natural extensions to linear regression, or logistic regression, that help to control overfitting by restricting the parameters of the model: least absolute shrinkage and selection operator (LASSO) regression (Tibshirani, 1996) and ridge regression (Hoerl and Kennard, 1970).

A penalty term can be applied by subtracting it from the log-likelihood. The inclusion of the square of the L_2 norm with the log-likelihood gives

$$\ell(\beta) = \sum_{i=1}^{N} \{ y_i log(P(X)_i) + (1 - y_i) log(1 - P(X)_i) \} - \lambda \sum_{j=1}^{p} \beta_j^2, \tag{2.14}$$

where β is a p-length vector of estimated coefficients, excluding the intercept. This produces a ridge-penalised logistic regression which shrinks all coefficients but the intercept. This introduces a new tuning parameter, λ, which takes values of 0 or above and is chosen through cross-validation. If $\lambda = 0$, the L_2 norm is nullified. As the hyperparameter increases toward

infinity, the ridge penalty forces coefficients closer to zero than the maximum likelihood estimates; however, none will actually reach 0 until $\lambda = \infty$, at which point $||\beta||_2^2 = 0$. It therefore shrinks coefficients but does not perform predictor selection.

2.3.2.4 LASSO

Swapping the L_2 for the L_1 norm produces a LASSO-penalised logistic regression,

$$\ell(\beta) = \sum_{i=1}^{N} \{y_i log(P(X)_i) + (1 - y_i) log(1 - P(X)_i)\} - \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.15}$$

where $\beta \in \mathbb{R}^p$ is a vector of coefficient estimates, excluding the intercept. The LASSO was first proposed by Tibshirani, 1996. It addresses the principle issue with ridge regression, that regardless of the number of predictors or the size of λ (below ∞), all predictors will be included in the resulting model. By contrast, the LASSO penalty may restrict some coefficients to be zero, and so includes predictor selection in modelling. It is therefore a sparse model and an example of embedded predictor selection. As with ridge, penalisation increases with λ, with coefficients equal to the maximum likelihood estimates when $\lambda = 0$, and the values of coefficients decreasing toward zero as λ increases. A higher λ further reduces the number of coefficients with non-zero values.

Optimisation for ridge and LASSO in equations 2.14 and 2.15 can be rewritten as maximisation of log-likelihood, subject to the constraint that the L_1 or L_2 norm must be less than some value, s, where each value of s has a corresponding value for λ. This reformatting allows us to visualise the optimisation in 2-dimensional space when considering only 2 predictors. Figure 2.12 demonstrates a situation where s, and hence λ, takes on values that forces the maximum likelihood estimates for the two coefficients to not overlap with the constraints from L_1 or L_2 penalties. The values for $\hat{\beta}_1$ and $\hat{\beta}_2$ for the 2 predictors are constrained to the point where the contour plot for $\hat{\beta}$ meets the blue diamond or sphere. The shape of the constraints makes it more likely that interaction between the contour and the diamond (LASSO) is on an axis, therefore setting the value of β_1 to 0, while the sphere (ridge) meets the contours at a point away from the axis (James et al., 2013). $\hat{\beta}$ here gives the least-squares estimate from linear regression when $p = 2$. However, this intuition extends to logistic regression applied to higher-dimensional problems. A third extension, the elastic net, is a generalisation of the LASSO and ridge which combines both penalties, but is not considered here.

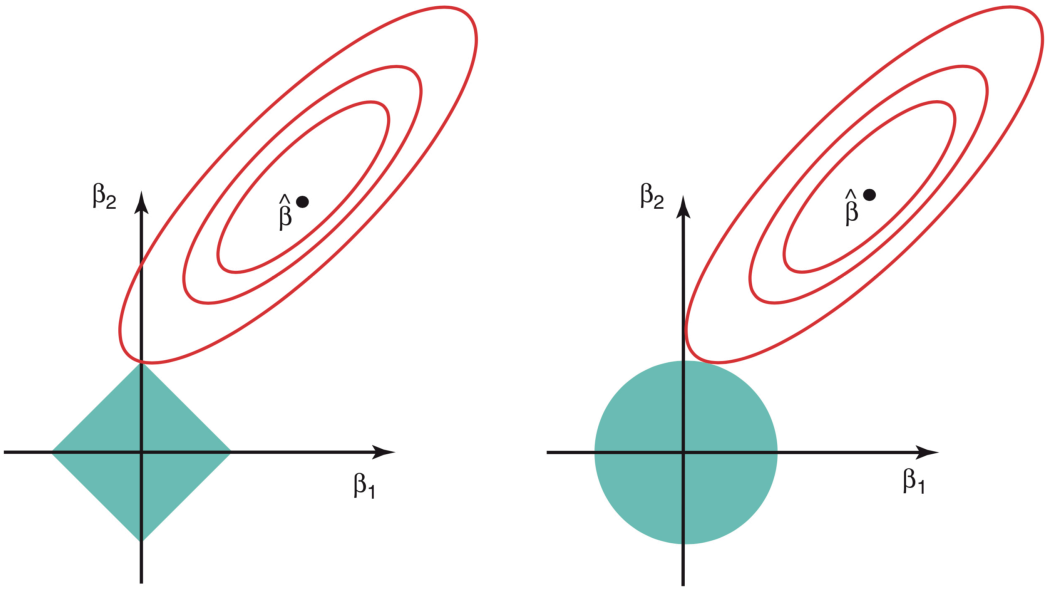


FIGURE 2.12: Optimisation for ordinary least squares linear regression with ridge or LASSO penalties on a 2-dimensional problem. The blue diamond and sphere show the values that β_1 and β_2 are constrained to take under the L_1 and L_2 penalties, respectively, for a given s. The red area surrounding $\hat{\beta}$ illustrates a contour plot, often used for illustrating an optimisation minimum, where points in a single ring take the same value. The region where the contour meets the green region provides the values of β_1 and β_2 which minimise the penalised least squares, while the point denoted $\hat{\beta}$ gives the least squares estimates. Adapted from (James et al., 2013).

2.3.2.5 Implementation

All machine learning and statistical approaches are implemented in the popular open source Python package, scikit-learn. LASSO and ridge-penalised logistic regression is parameterised differently in scikit-learn than listed given here, and in other popular packages, such as glmnet. Logistic regression is implemented with multiple options for solvers in scikit-learn. Choosing 'newton-cg' with no penalty produces an unpenalised logistic regression in the same manner as that from the R statistical programming language. Use of both L_1 and L_2 penalties with logistic regression is only possible with the 'liblinear' solver, which performs well in high dimensions. Here shrinkage is parameterised by C, equivalent to $\frac{1}{\lambda}$, which provides stronger regularisation with smaller values; optimisation is performed by coordinate descent.

2.3.3 Support vector machines

The support vector machine (SVM) is a classifier in the true sense, in that it directly considers class assignment, rather than attributing probabilities of class membership. It first emerged in the 1990s (Cortes and Vapnik, 1995) and received widespread use in the biological sciences for micro-array classification in the early 2000s. The technique forgoes probability to produce a geometrically-derived classification. The method works only on numerical data, where each predictor is a dimension in space: a dataset with p predictors

will have p dimensions. A hyperplane is fit to the training data with the aim of separating the two classes. The hyperplane that best separates the two classes in the training data is then used to predict classes in the test data.

The SVM is best understood by reviewing its foundations: the linearly separating hyperplane, the max-margin classifier, and the support vector classifier.

2.3.3.1 Separating hyperplanes

If trained on a dataset with 2 predictors, the hyperplane is simply a 1-dimensional line, and with 3 predictors it becomes a 2-dimensional plane. In higher feature space it becomes more difficult to visualise, but the same principle applies. A hyperplane in p-dimensions is given as

$$b + w_1 X_1 + w_2 X_2 + \cdots + w_p X_p = 0, \tag{2.16}$$

where b is the bias term, w is a weight vector of length p and $X_1, \ldots, X_p$ are the predictors. Each observation $x_i \in \mathbb{R}^p$, $\forall i = 1, \ldots, N$ is assigned to a class, $y_i \in \{-1, 1\}$, depending on which side of the hyperplane it falls. Class assignment can therefore be denoted by

$$h(x) = \text{sign}(b + x^T w), \tag{2.17}$$

where values which deviate further from 0 indicate greater confidence that an observation belongs to the class. Assuming the outcome is linearly separable, and a hyperplane can be found which separates the two classes, an issue is encountered: how to choose the best linearly separating hyperplane when an infinite number exist?

2.3.3.2 Max-margin classification

The maximal margin classifier is a precursor of the SVM that applies the same principle of fitting a hyperplane to the data, but can only be used for linearly-separable classes. For any given hyperplane, the margin is the area between it and the surrounding data points, calculated as perpendicular distance from the hyperplane to the point. The selection of the best hyperplane for separating the data is done by maximising the size of the margin, so producing the largest possible gap either side of the hyperplane.

To achieve max-margin classification, we seek to maximise the margin of width $2M$, where $M = \frac{1}{||w||}$, $||w||$ is the Euclidean norm of the weights $w_1, \ldots, w_p$, and the weight vector w is orthogonal to the hyperplane (Figure 2.13). The optimisation problem can be denoted as

$$\begin{aligned} &\min_{w,b} ||w|| \\ &\text{subject to } y_i(b + x_i^T w) \geq 1, \forall i, \end{aligned} \tag{2.18}$$

where $y_i(b + x_i^T w) \geq 1$ ensures observations lie on the correct side of the hyperplane. This formulation produces a convex optimisation problem that can be solved with quadratic programming. The resulting hyperplane is called the maximal margin hyperplane, or optimal separating hyperplane. Significantly, only the points which touch the margin, called *support vectors*, contribute to its placement.

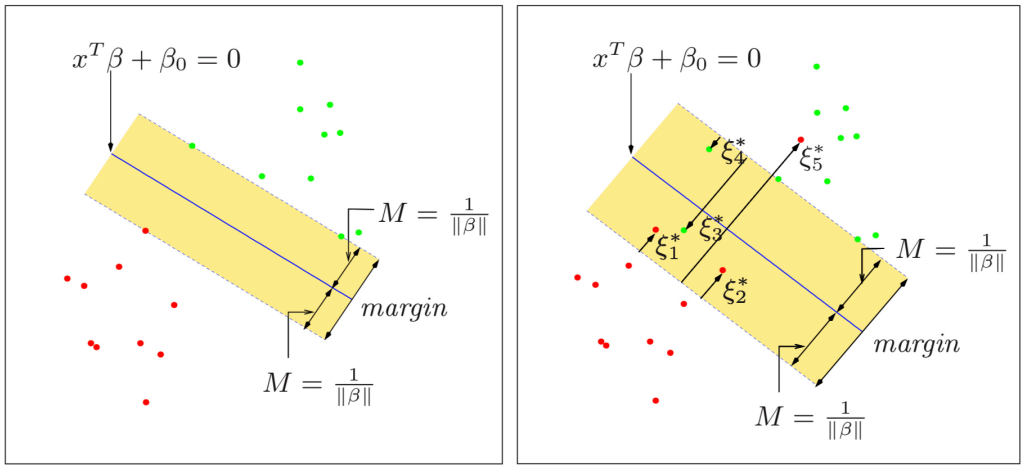


FIGURE 2.13: The margin for a max-margin classifier (left) and a support vector classifier. The latter allows for margin violations, shown here as $\xi_1, \ldots, \xi_5$. The hyperplane, shown in blue is the equation of the line in p-dimensional space. The margin, shown in yellow, has width $2M$ where $M = \frac{1}{||w||}$. Adapted from Hastie, Tibshirani, and Friedman, 2009.

Use of this optimal hyperplane produces the maximal margin classifier, which has the agreeable property that the hyperplane has the greatest possible distance between itself and observations in the training data, and so is likely to generalise better to the testing data than alternative separating hyperplanes.

2.3.3.3 Support vector classification

Classes are rarely linearly-separable in biomedical applications; for any hyperplane that can be fit to the data there will be points that lie on the 'wrong' side. The solution to this is to use a 'soft' margin, where vectors which violate the boundary can be tolerated to some level. The implementation of this for non-separable cases is called the *support vector classifier* or *soft margin classifier*. To account for these violations, errors are defined ($\xi = \xi_1, \ldots, \xi_N$) for all N support vectors, also known as *slack variables*.

Allowing for margin violations, the optimisation of the support vector classifier is taken as

$$\begin{aligned} &\min_{w,b} ||w|| \\ &\text{subject to} \\ &y_i(b + x_i^T w) \geq 1 - \xi_i, \forall i \\ &\xi_i \geq 0, \sum_{i=1}^{n} \xi_i \leq C. \end{aligned} \tag{2.19}$$

We can see from this that the hyperparameter that denotes the cost, C, provides an upper bound for the sum of the violations to the margin and hyperplane which are deemed acceptable.

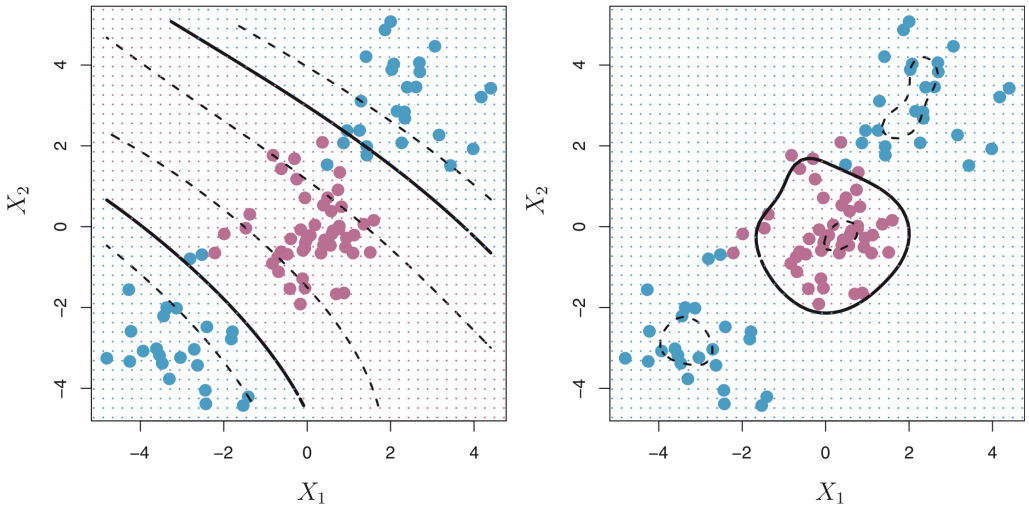


FIGURE 2.14: Application of a 3-degree polynomial kernel (left) and radial kernel (right) on a 2-dimensional problem when a linear boundary would not be sufficient. Adapted from James et al., 2013.

2.3.3.4 The support vector machine

The SVM uses a maximal margin hyperplane with a soft margin, making it similar to the support vector classifier, but able to handle non-linear boundaries. It does this using a powerful and widely-used technique in machine learning: kernel functions.

An intuitive solution to solve a non-linear problem is to transform predictors to a higher-dimensional space and use a linear classifier. The decision boundary will be linear in the high-dimensional space in which it was learned, but non-linear in the original space. The brilliance of kernel functions is that they introduce a method for performing such an operation without needing to transform predictors. This is made possible by first noting that the SVM optimisation problem can be represented in its dual formulation using

$$\sum_{i=1}^{N} \alpha_i y_i \langle x_i, x_{i'} \rangle + b, \tag{2.20}$$

for any two observations x_i and $x_{i\prime}$, where α is a vector taking non-zero values only for support vectors and $\langle x_i, x_{i'} \rangle$ denotes the inner product. It can be solved using only the inner products of all pairs of observations in the training set (Hastie, Tibshirani, and Friedman, 2009). Similarly, classifying new observations requires only the inner product between the new observation and all training observations. The support vector machine replaces $\langle x_i, x_{i'} \rangle$ with a kernel function, K, which calculates the inner product in the transformed space. This allows for a non-linear decision boundary in the predictor space (Figure 2.14) and makes finding the optimal hyperplane computationally efficient even in high (or infinite) dimensions. A kernel function takes the form

$$K(x_i, x_{i'}) = \langle g(x), g(x') \rangle. \tag{2.21}$$

By far the most popular kernels are the linear, polynomial and radial:

$$\text{Linear kernel: } K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j} = \langle x, x' \rangle, \tag{2.22}$$

$$\text{Polynomial kernel: } K(x_i, x_{i'}) = (1 + \langle x, x' \rangle)^d, \tag{2.23}$$

$$\text{Radial kernel: } K(x_i, x_{i'}) = exp(-\gamma ||x - x'||^2), \tag{2.24}$$

where d, the degree of the polynomial, and γ, which controls localisation of the decision boundary, are hyperparameters for the polynomial and radial kernels respectively. Only the linear and radial kernels are considered here due to their overwhelming popularity and contrasting behaviour. SVMs making use of the radial kernel are referred to as radial basis function (RBF) SVMs.

2.3.3.5 Loss function

Previous sections have considered the SVM from a max-margin perspective. It can also be viewed from a penalisation perspective. This requires reformulating the optimisation using the hinge loss

$$\ell_{hinge}(b, w; y_i, x_i) = max(0, 1 - y_i(w^T x_i + b)), \tag{2.25}$$

to give the L_2-regularised L_1 hinge loss,

$$\min_{w,b} \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{N} \ell_{hinge}(b, w; y_i, x_i). \tag{2.26}$$

It takes the typical form of a loss function with a penalty term, and is equivalent to equation 2.19 (Aggarwal, 2015). Equation 2.26 illustrates that the standard formulation of the SVM implements L_2 penalisation of the weights. Squaring equation 2.25 gives the alternative L_2-penalised L_2 hinge loss. Further discussion of loss functions is given in section 2.3.6.1.

2.3.3.6 Hyperparameter tuning

Tuning the linear SVM involves altering C, which controls how soft the margin of the hyperplane is. C is a positive tuning parameter usually set to 1 by default. At $C = 0$, no margin violations are tolerated, and only a linearly-separable problem can be solved. Increasing C creates a wider margin that is more frequently violated by support vectors. This means the position of the hyperplane depends on more individuals, giving it higher bias and lower variance (as it depends on so many individuals, a small change in the training set is less likely to cause a large shift in the hyperplane, and consequently any predicted classes). Lowering C forms a smaller margin that is less tolerant of violations; it depends on far fewer instances in the training data, so has higher variance and lower bias.

The RBF SVM uses the γ hyperparameter, in addition to C. γ controls how localised and hence non-linear the hyperplane is. The default value is typically $\frac{1}{N}$, where N is the number of samples. A higher value for γ means the hyperplane becomes more linear and depends on more distant support vectors to form its shape, while a lower value means only points much closer are likely to have a strong effect, making it more non-linear and localised. In practice, both C and γ are used to control model regularisation in the bias-variance trade-off. A low C and γ can cause higher variance and increased bias, while increasing them both may improve generalisability of the model up to a point.

2.3.3.7 Implementation

SVMs are implemented in scikit-learn, which makes use of the liblinear (Fan et al., 2008) and libSVM (Chang and Lin, 2011) packages. The output of SVMs is a class assignment or distance from the hyperplane, which may be converted to a probability using Platt scaling (Platt et al., 1999); output was kept as distance from the hyperplane by setting 'probability' to 'False' unless indicated otherwise.

The SVM is often touted as one of the best "out of the box" classifiers, and is particularly useful for high-dimensional data such as that found in genomics. However, the lack of interpretable weights in non-linear kernels often earns it the label of a "black box". This may partly explain why it has often found competition in a class of methods that are easier to intuitively grasp and can also provide a helpful in-built measure of predictor importance: trees.

2.3.4 Decision trees

While logistic regression and support vector machines take probabilistic and geometric approaches to prediction respectively, trees are a fundamentally algorithmic operation with an intuitive geometric interpretation. Decision trees alone can be a useful predictor, but they are perhaps most widely known for their use in two contrasting ensemble approaches. This section covers the foundation of decision trees before progressing to random forests and gradient boosting.

Decision trees have been in use for regression and classification problems since the 1960s and 70s respectively as the automatic interaction detection (AID) and THAID programs produced by Morgan, Sonquist and Messenger (Breiman et al., 1984). However, modern classification trees owe much of their origin to Jerome Friedman and Leo Breiman. Together they independently re-implemented trees, and later jointly introduced classification and regression trees (CARTs). These are now the most common form of decision tree and can be used for both classification and regression. Many alternatives to CARTs have been considered. The most prominent of these are iterative dichotomiser 3 (ID3) (Quinlan, 1986) and C4.5 (Quinlan, 2014), both developed by Ross Quinlain. ID3 made use of entropy for considering splits and could be applied to classification using categorical predictors only, while C4.5 grew out of ID3, using the gain ratio for node splitting, extending handling of data types and introducing tree pruning. Only CARTs are expanded on here, due to their frequent use and incorporation into ensembles.

Decision trees use an iterative, greedy procedure to partition the predictor space into regions. In CARTs, each partition, or split, separates the observations into two groups, and continues recursively, each time further separating the newly formed groups. This process is known as recursive binary splitting, and each of these groups or regions are called nodes or leaves. A node can also be labelled as the 'root' node if it is the first node to be considered at the top of the tree, 'internal' if it is further split in two, or 'terminal' if it is not, while 'branches' connect the nodes together. For each predictor X_j, every possible cut-point s is considered for splitting. The reduction in error in the resulting splits is evaluated for each cut-point. From all predictors and cut-points, the split that maximises node purity is chosen for partitioning the data. The procedure is repeated for each of the resulting regions. This continues until a stopping criterion is reached.

2.3.4.1 Splitting criteria

The choice of splitting criterion is one possible hyperparameter in decision trees (Figure 2.15). For classification, the logical choice might be simply the classification error (E),

$$E = 1 - \max_k(\hat{p}_{mk}), \tag{2.27}$$

where $\hat{p}_{mk}$ refers to the proportion of observations in the mth region of the kth class. However, this has proven too insensitive to be useful (James et al., 2013). Instead, the Gini index (G) and entropy measures (D) are commonly used. The Gini index is given as

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{2.28}$$

and approaches zero when the majority of observations in a region are the same class. It therefore measures how pure a node is, taking on a lower value as the node becomes more pure, and is minimised to give the estimated best split. The most common alternative, entropy, is simply a generalisation of Shannon's entropy for more than two classes, also known as the cross-entropy or deviance, and is given by

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}. \tag{2.29}$$

As with the Gini index, entropy gets closer to zero as as nodes become more pure, and so the cut-point with the smallest entropy is used to split the node.

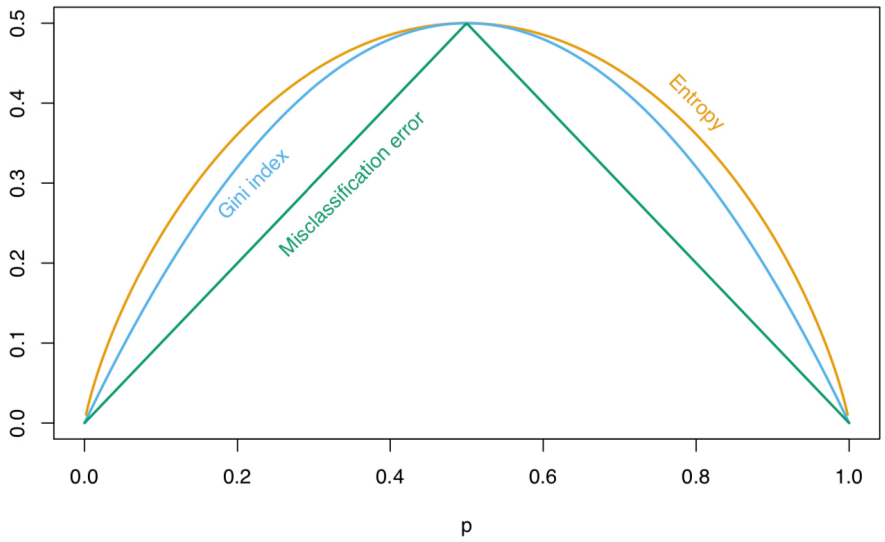


FIGURE 2.15: Splitting criteria in CARTs. Curves show the values that classification error, Gini index and rescaled entropy take for different p, the proportion of observation in a region taking the desired class. Adapted from (Hastie, Tibshirani, and Friedman, 2009).

As values increase as purity decreases, criteria are labelled as measures of node impurity. For all three measures, the splitting criterion is actually calculated on the potential new daughter nodes or regions, with the sum of the impurity function for both potential nodes minimised; i.e. for the candidate daughter regions

$$R_1(j,s) = X|X_j \leq s \quad \text{and} \quad R_2(j,s) = X|X_j > s, \tag{2.30}$$

for the j^{th} predictor and cut-point s, the sum of each region is minimised using

$$\min_{j,s}[c(R_1(j,s)) + c(R_2(j,s))], \tag{2.31}$$

where c is the classification error, Gini index or entropy evaluated for all observations in the mth region $x_i \in R_m(j,s)$. Alternatively, the difference between the impurity of the parent node and the sum of the impurity of the daughter nodes can be maximised. Many alternative splitting criteria have been suggested, such as the Gain ratio and normalised information distance; however, only Gini index and entropy have enjoyed widespread adoption in CARTs (Aggarwal, 2015).

2.3.4.2 Predictions from decision trees

The final collection of all splitting rules can be visualised in a tree-like structure which shows the recursive nature of the splits. At prediction-time, each new observation is traced down the tree, following the splitting rules, until a terminal node is reached. Predictions from terminal nodes are made by taking a majority vote or average of the outcome for the training-set observations.

Trees are an attractive option because of their high level of interpretability, speed of computation, robustness to distributions due to rank-based splitting, ability to capture interactions and the option of being used in both classification and regression. However, trees struggle to capture additive effects (Hastie, Tibshirani, and Friedman, 2009), an aspect of some concern for traditional quantitative genetic models of polygenic traits. They also show bias toward predictors with high cardinality when choosing splits, though this can be addressed through the use of conditional inference trees and forests (Hothorn, Hornik, and Zeileis, 2006; Strobl et al., 2007). Furthermore, they are liable to overfit, showing low bias and high variance, and consequently poor prediction, a feature Breiman referred to as "instability" (Breiman, 1996). Modifications have been developed to control such overfitting, mainly by limiting tree depth or pruning-back fully grown trees. The former requires making tree-depth a hyperparameter that is tuned by cross-validation to ensure the tree is large enough to capture the structure of the data without being so large as to capture noise as well. Controlling tree depth through pruning similarly involves tuning a hyperparameter, here by indexing all sub-trees of the fully grown tree and choosing the best index through cross-validation. This relies on the intuition that restricting tree depth up-front partially ignores the greedy nature of trees, and allowing them to first grow fully enables them to possibly find more informative splits further down the tree. Despite these efforts, trees still show worse prediction than modern machine learning methods. Rather than optimising the structure of a single tree, it

is far more common for them to be combined together in an ensemble which improves over the prediction of any individual tree. Such an approach is the basis for random forests.

2.3.5 Random forests

The high variance of fully-grown decision trees can be addressed by bagging (Algorithm 1), a simple but powerful technique. Bagging takes repeated samples of a dataset, with replacement, builds a high-variance model on each subsample and averages the predictions from all models (Breiman, 1996). The resulting ensemble has greatly reduced variance compared to any of the individual models, typically decision trees, and is the foundation on which random forests are built. The key insight Breiman made was that variance can be reduced even further if the correlation between individual trees is also reduced, and that this can be done by random subsampling of both observations and predictors.

Algorithm 1: Bagging, adapted from (Aggarwal, 2015)

input : Training data $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N} (x_i \in \mathbb{R}^p, y_i \in \mathcal{Y})$

output: An ensemble classifier g

for $t \leftarrow 1$ *to* T *trees* **do**

  Construct a sample data set $\mathcal{D}_t$ by randomly sampling with replacement in $\mathcal{D}$

  Learn base classifier h_t based on $\mathcal{D}_t$

end

return $g(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{t=1}^{T} I(h_t(x) = y)$

While bagging is a generic method that can be applied successfully to any classifier with low bias and high variance, random forests take advantage of the specific structure of decision trees. As with bagging, the data for building each tree is bootstrapped from the original dataset. In random forests, however, at each split the predictors are also subsampled, serving to further decorrelate trees (Algorithm 2). The subsampling fraction is another tune-able hyperparameter for random forests. However, for classification it is typical to take a random subsample of m predictors from all p which are available, where $m = \sqrt{p}$ is normally used (Breiman, 2001). Setting m to be lower than this further decorrelates trees, while setting it to the maximum of 1, taking all predictors at each split, is equivalent to bagging.

The use of bootstrapping in random forests has the side-effect of creating a set of observations which were not used in training for each tree. These are termed out-of-bag (OOB) observations and can be used to provide an estimate of the test error for hyperparameter tuning or to monitor the value of including additional trees. On average, the bootstrapping procedure will allocate around two thirds to use in training a single tree, leaving the remaining third available. To obtain predictions for an observation used in training, we simply make use of all trees in which it is not used (i.e. where the observation is OOB). This is around $T/3$ predictions, where T is the number of trees built, from which a majority vote can be taken, or average probability in classification. The OOB error approaches the LOOCV error as T increases. This makes it a useful option for particularly large datasets, but for smaller

ones on which many classifiers are compared, it is often foregone in favour of using 5 or 10-fold cross-validation because the computational cost is not too great, and classifiers can be compared together under a common validation framework.

Algorithm 2: Random forest, adapted from (Aggarwal, 2015)

input : Training data $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N} (x_i \in \mathbb{R}^p, y_i \in \mathcal{Y})$

output: An ensemble classifier g

for $t \leftarrow 1$ *to* T *trees* **do**

Construct a sample data set $\mathcal{D}_t$ by randomly sampling with replacement in $\mathcal{D}$

Learn a decision tree h_t by applying LearnDecisionTree($\mathcal{D}_t$, iteration = 0, ParentNode = root):

If stop criterion is satisfied, return

Randomly subsample predictors in the whole predictor space $\mathbb{R}^p$ to get a new data set $\hat{\mathcal{D}}_{current}$ = RandomSubset($\mathcal{D}_{current}$)

Find the best feature q^* according to impurity gain

Split data $(\mathcal{D}_L, \mathcal{D}_R) = split(\mathcal{D}_{current}, q^*)$

Label the new parent node $v = parent.newchild(q^*)$

Conduct

$LearnDecisionTree(\mathcal{D}_L, iteration = iteration + 1, ParentNode = v)$

and

$LearnDecisionTree(\mathcal{D}_R, iteration = iteration + 1, ParentNode = v)$

end

return $g(x) = \underset{y \in \mathcal{Y}}{\text{argmax}} \sum_{t=1}^{T} I(h_t(x) = y)$

After training trees separately, they are taken together to provide a prediction. To predict the outcome for a new observation, it is passed down each tree making up the forest until it reaches the terminal node. Breiman originally proposed taking a majority vote of each decision tree in the ensemble (Breiman, 2001). However, scikit-learn averages probabilities (the proportion of training observations belonging to the positive class) across terminal nodes. In addition, Breiman suggested trees should be fully-grown in a random forest, but subsequent analysis suggests restricting tree depth is likely to improve prediction (Segal and Xiao, 2011).

Random forests have remained a popular method in bioinformatics and machine learning more generally. This is because they persist many of the appealing properties of decision trees, such as importance scores and invariance to scale or transformation, but their implementation can be easily parallelised and their predictive ability is competitive with other top-performing approaches. They are, however, less interpretable than decision trees. In addition, they are not the only approach to building ensembles of decision trees for improved prediction. While random forests mostly average over fully-grown trees, an alternative approach which sequentially combines many shallow trees has become increasingly popular in recent years following computational improvements and the growth of online machine

learning competitions.

2.3.6 Gradient boosting machines

Random forests were initially developed in response to boosting (Cutler, Cutler, and Stevens, 2012), an alternative ensembling methodology. Boosting creates an ensemble, or 'committee', through the combination of 'weak' base learners, where accuracy has a high probability of being slightly greater than chance, to create a single 'strong' learner that has accuracy competitive with other modern approaches. Unlike the parallel algorithms of bagging and random forests, boosting is done sequentially. In a regression context, each new base learner is fit to the residuals of the current ensemble, each time making small improvements to the learned function. In this sense it follows a general principle in machine learning that learning slowly via smaller steps gives improved prediction over a single, fast model such as a fully-grown decision tree. Boosting is a general method that can be applied to any base learner; SVMs and logistic regression have been proposed as appropriate for classification, but shallow decision trees are the most commonly used. Boosting is introduced through AdaBoost, followed by a description of gradient boosting and modern implementations.

2.3.6.1 AdaBoost

Boosting first became popular after appearing in a classification context as AdaBoost (Freund and Schapire, 1997), following separate earlier work by both authors (Schapire, 1990; Freund, 1995), and inspired by research from Kearns and Valiant on improving the performance of a weak learner. As with other boosting methods, AdaBoost sequentially adds weak learners then takes a weighted majority vote to predict from the ensemble. During learning, weights are applied to the observations. These are initially equal, but get updated after iteration, also known as a step or boosting round, where a new tree is added. After each step, weights are modified to focus increasingly on the misclassified observations.

AdaBoost was the first method to clearly demonstrate the ability to 'boost' an arbitrary learner. It was later formalised as a forward stagewise additive model with an exponential loss function that approximates an additive basis expansion (Friedman, Hastie, Tibshirani, et al., 2000). Here, the additive basis expansion is a weighted summation over the weak learners,

$$g(x) = \sum_{m=1}^{M} \beta_m b(x; \gamma_m), \tag{2.32}$$

with β as the vector of coefficients for expansion and $b(x; \gamma_m)$ a basis function on x with parameters γ. Instead of fitting equation 2.32 through maximum likelihood, for example, which can be computationally expensive, boosting adds basis functions sequentially, each time choosing a new weight, without updating previous weights in β. In this sense it is additive, as it is a summation over the basis functions (weak learners), and forward in

that each new weight and basis function to be added are solved alone, without ever moving 'backwards' to modify previous weights in response. It also stagewise, as rather than learning a single function g to approximate f, one is composed from parts learned in stages.

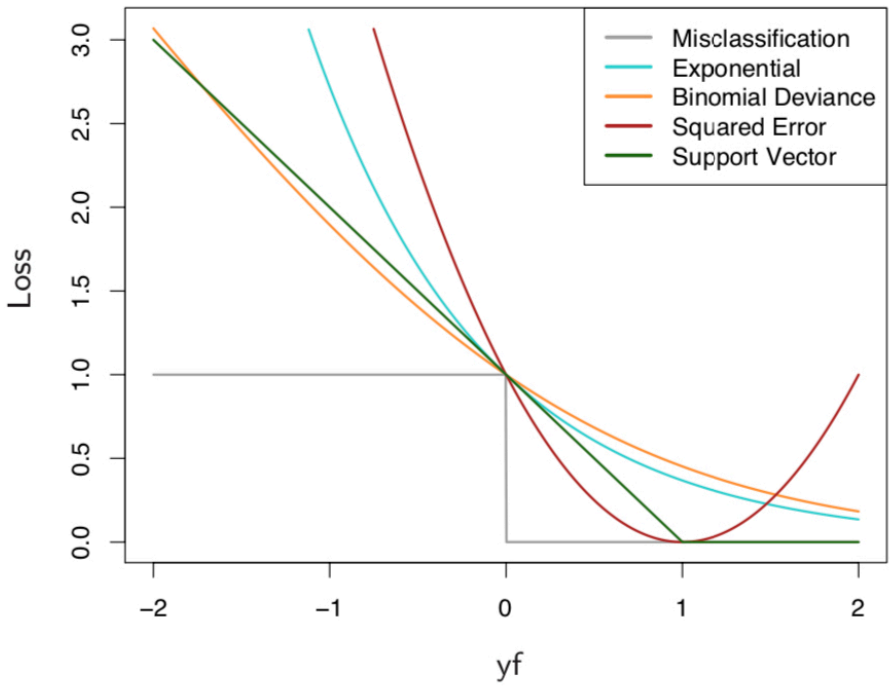


FIGURE 2.16: Loss functions scaled for comparison where the outcome $y \in \{-1, 1\}$. The x axis, the product of y and $f(x)$, is the "margin", so that if $yf(x) > 0$ then classification is correct. The squared error is the least squares estimate, binomial deviance is the negative log likelihood, support vector is the hinge loss, and exponential loss is that used by AdaBoost. Misclassification loss applies only a unit penalty for incorrect classifications. Adapted from (Hastie, Tibshirani, and Friedman, 2009).

It has been demonstrated that AdaBoost is equivalent to equation 2.32 learned through an exponential loss function (Friedman, Hastie, Tibshirani, et al., 2000). The use of an exponential loss is computationally useful but may be considered sub-optimal when used with noisy data that have weak predictors or imperfect class labels. The negative log-likelihood (binary cross-entropy or deviance) or hinge loss (used in SVMs) are similar in that they place a linearly-increasing penalty on observations which are misclassified (Figure 2.16), and to a lesser extent a decreasing one on those correctly classified (with the Hinge loss not considering correctly labelled observations). However, exponential loss applies a significantly greater penalty on the most misclassified observations, while hinge loss and binomial deviance rely on a wider spread of the instances. This makes hinge and binomial deviance more robust when the observations most difficult to classify may be impossible to improve upon using available predictors (Hastie, Tibshirani, and Friedman, 2009), or may have mislabelled classes; these properties suggest AdaBoost may be less suited to prediction from common variants in psychiatry. The squared error is also inappropriate for classification in general, as it places increasing penalty on correctly classified observations. Ideally, loss

functions for classification should approximate the misclassification error (shown in grey for Figure 2.16); the hinge and binomial deviance loss functions are continuous monotonic examples of this.

Algorithm 3: GBMs, adapted from (Hastie, Tibshirani, and Friedman, 2009)

input : Training data $\mathcal{D}\{x_i, y_i\}_{i=1}^{N}(x_i \in \mathbb{R}^p, y_i \in \mathcal{Y})$

output: An ensemble classifier g

Initialise $h_0(x) = \text{argmin}_\gamma \sum_{i=1}^{N} \ell(y_i; \gamma)$.

for $m \leftarrow 1$ *to* M *iterations* **do**

  for $i \leftarrow 1$ *to* N *observations* **do**

$$r_{im} = -\left[\frac{\partial \ell(y_i, h(x_i))}{\partial h(x_i)}\right]_{h=h_{m-1}} \tag{2.33}$$

  end

  Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm}, j = 1, 2, \ldots, J_m$.

  for $j \leftarrow 1$ *to* J_m *terminal nodes in the mth tree* **do**

$$\gamma_{jm} = \text{argmin}_\gamma \sum_{x_i \in \mathbb{R}_{jm}} \ell(y_i, h_{m-1}(x_i) + \gamma) \tag{2.34}$$

  end

  Update $h_m(x) = h_{m-1}(x) + \eta \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

end

return $g(x) = h_M(x)$

2.3.6.2 Gradient boosting in regression

The result of statistical exploration of the success of AdaBoost was the creation of a boosting method for learning via differentiation of an arbitrary loss function and base learners that is more robust to noise. The method was termed gradient boosting, and the classifiers gradient boosting models or gradient boosting machines (GBMs) (Friedman, 2001). In these, an approximate method is used for solving the forward stagewise model by minimising

$$\ell(h) = \sum_{i=1}^{N} \ell(y_i, h(x_i)), \tag{2.35}$$

where $h(x)$ is an additive combination of trees, as in equation 2.32, and $\ell(h)$ is the loss function. We seek the h that minimises $\ell(h)$ in order to obtain g. Algorithm 3, as given by Hastie, Tibshirani, and Friedman, 2009, describes the details of the traditional procedure for gradient boosting for regression problems, using a generic loss function as in equation 2.35. In the context of regression, typically half the squared error is used as the loss function to make taking the derivative easier. This procedure is started by initialising $h_0(x)$ as the mean response; this may be viewed as a single terminal node.

Following this, the procedure iterates through all M trees to compute the necessary steps (Algorithm 3). First, pseudo-residuals, r_{im}, are obtained from the negative gradient of the loss function for each observation i in the current tree m (equation 2.33). The derivative for regression can be simplified to $y_i - \hat{y}_i$, where $\hat{y}_i$ is the predicted value from the previous boosting round. Second, the vector r_{im} is taken as the response in fitting a regression tree. The tree partitions the feature space into J terminal regions, R_j, where $j = 1, 2, \ldots, J$. Third, the values γ_{jm} are computed from the j_{th} terminal nodes in the current tree, m, where γ is the predicted value for each observation x_i which reaches that terminal node, R_j (equation 2.34). For regression, this is simply the mean value in R_j. While the negative gradient in equation 2.33 gave the direction of steepest descent in which to move, equation 2.34 gives the step size. Finally, the function $h(m)$ is updated to combine previous trees with the output from the new tree, scaled by the learning rate. The learning rate therefore reduces the fit to the pseudo-residuals. Each tree takes a small step toward better prediction; the smaller the learning rate, the more slowly the model converges.

2.3.6.3 Gradient boosting in classification

For a binary outcome, the procedure follows that for regression quite closely. Twice the negative log-likelihood is taken as the loss function to minimise for classification,

$$\ell(\beta) = -2\sum_{i=1}^{N} y_i log(P(X)_i) + (1 - y_i) log(1 - P(X)_i), \tag{2.36}$$

similar to the maximisation of the log-likelihood (equation 2.10) shown for logistic regression. The initial estimate, $h_0(x)$, is taken to be the log(odds) of the outcome, $log(\frac{y}{1-y})$. The derivative taken to calculate the pseudo-residual for binary classification, r_{im}, given in equation 2.33, can again be simplified to be observed minus predicted; as predictions are on the log(odds) scale, the logistic function is used. For the first tree, $h_1(x)$, this is the outcome minus the logistic transformation of the log(odds) from $h_0(x)$. For the following step in classification, a regression tree is fit to the residuals. However, for equation 2.34 a slightly different approach is taken to obtain each γ_{jm} by using the log(odds) from each region R_j, as

$$\gamma_{jm} = \frac{\sum_{x_i \in \mathbb{R}_{jm}} w_{im}(y_i - p_i)}{\sum_{x_i \in \mathbb{R}_{jm}} w_{im} p_i (1 - p_i)}, \tag{2.37}$$

where p_i is the logistic transformation of the value in $R_j m$ and w_{im} are the sample weights in tree m. The final prediction is the sum of the initial log(odds) from $h_0(X)$ and the subsequent log(odds) produced by individual trees $1, 2, \ldots, M$, each weighted by the learning rate, η. The predicted probability is obtained from the this by the logistic function, $\frac{e^{h_M(x)}}{1+e^{h_M(x)}}$. Were an exponential loss function to be used instead, the resulting algorithm would be AdaBoost.

An important difference between standard CARTs and those used in gradient boosting is the splitting criterion. A typical CART for regression will use the mean squared error (MSE) for splits, $\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_i)^2$, for N observations, where $\bar{y}$ is the mean outcome in node m. In gradient boosting, the "proxy gain" is used to quickly find the best cutpoint, as the optimal split minimises both of these measures; MSE is then calculated once for the best chosen split. The gain is given by

$$Gain = \frac{G_{jmL}^2}{n_{jmL}} + \frac{G_{jmR}^2}{n_{jmR}}, \tag{2.38}$$

where G is the sum of the gradient of the loss function in the left (L) or right (R) hand daughter leaf from node j of tree m.

2.3.6.4 eXtreme Gradient Boosting

Modern implementations of gradient boosting use the method described in Algorithm 3 in addition to subsampling of observations, following work by Friedman demonstrating a further reduction in overfitting (Friedman, 2002). An additional hyperparameter controls the maximum depth of trees. This bounds the possible interactions that can be detected by gradient boosting, whereby a stump, with only 2 terminal nodes, can only detected main effects. More generally, a tree of depth k can detect $k-1$ order interactions. Together, learning rate, subsampling and tree depth are the main tunable hyperparameters in classic GBM implementations. Generally, more shallow trees with a slow learning rate (below 0.1) give improved prediction.

One of the most popular and effective implementations is eXtreme Gradient Boosting (XGBoost), a highly-optimised implementation of regularised GBMs (Chen and Guestrin, 2016). XGBoost makes several key changes to decrease overfitting.

First is to formalise regularisation in gradient boosting by adding penalty terms to the objective function. As with other methods introduced so far, the objective function is made up of the loss and the regularisation terms. Following the convention in Chen and Guestrin, 2016, the objective function is defined as

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \ell(\hat{y}_i, y_i) + \sum_{m=1}^{M} \Omega(h_m), \tag{2.39}$$

where Ω controls model complexity and θ represents the additive boosting model which produces y_i for each observation x_i. The unpenalised loss function is the second-order Taylor expansion of the loss after removing constant terms,

$$\ell(h_m) = \sum_{j=1}^{J_m} \sum_{i \in R_{jm}}^{N} [grad_{im} w_{jm} + \frac{1}{2} hess_{im} w_{jm}^2], \tag{2.40}$$

.

where $grad_i$ is first order derivative (gradient) and $hess_i$ is the second order derivative (Hessian) of the loss for observation i, and w_{jm} is the score in node j for iteration m. This is equivalent to

$$\ell(h_m) = \sum_{i=1}^{N} [G_{jm} w_{jm} + \frac{1}{2} H_{jm} w_{jm}^2], \tag{2.41}$$

where G_{jm} and H_{jm} are the sum of the gradient and Hessian at the mth iteration in node j. Optimisation of the loss is only a function of the gradient and the Hessian, unlike for GBMs. Applying both L_1 and the squared L_2 norm, the regularisation term is

$$\Omega(h_m) = \gamma T + \frac{1}{2}\lambda ||w||_2^2 + \alpha ||w||_1, \tag{2.42}$$

for T leaves in a tree, where w is a vector of scores in the terminal nodes (Chen and Guestrin, 2016). Complexity is therefore controlled at the tree-level using three hyperparameters. γ limits the size of the tree by setting the minimum gain required to split. α and λ adjust the strength of the L_1 and L_2 penalties on the terminal node values, respectively. XGBoost allows for use of either L_1, L_2 or both, similar to elastic nets.

Second, the gain function is updated as

$$Gain = \frac{1}{2}\left[\frac{G_{jmL}^2}{H_{jmL} + \lambda} + \frac{G_{jmR}^2}{H_{jmR} + \lambda} - \frac{(G_{jmL} + G_{jmR})^2}{H_{jmL} + H_{jmR} + \lambda}\right] - \gamma, \tag{2.43}$$

for the L_2-penalised loss. It is not feasible to search all possible trees to optimise equation 2.41 at each step. Instead, trees are built using the modified gain function above to optimise in a level-wise manner for each tree. This uses both the gradient and Hessian to improve the estimate of the direction of steepest descent. This is in contrast to standard GBMs, which use only the gradient to perform splits (equation 2.38).

Third is the introduction of column subsampling to gradient boosting, similar random forests, on either the tree or at each level in the tree; standard GBM implementations now often also implement tree-level predictor subsampling. There are a range of additional hyperparameters available in XGBoost to control other aspects of learning, such as fine-tuning aspects of tree growth or including additional regularisation techniques. XGBoost also implements several performance enhancements not covered here, including optimisation for sparse data, caching for out-of-core computation, an approximate split-finding algorithm, and blocking for improved parallelisation. These allow for more thorough hyperparameter search when computation is expensive.

2.3.6.5 Comparison to random forests

Though random forests and gradient boosting seem superficially similar, the techniques actually represent contrasting approaches. Random forests are a tree-specific method that averages many models with low bias and high variance as a natural extension to bagging. They improve prediction primarily by reducing variance; generally, adding more trees cannot make a model worse. In contrast, boosting is a generic modelling approach which often works best with shallow decision trees as weak learners. This is because it combines many models with high bias and low variance. It primarily improves prediction by reducing bias, and adding more trees only improves prediction up to a point, after which it can get worse.

The two do share the same fundamental idea that individual models can be combined to improve prediction in a way that focuses on reducing bias or variance. This has proven to be a particularly useful paradigm for increasing prediction performance. As computing power has increased, and prediction modelling competitions gained in popularity, algorithms have moved beyond simply combining small, fast models like decision trees. The next section addresses a more general approach.

2.3.7 Averaging, stacking and blending

Random forests and gradient boosting provide large gains over their composite learners. Where predictive ability is the main concern, smaller gains can be made by considering ensembles of more complicated models. These take reasonably competitive, predictive models such as SVMs and random forests and combine them together into a single score or model. Their success relies on taking diverse techniques which make different assumptions in model-building and combining these together, such they collectively span the hypothesis space of approaches for estimating f on a given dataset. They follow similar logic to bagging and random forests: taking an average of less-correlated models reduces the variance and improves prediction.

Ensembles can be partitioned into the two operations of generating base learners on the training data, and aggregating them into a final model (Hastie, Tibshirani, and Friedman, 2009). It has become more possible and popular to use ensembles in recent years as online machine learning competitions such as Kaggle have flourished, where small improvements in discriminative ability determine rankings of competitors, and the average computational budget available to practitioners has increased. Despite this, the concept of creating ensembles is certainly not new. Error-correcting codes, proposed by Dietterich and Bakiri, 1994, for improved prediction in multi-class problems, provide intuition around the improvements seen in model averaging. Error correcting codes employ a distributed representation of each class in a multi-class problem. A p-length code for each output is devised, as shown in Table 2.2, such that each class is uniquely represented by its code, and p separate binary models are trained to predict each column of the coding. The final prediction is taken by combining the output of each binary model to give a predicted code. Later work showed that this method

was partly effecting because it trains p binary models, instead of a single multi-class model, and combines them, reducing the variance of predictions (Kong and Dietterich, 1995).

Class	A	B	C	D	E	F
0	0	0	0	1	0	0
1	1	0	0	0	0	0
2	0	1	1	0	1	0
3	0	0	0	0	1	0
4	1	1	0	0	0	0
5	1	1	0	0	1	0
6	0	0	1	1	0	1
7	0	0	1	0	0	0
8	0	0	0	1	0	0
9	0	0	1	1	0	0

TABLE 2.2: Distributed codes for a 10-class digit-recognition problem. Each class is represented by a 6-digit code. 6 binary classifiers are trained with one of the columns as the outcome. As codes overlap, all columns except F involve multiple classes. Their combined prediction produces a 6-digit code, which is compared to the true codes by Hamming distance (number of mismatches). Adapted from Dietterich and Bakiri, 1994.

The 3 most common approaches for combining competitive models are considered. The first is a simple averaging of predictions. The most basic form of this is taking a majority vote (hard voting), but more commonly predicted probabilities are averaged (soft voting). Individual classifiers may also be weighted by a learned or pre-chosen measure. While effective and easy to implement, averaging relies on all models producing a probability, as opposed to the decision boundary used in SVMs, for instance. It also requires all models to be

well-calibrated.

Algorithm 4: Stacking, adapted from (Aggarwal, 2015)

input : Training data $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N} (x_i \in \mathbb{R}^p, y_i \in \mathcal{Y})$

output: An ensemble classifier g

Step 1: Learn first-level classifiers

for $t \leftarrow 1$ *to* T *base classifiers* **do**

  Learn a base classifier h_t based on $\mathcal{D}$

end

Step 2: Construct new data sets from $\mathcal{D}$

for $i \leftarrow 1$ *to* N *observations* **do**

  Construct a new data set that contains $\{x_i', y_i\}$, where $x_i' = \{h_1(x_i), h_2(x_i), \ldots, h_T(x_i)\}$

end

Step 3: Learn a second-level classifier

Learn a new classifier h' based on the newly constructed data set

return $g(x) = h'(h_1(x), h_2(x), \ldots, h_T(x)$

A solution to these problems is to use a meta-estimator. This is termed stacking, and was first introduced before the bagging or random forest procedures described by Breiman (Wolpert, 1992). The idea is simply taking the predicted classes or probabilities for each model, where these should have reduced correlation for better improvements, and 'stacking' them together as meta-features or meta-predictors. These are used as the training data for a meta-estimator such as logistic regression, decision tree or random forest. Regression has been suggested as the most appropriate choice for meta-estimator (Ting and Witten, 1999). The models which produce the meta-predictors are termed first-level classifiers or base estimators and the outer model which combines these is known as the second-level classifier or meta-estimator. In practice, it is much more common to take probability of class membership than class assignments, as the former has been shown to be essential for successful stacking in classification (Ting and Witten, 1999). The predictions from the

trained meta-estimator are taken as the final predictions from the whole ensemble.

Algorithm 5: Stacking with K-fold cross-validation, adapted from (Aggarwal, 2015)

input : Training data $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N} (x_i \in \mathbb{R}^p, y_i \in \mathcal{Y})$

output: An ensemble classifier g

Step 1: Adopt cross-validation approach in preparing a training set for second-level classifier

Randomly split $\mathcal{D}$ into K equal-size subsets: $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$

for $k \leftarrow 1$ *to* K *folds* **do**

  Step 1.1: Learn first level-classifiers

  for $t \leftarrow 1$ *to* T *base classifiers* **do**

    Learn a base classifier h_{kt} from $\mathcal{D} \backslash \mathcal{D}_k$

  end

  Step 1.2: Construct a training set of second-level classifier

  for $x_i \in \mathcal{D}_k$ **do**

    Get a record $\{x_i', y_i\}$, where $x_i' = \{h_{k1}(x_i), h_{k2}(x_i), \ldots, h_{kT}(x_i)\}$

  end

end

Step 2: Learn a second-level classifier

Learn a new classifier h' from the collection of $\{x_i', y_i\}$

Step 3: Re-learn first level classifiers

for $t \leftarrow 1$ *to* T *base classifiers* **do**

  Learn a classifier h_t based on $\mathcal{D}$

end

return $g(x) = h'(h_1(x), h_2(x), \ldots, h_T(x))$

A concern when using stacking is multicollinearity, wherein one variable in a dataset can be predicted by other variables using a linear model. This has been suggested to be an issue for stacking, resulting in an ensemble which does not out-compete base classifiers (Dumancas and Bello, 2015). However, the principle concern in stacking classifiers is preventing information leakage from the base classifiers to the stacked meta-estimator. This can easily occur if a simple split-sample approach is taken, where the base estimators are fit to the training data, and give predictions on that same data (Algorithm 4). These training set predictions are then used as input to learn the meta-estimator in the training set. Each base estimator gives predictions for the test set, which feeds into the meta-estimator to give the final test set predictions. Such an approach can lead to overfitting.

Applying cross-validation to stacking reduces the risk of overfitting (Aggarwal, 2015). A single train-test split is first applied, followed by cross-validation in the training fold. Algorithm 5 describes this improved procedure, wherein the predictions from the base estimators in the training split are the test-fold predictions from cross-validation. The stacked predictions from base estimators are therefore all predictions on unseen data, from distinct but correlated models. The base estimators must then be refit to the whole training set before predicting

on the test set. This runs less risk of leading to overfitting as it avoids the meta-estimator giving more weight to more flexible models which may have overfit the training data (Hastie, Tibshirani, and Friedman, 2009). In small datasets, this process can be adjusted for nested cross-validation by simply repeating the split-sample approach K times on K equal-sized non-overlapping outer folds. Alternatively, a hold-out set may be reserved specifically for training a meta-estimator. This is sometimes termed blending (Töscher, Jahrer, and Bell, 2009), but is less frequently used as it requires a much larger dataset to be possible. The use of stacking, blending or averaging will likely incur reduced interpretability of the model, but with the possible gain of improved prediction.

An additional concern in stacking is multicollinearity, wherein multiple predictors show relationships. While machine learning methods are often noted for not having strict requires for multicollinearity that are rpesent in classicial statistical approaches, stacking has been shown to be affected by this ().

2.3.8 Neural networks

Deep learning is a subset of machine learning focused on using neural networks. As with gradient boosting, neural networks make use of sequential optimisation of an objective function. They differ by applying a network of weights to the input which are iteratively updated through backpropagation to learn associations with the outcome (LeCun, Bengio, and Hinton, 2015). Specific notation is adopted for discussing neural networks:

- L - the number of layers in a network
- s_l - the number of neurons in the l^{th} layer
- z_l^m - the linear combination of inputs and weights for neuron m in layer l
- g_l^m - the activation function of neuron m in layer l
- a_l^m - the output of neuron m in layer l
- w_l - a matrix of weights from layer l to layer $l+1$
- b_l - the bias term for layer l
- T - the number of gradient updates
- w_l^t the weight matrix for the l^{th} layer in the t^{th} update

2.3.8.1 Neurons

The fundamental unit of a neural network is the neuron, shown in the centre of Figure 2.17. A neuron, or node, is a single unit of computation with two functions. The first takes a linear combination of inputs z,

$$z = w^T x + b, \tag{2.44}$$

using learned weights w, and adds a bias term b. The second applies an non-linear activation function to transform the output. The single-neuron model in Figure 2.17 is known as a perceptron.

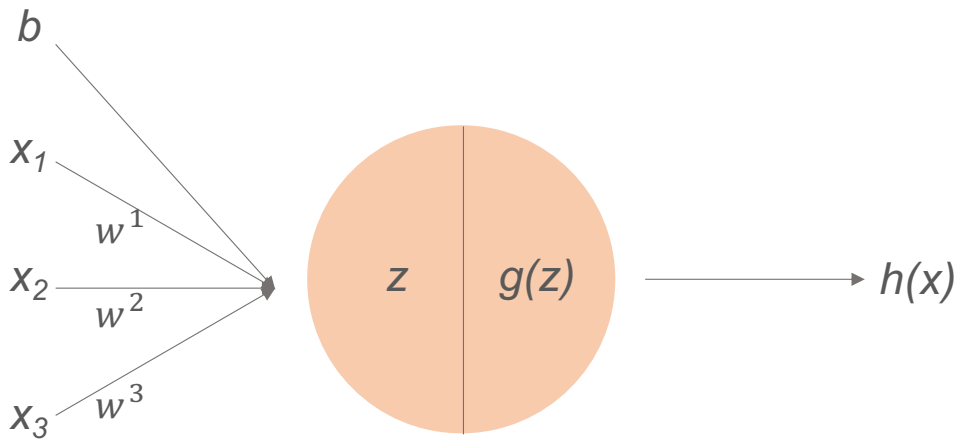


FIGURE 2.17: A perceptron. A linear combination of weights and predictors is added to the bias, before being passed through a non-linearity (activation function).

Many activation functions are available in neural networks. The absence of a non-linearity, sometimes termed a linear activation function, applies no transformation to the output. A network with only linear activations can only learn a linear function, no matter how deep the network. In binary classification, a sigmoid function is used at the terminal node. This is applied to the scalar from z to give the final output of a neuron, a, as

$$a = g(z) = \frac{1}{1 + e^{-w^T x + b}} \tag{2.45}$$

While sigmoid is used for the output layer for binary classification, the hyperbolic tangent activation function,

$$g(z) = tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}, \tag{2.46}$$

and rectified linear unit (ReLU) activation function,

$$g(z) = \begin{cases} z & \text{if } z > 0 \\ 0 & \text{if } z \leq 0 \end{cases}, \tag{2.47}$$

are often used for layers within the network. ReLU is a piecewise linear function which becomes the identity function when $z > 0$. It is perhaps the most popular activation function as computation is cheap - the derivative is only ever 0 or 1 - and it solves the "vanishing gradient problem" when training deep neural networks, an issue where the gradient of the

tanh and sigmoid functions can become vanishing small, preventing the loss being propagated through the network. Many modifications to these exist, including leaky ReLU and exponential linear unit (ELU). Leaky ReLU is a suggested fix for the "dying ReLU" problem, whereby the computation for z is negative, causing ReLU to output zero for $g(z)$, resulting in a gradient of zero and no further learning for that neuron. Leaky ReLU changes g to a small value when $z < 0$ by using

$$g(z) = \begin{cases} z & \text{if } z \geq 0 \\ \alpha z & \text{if } z < 0 \end{cases}, \tag{2.48}$$

where α is a constant, typically around 0.1. ELU also adapts ReLU by taking the exponent of negative inputs and scaling them, using

$$g(z) = \begin{cases} z & \text{if } z \geq 0 \\ \alpha(e^z - 1) & \text{if } z < 0 \end{cases}. \tag{2.49}$$

Whether there is a benefit from using these alternatives is not clear, and researchers often stick to ReLU for fast computation with large networks.

2.3.8.2 Networks

Neurons are combined together into networks as connected layers (Figure 2.18). In the simplest form, neurons in each layer take inputs from the previous layer and pass outputs to the next. This is initiated with the inputs to the network, which is not counted as a layer here, so that a network consisting of inputs, 1 hidden layer and an output layer is considered to be a 2-layer network. In the final layer the node, or nodes, provide the prediction. Each neuron in a layer is connected only to neurons in the adjacent layer, but not to other neurons in the same layer. This structure is a feed-forward network, and may be referred to as "dense". Between the input and output layers, so-called hidden layers may be used. Use of one or more hidden layers typically earns the moniker "deep" learning. The feed-forward network described here is a multi-layer perceptron.

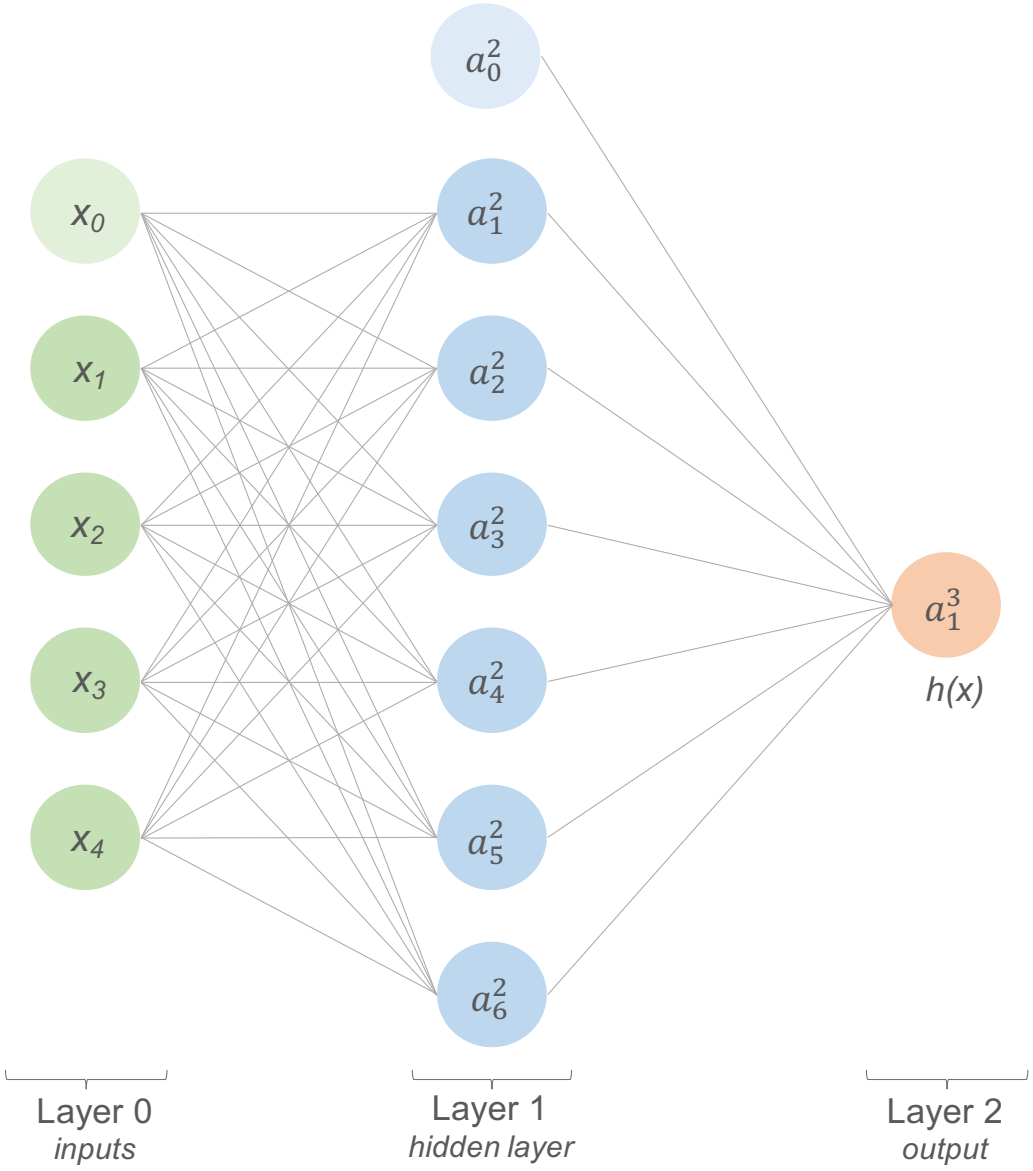


FIGURE 2.18: Neural network architecture, where each circle is a neuron and lines denote weights. Inputs are connected to a hidden layer through weights. Each predictor connects to each node in the next layer in a dense multi-layer perceptron. The hidden layer feeds forward to the output layer, a^3. Bias terms are present for inputs, x_0, and hidden layers, a_0^2

Taking the computation from a single neuron in equation 2.45, computation of layers can be considered. For a given layer l with s_l units, the matrix of weights w_l connecting it to a layer $l+1$ with s_{l+1} units is an $s_{l+1}xs_l$ matrix, excluding the bias. In Figure 2.18, the matrix of weights $w_1 \in \mathbb{R}^{6x4}$ connects the input vector $x = [x_1, x_2, x_3, x_4]^T$ to layer 1 (the hidden layer). The output from this, $z_2 = w_1 x + b_1$, undergoes element-wise transformation to give $a_2 = g(z_2)$ using the chosen activation function, where $a_2 \in \mathbb{R}^6$, excluding the bias term.

Passing values through the network, from left to right in Figure 2.18, is known as forward propagation. The choice of number of neurons and number of layers affects the size of the

network and its capacity to learn. The number, layout and connections between neurons is known as the architecture of a network. There is no clear rule for choosing an architecture. The input and output layers are set by the data. Rules-of-thumb for hidden layers typically rely on keeping the number of layers small for tabular data with a small-to-moderate sample size, often 1 to 3, and constraining the number of hidden neurons to limit the risk of overfitting. The geometric pyramid rule suggests that number of hidden neurons should decrease as layers get deeper in the network, and that a single hidden layer with $\sqrt{s_1 * s_L}$ neurons is acceptable (Masters, 1993), where s_1 and s_L are the size of the input and output layers respectively. Other sources suggest a good approximation can be achieved by choosing the size of a hidden layer to be between s_1 and s_L, or between $\frac{2}{3}s_l$ and $2s_l$ (Heaton, 2008). Both sources suggest such heuristics achieve a reasonable starting point, but multiple architectures should be tried for a dataset.

Neural networks are not a recent invention. *In silico* neuronal models have existed since the mid-20$^{\text{th}}$ century as a linear combination of binary values (McCulloch and Pitts, 1943); Rossenblatt's perceptron and its associated learning procedures followed in the ensuing decades. Despite such early work, interest in applying neural networks has waxed and waned through several AI "winters", periods of diminished interest and funding that occurred principally in the 1970s to 1990s. Modern neural network architectures enjoy huge success due to a surfeit of funding, open source software, computational power and, in many fields, data. Preceding all this was a fundamental development in the 1980s that provided an effective mechanism for training larger neural networks.

2.3.8.3 Gradient descent

Neural networks are trained by gradient descent, which involves updating weights through backpropagation (Rumelhart, Hinton, and Williams, 1985). In order to achieve this the gradient of the objective function must be determined. The negative gradient gives the direction of steepest descent in which to step. First, weights are randomly intialised. These are typically drawn from a standard normal distribution before reducing the variance to an appropriate scale, for example by $\sqrt{(\frac{2}{N})}$ when using ReLU (He et al., 2015), where here N is the number of weights. Next, the partial derivative of each node in the L^{th} layer is calculated with respect to the loss function. Once the gradient has been computed, weights are updated,

$$w_1^{t+1} = w_1^t - \eta \partial w_1 \tag{2.50}$$

$$b_1^{t+1} = b_1^t - \eta \partial b_1 \tag{2.51}$$

where η is the learning rate that controls the size of the step. η is a hyperparameter that is tuned in training, with values typically chosen from a logarithmic scale in the range 10^{-5}

to 10^{-1} (Goodfellow et al., 2016). The partial derivative is then calculated for each node in the $L-1$ layer, followed by a weight update, then layer $L-2$ and back to the first layer, each time propagating results backwards through the network.

Data can be passed to the network in mini-batches, often with 2^n observations (e.g. 16, 32, 64, 128). Scaling of each mini-batch, rather than on the entire training set, improves speed of convergence (Ioffe and Szegedy, 2015). Typically a z-transformation is applied to bring predictors into the same range.

Gradient descent is repeated for a number of epochs (complete passes through the data) until the loss function begins to converge. This can be monitored by plotting the loss against epochs, or applying a test designed to automatically check for convergence. The number of epochs to train for depends on the learning rate and the problem. A smaller learning rate will require a larger number of epochs to reach convergence.

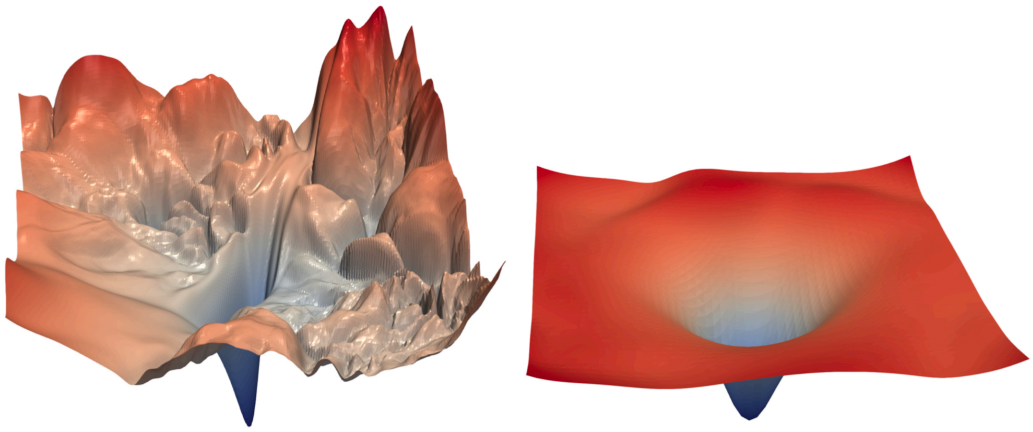


FIGURE 2.19: Visualisation of the loss landscape. Smoothness can be affected by different choices of hyperparameters during learning. The figure illustrates this for a type of convolutional neural network, ResNet-56 (left), compared to the same network with skipped connections (right). Adapted from (Li et al., 2017a)

Variations on gradient descent can be used to alter the number of observations used to compute a step, or apply a constant or adaptive learning rate. Instead of computing the loss for a whole dataset (batch) or chunk of it (mini-batch), stochastic gradient descent (SGD) takes a single observation at a time. This greatly increases computation speed but also causes much more variation in the gradients; steps are more affected by noise, move in opposite directions and take much longer to approach convergence. An adaptation to gradient descent which improves speed and often performance is momentum (Sutskever et al., 2013). This applies exponentially weighted averaging over rounds of gradient descent to smooth-out the steps taken, resulting in a more direct path to the optimum. Use of momentum adds another parameter to adjust the strength of the weighted average, with values typically set between 0.8 and 0.99.

Architecture and hyperparameter choices made in training can vastly change the landscape for the loss function (Figure 2.19). Appropriate choice of learning rate and momentum can

attain convergence on local minima in a multilayer perceptron, but does not ensure the global minimum is reached. Adaptive learning rates can be used to help speed up or decrease the size of the steps in gradient descent depending on the landscape of the loss function. Adaptive moment estimation (Adam) (Kingma and Ba, 2014), adaptive gradients (AdaGrad) (Duchi, Hazan, and Singer, 2011) and root mean squared propagation (RMSprop) (Tieleman and Hinton, 2012) and other techniques apply such learning rate schedulers, allowing for much larger initial steps to be taken, followed by smaller steps as the loss approaches convergence. The most popular and successful of these, Adam, combines momentum and the adaptive learning rate method from RMSprop.

Gradient descent performed in training neural networks is similar to that used for training gradient boosting machines. The key difference between these is that neural networks use gradient descent to update the weights directly, and so employ it on the input space. GBMs update the residuals for each round of boosting based on the gradient, performing gradient descent in the function space. Neural networks use gradient descent to train a single strong model, while GBMs use it to boost the prediction from multiple weak models.

The ability to switch-out different activation functions and objective functions highlights a small fraction of the flexibility of neural networks. For instance, a single neuron with a linear activation function and a hinge loss is equivalent to a linear SVM; alternatively, attaching a sigmoid function and minimising the log-likelihood is equivalent to logistic regression, though both are trained here through gradient descent.

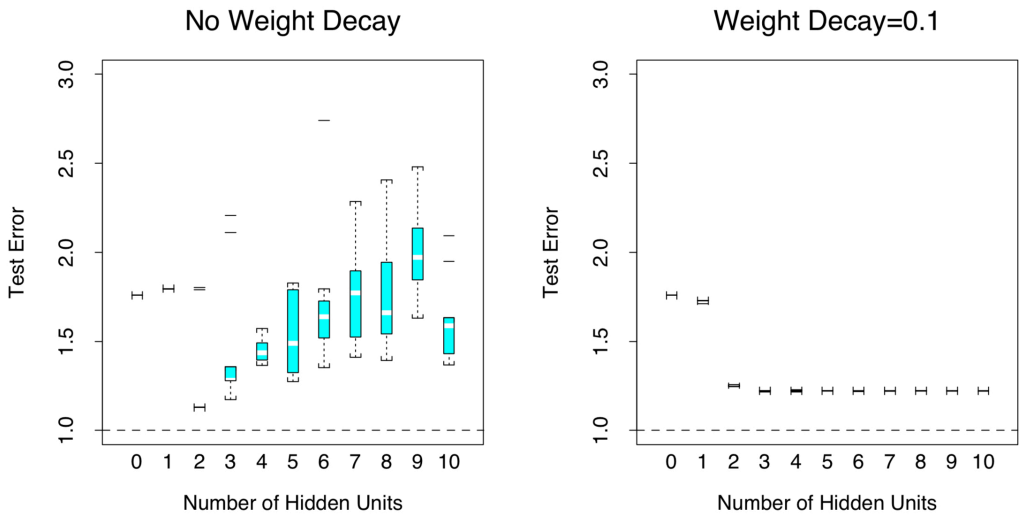


FIGURE 2.20: Regularisation by weight decay for simulations with 2 predictors, 100 train observations and 10,000 test observations. Test error is shown for a neural network with one hidden layer with (right) and without (left) L_2 regularisation. The results show the distribution across 10 random weight intialisations, for varying number of hidden neurons. 0 nodes is a network with no hidden layer. Adapted from (Hastie, Tibshirani, and Friedman, 2009).

2.3.8.4 Regularisation

Large networks can easily overfit in small datasets. Several methods are available for controlling this. As with other models described so far, an L_1 or L_2 penalty can be used to constrain optimisation of the weights. This is termed weight decay. L_2 regularisation is the most common (Goodfellow et al., 2016); it is often reported to give improved prediction over L_1, and allows for more stable predictions when deeper architectures are used (Figure 2.20). Values for weight decay are normally chosen from a logarithmic scale between 0 and 0.1.

An alternative method is dropout (Srivastava et al., 2014). This randomly zeros-out neurons in a hidden layer, with a pre-specified probability. Different neurons are zeroed-out in a layer for each iteration of gradient descent, forcing the model to be more robust by learning alternative routes through the network. The most common implementation of this is inverted dropout, whereby output a_l from a layer is scaled relative to the probability of dropout, to account for the decreased value from z_l due to the zeroed-out nodes. Dropout is simple and effective, and may be used alone or in addition to other regularisation techniques like weight decay. At prediction time, the whole network is used. A further option is early stopping. Here, convergence is monitored in a hold-out set, and optionally also in the training set. Training is stopped if no improvement is seen within a set number of epochs, often labelled the "patience", in the hold-out set. This allows learning to be halted before overfitting occurs, but is only possible in large datasets where a portion of training data can be set aside for monitoring.

Though the multi-layer perceptron described here can be increased to a large number of hidden nodes and layers, it remains the simplest architecture. Much more complicated structures have been developed for specific applications. Principal among these are the convolutional neural network (CNN) and the recurrent neural network (RNN). CNNs combine 1, 2, or 3-dimensional convolution layers with pooling layers to greatly reduce the number of parameters in a model, culminating in a fully-connected feed-forward network. They are specialised for image processing but have been adapted to other fields. RNNs maintain an internal state to process sequence data, such as language models.

2.4 Summary

Machine learning comprises a broad array of methods for leaning from data. This chapter has introduced commonly-used methods which approximate a target function f using g. Each learning algorithm $\mathcal{A}$ makes assumptions to produce a final competitive prediction. The chapter has also detailed the methods of developing, validating and comparing models.

The most competitive machine learning models are used in Chapters 4 and 5: ridge, LASSO, linear SVM, RBF SVM, random forest, XGBoost and neural networks. These are compared

to polygenic risk scores whenever estimates of effect sizes are available and logistic regression where $p < n$. Decision trees and AdaBoost were both described in this chapter, but are not implemented in analyses; decision trees are liable to high variance and low bias, while the exponential loss function in AdaBoost makes it unlikely to perform well in genetic data for psychiatric disorders. While chapter 4 aims to compare many methods to identify which perform best under different conditions, chapter 5 seeks to maximise prediction of schizophrenia from real data, and so also incorporates stacking and additional assessment of the predictive models.

With few exceptions, no machine learning method has a clear theoretical basis for being deemed superior or inferior for genetic prediction. The next chapter systematically reviews the literature to assess the predictive performance of supervised machine learning methods in psychiatric genetics.

Chapter 3

Systematic Review of Machine Learning Methods for Genetic Prediction of Psychiatric Disorders

3.1 Introduction

Machine learning represents a contrasting approach to traditional methods for genetic prediction. It has increased in popularity in recent years following breakthroughs in deep learning (Glorot, Bordes, and Bengio, 2011; Hinton et al., 2012; Krizhevsky, Sutskever, and Hinton, 2012; Sutskever, Vinyals, and Le, 2014), and the scaling-up of datasets and computing power. The ability to function in high dimensions and detect interactions between loci (Cordell, 2009) without assuming additivity makes such methods an attractive option in statistical genetics, where the effects of myriad factors on an outcome is difficult to prespecify. Calls to address the complexity of disorders like schizophrenia with machine learning have also become more frequent (Krystal et al., 2017; Schnack, 2017; Tandon and Tandon, 2018). However, the predictive performance of machine learning methods in psychiatric genetics is unclear, and a recent review of clinical prediction models across various outcomes and predictors found them to be no more accurate than logistic regression (Christodoulou et al., 2019); it is therefore timely to review their predictive performance in psychiatry.

Genome-wide association studies, genetic prediction and psychiatry have each been reviewed with respect to machine learning. Chen and Ishwaran, 2012, give a detailed history of the algorithmic modifications that have been made to random forests for application to genetic data. A methodological focus is also taken in a review of machine learning in psychiatry, with attention paid to the principles of machine learning, which methods are available and application to large psychiatric datasets (Iniesta, Stahl, and McGuffin, 2016). A more recent review of deep learning, where the authors cover common architectures of neural networks, highlights examples of neuroimaging and data from health records or electronic devices being used for prediction in psychiatry (Durstewitz, Koppe, and Meyer-Lindenberg, 2019). By contrast, two further reviews look across diseases to give an overview of machine learning with respect to interactions, predictor selection and regularisation (Okser, Pahikkala, and Aittokallio, 2013; Okser et al., 2014). Single nucleotide polymorphism (SNP)-based

prediction has also been reviewed across diseases in two articles. An earlier publication aimed to systematically review machine learning on data from genome-wide association studies (GWAS) (Kruppa, Ziegler, and König, 2012). They identify 115 papers, 91 of which apply machine learning (ML) to candidate loci; none of the articles included by Kruppa, Ziegler, and König, 2012, overlap with the present chapter. Prediction from SNPs has also been reviewed more recently (Ho et al., 2019). Here, the authors give an overview of penalised regression and tree-based models with examples from the literature, highlighting the differences between machine learning and polygenic risk scores.

Though several reviews have sought to address aspects of machine learning and genetics across diseases, psychiatry presents a distinct problem from somatic and neurological diseases as a result of genetic correlation between disorders (Anttila et al., 2018) and the risk of class mislabelling due to biological heterogeneity that may underlie symptom-based diagnoses (Kapur, Phillips, and Insel, 2012). In addition, only two previous reviews of machine learning and psychiatric disorders are systematic. The first of these evaluated 51 studies for the use of machine learning and any data-type to predict bipolar disorder (Librenza-Garcia et al., 2017). 5 genetic studies were included in the review by Librenza-Garcia et al., 2017, 2 of which are also included here (Pirooznia et al., 2012; Acikel et al., 2016). The remaining 3 were excluded from the current review as they exclusively used gene expression data, only used random forests to perform quality control, or combined multiple disorders into a single outcome. The second systematic review assessed 26 studies for use of machine learning in prediction of treatment-related outcomes in depression (Lee et al., 2018). 3 studies with a genetic predictor were included by Lee et al., 2018, but all were excluded in this chapter as they involved prediction of treatment response instead of a disorder.

Systematic reviews have become an essential facet of the medical literature, requiring authors to apply "systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyse data from the studies that are included in the review" (Higgins et al., 2019). Ideally, they involve an assessment of risk of bias (Moher et al., 2009), a systematic issue in a study which causes optimistic or distorted performance measures (Wolff et al., 2019). Given the potential for machine learning methods to overfit and the often complex modelling steps occurring within a pipeline, analysis of risk of bias is a crucial step in evaluating approaches. However, no previous reviews on machine learning in psychiatric genetics have systematically assessed within-study risk of bias.

3.1.1 Aims and objectives

The aim of this chapter is to establish the current state of machine learning methods for prediction in psychiatric genetics. The objectives are:

- Systematically search, extract, summarise and present literature on machine learning methods for prediction of psychiatric disorders from genetic data

- Report the ability of machine learning methods to discriminate between cases and controls
- Assess studies for inflation of performance measures through evaluating risk of bias in participants, predictors, outcome and analysis
- Summarise methods used by researchers in model development and validation

Literature was systematically reviewed related to the question: what is the ability of machine learning methods to predict psychiatric disorders using only genetic data? Discrimination, methodology and potential bias is reported for diagnostic or prognostic models and compared to logistic regression (LR) and polygenic risk scores (PRS) where available.

3.2 Methods

3.2.1 Search strategy

Medline via Ovid, PsychInfo, Web of Science and Scopus were searched for journal articles matching terms for machine learning, psychiatric disorders and genetics on 10th September 2019. Searches were broad, with terms for psychiatric disorders including schizophrenia, bipolar, depression, anxiety, anorexia and bulimia, attention-deficit hyperactivity disorder, obsessive compulsive disorder, Tourette's syndrome or autism. Terms for machine learning were also wide-ranging, including naïve Bayes, k-nearest neighbours (k-NN), penalised regression, decision trees, random forests, boosting, Bayesian networks, Gaussian processes, support vector machines and neural networks, but excluding regression methods without penalty terms, such as logistic regression. Searches were restricted to English language journal articles on humans, with no limits on search dates. All abstracts were independently reviewed for inclusion by two individuals. Full texts were assessed and independently screened against inclusion criteria if either individual had chosen to access them. Where conflicts occurred a third researcher was consulted as an arbiter. Duplicate reading of articles, extraction of discrimination, and assessment of risk of bias in studies was assisted by Karen Crawford; resolution of ties between assessors was assisted by Valentina Escott-Price, and all three parties contributed to the development of a data extraction form. All other work is the product of the author. An example search for Medline (Ovid) is given in appendix A (Table A.1).

3.2.2 Inclusion and exclusion criteria

Studies were restricted to cohort, cross-sectional or case-control designs of individuals for binary classification of a single DSM or ICD-recognised psychiatric disorder compared to unaffected individuals, where only genotyping array, exome or whole-genome sequencing data were used as predictors. Studies based solely on gene expression were excluded, but designs which made use of gene expression or functional annotations to inform models of genetic data were accepted. Studies which combined genetics and non-genetic data (such as

from neuroimaging) were included if they developed or validated a model using only genetic data. Models were only considered for inclusion if they contained two or more genetic predictors from more than one locus.

No further restriction was made on participants. Studies were excluded if they only predicted medication response, sub-groups within a psychiatric disorder or a psychiatric phenotype secondary to another disease. Psychiatric disorders were limited to those with demonstrated heritability and for which large association studies have been undertaken; neurological conditions with psychiatric comorbidities were excluded. A machine learning or statistical learning method was required to be used as the prediction model, with models only using ML for quality control or predictor selection not considered. Studies were also considered ineligible if they had a clear primary aim of drawing inference at the expense of prediction, if they developed a novel statistical method or only made use of unsupervised or semi-supervised methods. The review was registered to PROSPERO in advance (registration number CRD42019128820). Changes were made to the registered protocol to further restrict the review's scope, and to clarify inclusion and search criteria before completing database searches.

3.2.3 Extraction and analysis

A data extraction form was developed prior to screening of publications. Items from the critical appraisal and data extraction for systematic reviews of prediction modelling studies (CHARMS) checklist (Moons et al., 2014) were included as-is or modified, and additional items were included based on expert knowledge and relevance to the review topic, with reference to the genetic risk prediction studies (GRIPS) statement (Cecile et al., 2011) for items pertaining to genetic prediction studies. The complete form is given in the appendix (Table A.2). The form was piloted with five publications, containing 40 extracted ML models between them, and updated before being applied to all texts by simplifying questions and merging related items.

Where models were fit and internally evaluated before external validation in a single study, information was extracted for both internal and external validation. Internal validation is taken to be any form of evaluation on a subset of the same sample used for training, including splitting samples between training and test sets, bootstrapping and k-fold cross validation. Apparent validation, where training and testing are both done on the whole sample, is also recorded under internal validation for this chapter. Any form of internal validation is considered to be part of model development. External validation is understood as evaluation on an independent dataset, which differs in temporal, geographic or other aspects, and is not simply a splitting-off from the original sample. If multiple models were presented with subsampled predictors or participants, only main models presented in the text were extracted; if such a distinction was unclear, all models were selected for review.

The discrimination of machine learning methods was extracted as area under the receiver

operating characteristic curve (AUC), or c-statistic. Model performance measures for classification by accuracy, sensitivity and specificity were also extracted. Where area under the receiver operating characteristic curve (AUC) was only available graphically it was extracted from the figure using Plot Digitizer (*Plot digitizer*), and accuracy was calculated from the confusion matrix if not provided in-text. 95% confidence intervals for validation were estimated for AUC using Newcombe's method (Debray et al., 2019). Results were not meta-analysed due to sample overlap, present in at least half of studies (see Table A.4), which cannot easily be accounted for in the meta-analysis. Though the logit-transformed AUC provides a more stable inter-study measure of discrimination (Snell et al., 2018), AUC is presented here on the original scale for ease of interpretation and because a meta-analysis was not performed, so transformation was not required. Information on participants, predictors and model development and validation were also obtained, with LR or PRS models extracted when present. The presence of LR and PRS as comparators was not made a requirement due to their sparsity in the literature.

3.2.4 Risk of bias

Risk of bias (ROB) and applicability were assessed using the prediction model risk of bias assessment tool (PROBAST) (Wolff et al., 2019). PROBAST consists of 20 questions designed to signal where ROB may be present in either the development or validation of a model across 4 categories: participants, predictors, outcome and analysis (Table A.3). These include, for instance, questions on how missingness or complexities in study design were handled. Information on handling of population structure, a common confound in genetic association studies, was also extracted to aid ROB assessment (see Table A.2).

Where information was unavailable within a study, any references or links given for descriptions of datasets or methods were examined. The signalling questions from PROBAST remained unchanged; however, recommendations for assessing studies using genetics and machine learning were added to adapt the tool and keep consistency in answers across models and reviewers (appendix A). Reporting of the systematic review follows the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines (Moher et al., 2009).

3.3 Results

3.3.1 Selection

1,241 publications were identified through searches in Ovid Medline, PsychInfo, Scopus and Web of Science which included restrictions to English language journal articles (Figure 3.1). After merging and removing duplicates, 652 studies were assessed for inclusion. Of these, 63 full texts were assessed to determine eligibility. 14 publications were selected, with two merged following Cochrane recommendations (Higgins et al., 2019) as publications included

the same models on the same dataset. A final total of 13 studies were selected for inclusion, containing 77 distinct machine learning models.

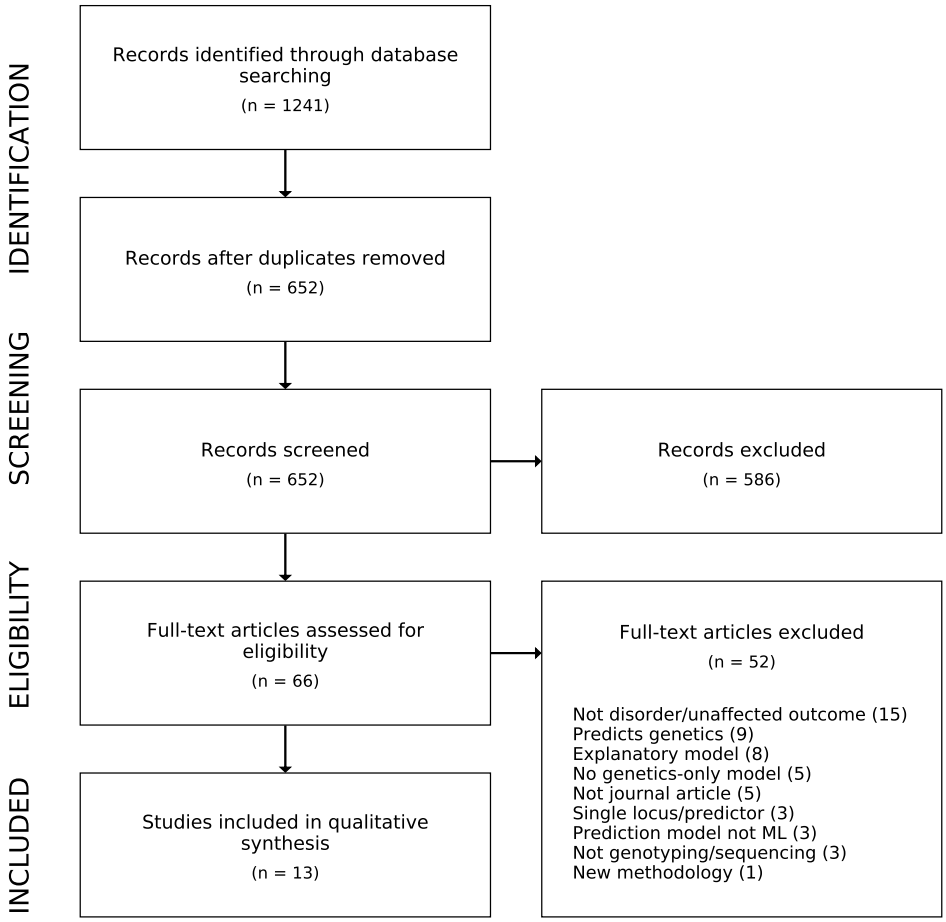


FIGURE 3.1: PRISMA flow diagram. Counts of publications are given for 'eligibility', 'screening' and 'identification', while counts of studies are used for 'included'. Two of the 14 selected publications were merged to give 13 studies for inclusion in the review.

3.3.2 Studies

A wide range of machine learning methods were applied to schizophrenia (7 studies, 47% of models), bipolar disorder (5 studies, 39% of models), autism (3 studies, 10% of models) and anorexia (1 study, 4% of models) (Table 3.1), with no studies identified for the 6 remaining disorders. Single nucleotide polymorphisms (SNPs) were the most common source of genetic data. Copy number variants (CNVs) and PRSs were each incorporated in models from a single study, and exome-sequencing data formed the basis of two studies. Datasets typically consisted of publicly-available GWAS; potential sample overlap was established for at least 7 studies (Table A.4). Briefly, 3 studies (Pirooznia et al., 2012; Guo et al., 2015; Vivian-Griffiths et al., 2019) included controls for the 1958 Birth Cohort (Power and Elliott, 2006) or the UK Blood Service (Consortium et al., 2007), 4 studies included controls from Knowledge Networks (Pirooznia et al., 2012; Li et al., 2014; Acikel et al., 2016; Chen et al., 2018), 2

studies used a Swedish population-based sample (Chen et al., 2018; Trakadis et al., 2019), and 3 studies used the same dataset, or provided a common reference for part of the dataset (Pirooznia et al., 2012; Li et al., 2014; Acikel et al., 2016). The remaining 6 studies (Aguiar-Pulido et al., 2010; Aguiar-Pulido et al., 2013; Yang et al., 2010a; Engchuan et al., 2015; Laksshman et al., 2017; Wang et al., 2018; Ghafouri-Fard et al., 2019) either gave unclear information, reported no previous reference for the dataset, or used datasets which appear to be separate from other studies. Where samples overlap, all models included in the review are distinct, using different predictors or modelling approaches. Additional overlap or cryptic relatedness may be present between studies.

Missingness was reported clearly in about half of all studies and models. When reported, it was most commonly handled by imputation after excluding genotypes with high missingness. Studies also reported complete-case analysis and inclusion of missing values in coding of predictors (Table A.5).

3.3.3 Machine learning methods

Support vector machines (SVMs) and neural networks were the most popular, followed by random forests and boosting (Table 3.2). SVMs were split roughly equally between using a linear kernel (3 studies, 7 models), a radial basis function (RBF) kernel (3 studies, 6 models), or an unreported kernel (3 studies, 6 models). Authors applying neural networks most commonly used multilayer perceptrons (3 studies, 6 models), an RBF network (2 studies, 5 models) or restricted Boltzmann machines (RBMs; 1 study, 9 models), with linear networks, convolutional neural networks (CNNs) and embedding layers each used once. Weak learners in boosted models were mainly decision trees, with the exception of a method which combined feature selection with the boosting of RBF-SVMs in AdaBoost (Yang et al., 2010a). Penalised regression was employed alongside linear and non-linear methods as least absolute shrinkage and selection operator (LASSO; 3 studies, 4 models) or ridge regression (1 study, 2 models).

Choice of model appears to be at least partly driven by availability: neural networks (RBF kernel), decision tables, BFTree and decision tree naïve bayes were only used in WEKA. 51% of all models were implemented in R or WEKA; Matlab and Python were preferred for neural networks (Table A.6).

3.3.4 Risk of bias

Risk of bias was assessed for each model within each study (Figure 3.2). All models displayed risk of bias, mostly in relation to participants (study design and inclusion/exclusion criteria), outcome (standardised definition and assessment of outcomes) and analysis. Within-study ROB for participants was due to the use of case-control studies. Predictors were mostly rated to have unclear or low ROB; instances of high ROB were limited to predictors which are

First Author (Year)	Disorder	Machine Learning Methods	Data	Models	Comparitors
Yang et al. (2010)	Schizophrenia	Adaboost (of SVM (RBF)), SVM (RBF)	SNPs	2	
Aguiar-Pulido et al. (2010; 2013)*	Schizophrenia	Adaboost, BFTree, DTNB, descision tables, SVM (kernel not reported), naïve Bayes, Bayesian networks, MDR, neural network (RBF, linear, perceptron), evolutionary computation	SNPs	12	
Pirooznia et al. (2012)	Bipolar Disorder	Bayesian networks, random forest, neural network (RBF), SVM (kernel not reported)	SNPs	16	PRS, LR
Li et al. (2014)	Bipolar Disorder, Schizophrenia	LASSO, Ridge, SVM (linear)	SNPs	6	
Engchuan et al. (2015)	Autism	Neural network (perceptron), SVM (linear), random forest, CIF	CNVs	4	
Acikel et al. (2016)	Bipolar Disorder	MDR, random forest, *k*-NN, naïve Bayes	SNPs	5	
Guo et al. (2016)	Anorexia nervosa	LASSO, SVM (RBF), GBM	SNPs	3	
Laksshman et al. (2017)	Bipolar Disorder	Decision tree, random forest, neural network (CNN)	Exomes	3	
Chen et al. (2018)	Schizophrenia	Neural network (perceptron)	PRS	4	PRS, LR
Wang et al. (2018)	Schizophrenia, Bipolar Disorder, Autism	Neural networks (cRBM)	SNPs, gene expression	9	LR
Ghafouri-Fard et al. (2019)	Autism	Neural network (with embedding layer)	SNPs	1	
Trakadis et al. (2019)	Schizophrenia	LASSO, random forest, SVM (kernal not reported), GBM (XGBoost)	Exomes	4	
Vivian-Griffiths et al. (2019)	Schizophrenia	SVM (linear, RBF)	SNPs	8	PRS

TABLE 3.1: Overview of studies. BFTree (best-first decision tree), CIF (conditional inference forest), cRBM (conditional restricted Boltzmann machine, CNN (convolutional neural network), DTNB (Decision table naïve Bayes), *k*-NN (*k*-nearest neighbours), LASSO (least absolute shrinkage and selection operator), LR (logistic regression, MDR (multifactor dimensionality reduction), RBF (radial basis function), SVM (support vector machine), PRS (polygenic risk score)). *Aguiar-Pulido et al., 2010 and Aguiar-Pulido et al., 2013 merged in extraction.

Machine Learning Methods	Studies	Models
SVMs	8	19 (25%)
Neural networks	7	23 (30%)
Random forests	5	8 (10%)
Boosting	4	4 (5%)
Penalised regression	3	6 (8%)
Bayesian networks	2	5 (6%)
Naïve Bayes	2	2 (3%)
Decision trees	2	2 (3%)
MDR	2	3 (4%)
Other	5	5 (6%)

TABLE 3.2: machine learning models. Boosting includes Adaboost and gradient boosting machines such as XGBoost. Boosting of RBF SVMs via Adaboost in (Yang et al., 2010a) is counted once under boosting. Other includes models seen only once: evolutionary computation, k-nearest neighbours (k-NN), conditional inference forests (CIF), decision tables and decision tree naïve Bayes (DTNB). Percentages are rounded to the nearest integer.

unavailable at the point of model use. Outcome definitions or measurements often differed between cases and controls.

Models displayed high ROB during analysis. This was often traced to inappropriate or unjustified handling of missingness and removal of enrolled participants prior to analysis, predictor selection using univariable methods and failure to account for overfitting. No studies reported calibration measures. In addition to PROBAST, information on population structure within studies was extracted (Table A.7). Most studies did not illustrate genetic ancestry across all observations in the current publication using dimensionality reduction, and none reported any evaluation of the final trained model for bias due to population structure. However, 2 studies (18% of models) visualised principal components for a subsample or showed a table of reported ancestry for participants (Acikel et al., 2016; Wang et al., 2018). Where ancestry was not addressed in a study, it was most often visualised in a referenced publication (55% of all models). 2 studies (13% of models) had no details or references which addressed genetic ancestry.

Across-study ROB was not formally assessed. For schizophrenia, bipolar and autism, studies with smaller numbers of cases in the development set report AUC less often, instead preferring classification metrics such as accuracy, sensitivity and specificity.

PROBAST encourages assessment of studies for applicability to the review question as this is often narrower than inclusion criteria (Wolff et al., 2019). Concern was identified for models in three studies (Pirooznia et al., 2012; Li et al., 2014; Wang et al., 2018). All others demonstrated either low concern or unclear applicability. Reasons for concern were attributable to outcomes which combined closely-related disorders, or the use of post-mortem gene expression data, whereas the review question focussed on models of single disorders with potential use in diagnosis or prognosis.

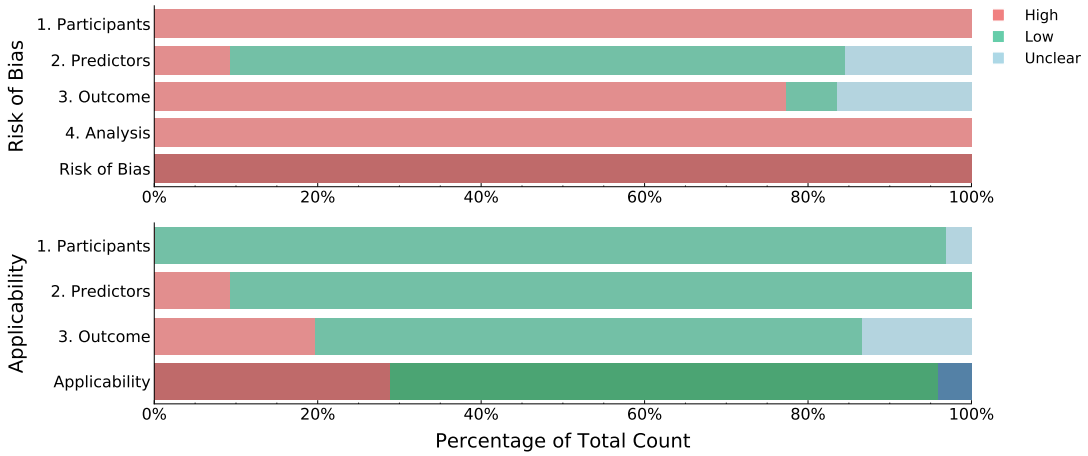


FIGURE 3.2: within-study risk of bias and applicability assessed by the prediction model risk of bias assessment tool (PROBAST). Colours indicate low, high or unclear risk of bias or applicability. Assessments were carried out for each validation of each prediction model in a study across 4 domains; the final rows give the overall assessment for risk of bias and applicability. Methodology is given further in sections 3.2.4 and A.1.2.

Methods	Studies	Models
Internal validation		
Cross-validation	8	44 (57%)
3-fold	1	4
4-fold	1	8
5-fold	2	8
10-fold	3	22
LOOCV	1	2
Split-sample	5	16 (21%)
34% train[1]	1	3
40% train[2]	1	3
70% train	1	4
80% train	1	2
90% train	1	4
Apparent validation	1	1 (1%)
Not reported[3]	1	16 (21%)
External validation		
External (temporal, geographic)[3]	1	16 (21%)
Partly external[4]	1	4 (5%)
Not performed	11	57 (74%)

Methods	Studies	Models

TABLE 3.3: validation. Percentages are given with respect to 77, the total number of models. Methodology for internal validation differed between models in a study (Purcell et al., 2014), which is counted in cross-validation (CV), split-sample and apparent. [1]Approximately equal three-way split between predictor selection, train and test, with 10-fold CV performed in the training fold for hyperparameter tuning. [2]40% train, 10% test, 50% final test. [3]No performance measures reported for internal validation, but discrimination for fully external validation reported (Daneshjou et al., 2017). [4]Control sample used in development and validation partially overlaps. LOOCV: Leave-one-out cross validation.

3.3.5 Model performance

Over half of all models assessed discrimination using AUC (58% models). A wide range of classification metrics and measures of model fit were also reported (Table A.8), with less than a quarter of models clearly reporting choosing a decision threshold *a priori* (Table A.9).

Around 79% of models, from 12 studies, reported some form of internal validation (Table 3.3). The majority of these were k-fold cross-validation (57% of all models; 8 studies), a resampling approach which involves testing a model on each of k independent partitions of a dataset, every time training on the remaining $k-1$ folds. 10-fold cross-validation (CV) was most commonly used, with just below half of all cross-validated models invoking repeats with different random splits. The remainder of studies using internal validation created a random split between training and testing sets (21% of all models; 5 studies), or applied apparent validation, where training and testing are both done on the whole sample (Acikel et al., 2016). A minority reported external validation (26% of models; 2 studies). Use of internal validation was not reported for 16 models from a single study (Pirooznia et al., 2012), but for which geographic and temporal external validation was given. External validation was reported for one other study, but with partly overlapping participants between development and validation sets (Chen et al., 2018).

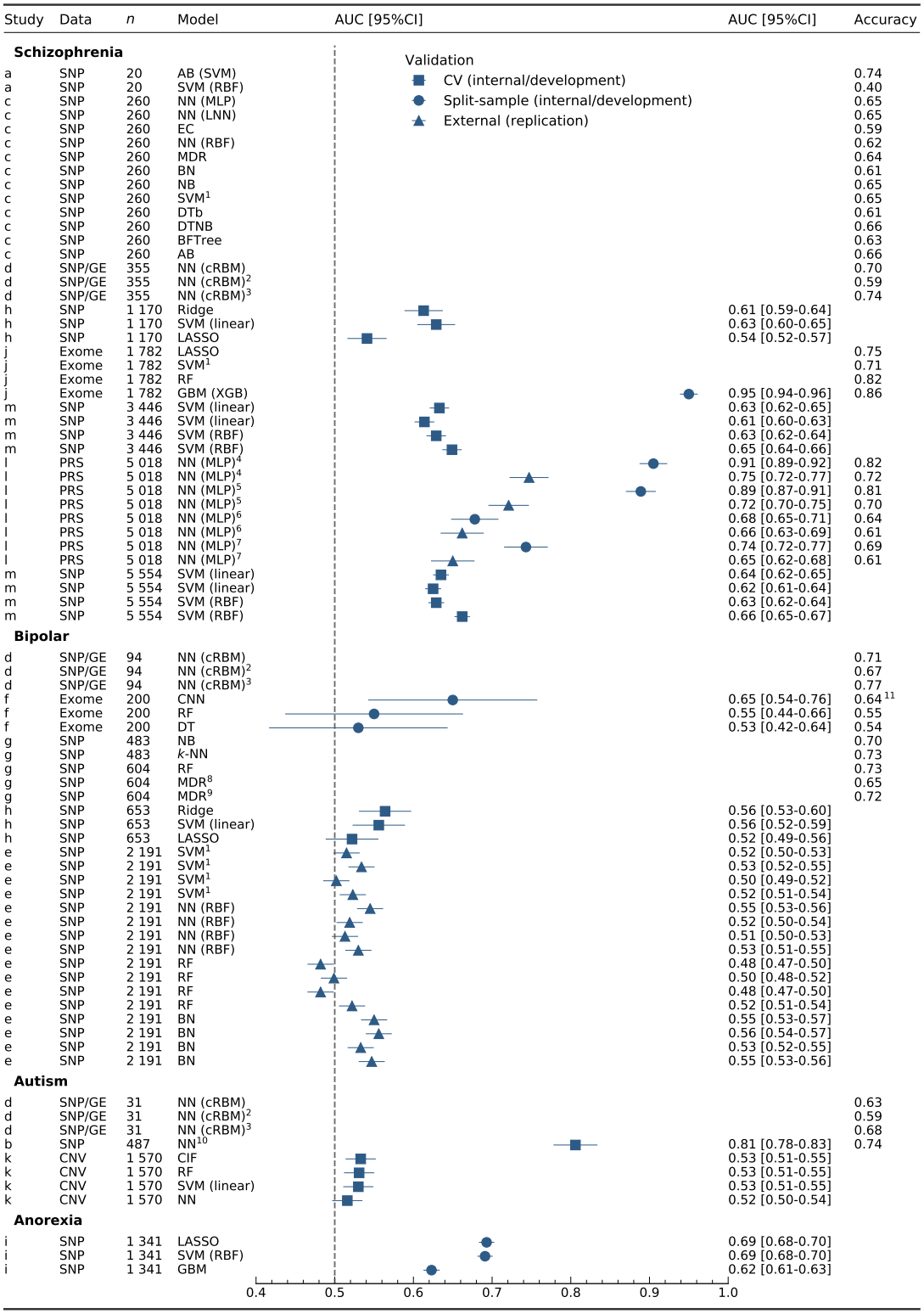


FIGURE 3.3: discrimination for all models. *n*: number of cases in training set. Studies: a (Yang et al., 2010a), b (Ghafouri-Fard et al., 2019), c (Aguiar-Pulido et al., 2010; Aguiar-Pulido et al., 2013), d (Wang et al., 2018), e (Pirooznia et al., 2012), f (Laksshman et al., 2017), g (Acikel et al., 2016), h (Li et al., 2014), i (Guo et al., 2015),

FIGURE 3.3: j (Trakadis et al., 2019), k (Engchuan et al., 2015), l (Chen et al., 2018), m (Vivian-Griffiths et al., 2019). [1]SVM kernel not reported. [2]Modified architecture with intermediate phenotypes in training set only. [3]Modified architecture with intermediate phenotypes for training and test sets. [4,5,6,7]Internal and external validation are shown for study l, where validations for the same model are denoted with the same number. [8]Two-way MDR. [9]Three-way MDR. [10]Neural network embedding layer. [11]Accuracy calculated from confusion matrix. AB: AdaBoost, BN: Bayesian networks, BFTree: best-first tree, CIF: conditional inference forest, cRBM: conditional restricted Boltzmann machine, CI: confidence interval, CNN: convolutional neural network, CNV: copy number variation, DTb: decision tables, DTNB: decision table naïve Bayes, DT: decision tree, EC: evolutionary computation, GE: gene expression, GBM: gradient boosting machine, *k*-NN: *k*-nearest neighbours, LASSO: least absolute shrinkage and selection operator, LNN: linear neural network, MDR: multifactor dimensionality reduction, MLP: multi-layer perceptron, NB: naïve Bayes, NN: neural network, PRS: polygenic risk scores, RBF: radial basis function, RF: random forests, SNP: single nucleotide polymorphisms, SVM: support vector machine, XGB: extreme gradient boosting.

Model performance varied by choice of statistical method, sample size and number of predictors within studies (Table A.10). Discrimination for models of schizophrenia (Figure 3.3) was extremely varied (0.54-0.95 AUC), with the highest AUC from exome data using XGBoost (0.95 AUC) (Trakadis et al., 2019). In this study, Trakadis et al. (2019) used counts of variants in each gene, after annotation and predictor selection, on participants with part-Finnish or Swedish ancestry (Purcell et al., 2014). Similarly high AUC (0.91 AUC) made use of multiple schizophrenia-associated PRS (Chen et al., 2018). However, the authors identify the presence of both the development and validation samples in the psychiatric genomics consortium (PGC) GWAS used to generate the schizophrenia PRS (Ripke et al., 2014), in addition to having overlapping controls between internal validation (model development) and external validation (replication) samples. All other schizophrenia models involved learning from SNPs (Yang et al., 2010a; Aguiar-Pulido et al., 2010; Aguiar-Pulido et al., 2013; Li et al., 2014; Vivian-Griffiths et al., 2019), with the exception of Wang et al., 2018 where gene expression data from post-mortem samples informed the weights in a conditional RBM trained on genotypes.

Predictive ability for bipolar disorder (Figure 3.3) was consistently lower than for schizophrenia, frequently overlapping with chance (0.48-0.65 AUC). Models were trained on genotypes, excepting a study (Laksshman et al., 2017) using exome data to train a CNN as part of the Critical Assessment of Genome Interpretation (CAGI) competition (Daneshjou et al., 2017), for which moderate discrimination was achieved (0.65 AUC).

Significantly fewer models were reported for autism (8 models, 3 studies) and anorexia (3 models, 1 study) (Figure 3.3). Varying predictive performance was illustrated in autism (0.52-0.81 AUC). High AUC (0.81 AUC) was shown for a single prediction model (Ghafouri-Fard et al., 2019), while models developed with a greater sample size by Engchuan et al. (2015) using CNVs were closer to or overlapping with chance (0.52-0.53 AUC) (Engchuan et al., 2015). The only models predicting anorexia nervosa had moderate discriminative ability between cases and controls (0.62-0.69 AUC) (Guo et al., 2015).

3.3.6 Logistic regression and polygenic risk scores

Three studies reported AUC for either logistic regression (5 models) or polygenic risk scores (12 models) alongside machine learning methods. PRS were weighted by summary statistics from a GWAS on the same disorder as the outcome and used as the sole predictor in a logistic regression model. Discrimination for ML is similar to logistic regression for two studies (Pirooznia et al., 2012; Chen et al., 2018) and worse than PRS for two studies (Pirooznia et al., 2012; Vivian-Griffiths et al., 2019). Where ML and LR perform better than PRS (Chen et al., 2018), multivariable models were built using multiple PRS for schizophrenia-associated traits as the predictors. Though discrimination shows some difference between model types, the number of studies for comparison is low and results are clustered by study and type of validation (Figure 3.4).

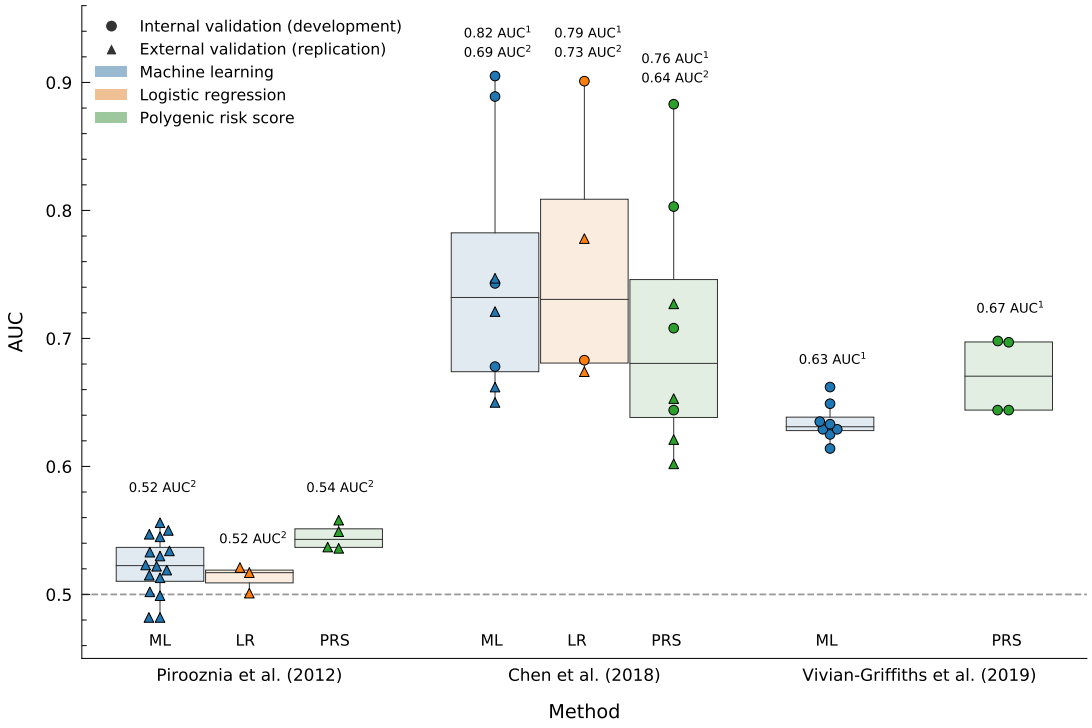


FIGURE 3.4: discrimination (AUC) for machine learning, logistic regression and polygenic risk scores. Internal validation (split-sample) and partly-external validation (with sample overlap) are reported for the same models in a single study (Chen et al., 2018). [1]Median AUC for internal validation (model development). [2]Median AUC for external validation (independent replication). Annotated scores are the median AUC for each model and study. Pirooznia et al. (bipolar disorder) and Vivian-Griffiths et al. (schizophrenia) show SNP-only models for LR and ML (Pirooznia et al., 2012; Vivian-Griffiths et al., 2019), while Chen et al. (schizophrenia) used multiple schizophrenia-associated trait polygenic risk scores as predictors (Chen et al., 2018). PRS model performance was extracted from a figure when unreported in-text (Vivian-Griffiths et al., 2019). AUC is shown only for 5 of the 9 reported logistic regression models; a fourth study compared ML and LR but did not report discrimination (Wang et al., 2018). AUC was not available for a logistic regression which was reported as attempted but not completed for one study (Pirooznia et al., 2012). AUC: area under the receiver operating characteristic curve, ML: machine learning, LR: logistic regression, PRS: polygenic risk scores.

3.3.7 Predictors

Coding of predictors was mostly unclear or unreported (7 studies, 55% of models). Coding was unclear if it was implied through the description of the type of classifier or software but not clearly articulated for the reported study. PRS were continuous (Chen et al., 2018) while counts of variants-per-gene or genes-per-gene-set were used for exomes and CNVs respectively (Trakadis et al., 2019; Engchuan et al., 2015). SNPs were coded under an additive model, a z-transformation of additive coding, or one-hot encoded (one predictor per genotype at a locus) (Table A.11). GWAS summary statistics from external datasets were also used in the selection, weighting or combining of predictors (9 studies, 64% models; Table A.12).

Predictor selection was adopted by most (12, 73% of models) and limited to filter-based selection, used prior to modelling, and embedded selection, an integral part of the prediction model (Table A.13). The latter involved LASSO regression, or ensembles and hybrids of decision trees and decision tables, in addition to a modified AdaBoost (Yang et al., 2010a). Filters were based on internal or external univariable association tests (GWAS). Embedded and wrapper-based methods, which typically 'wrap' a model in forward or backward-selection, were both also used prior to any predictive modelling. Modification of predictors using information from the test set was the most common cause of information 'leaking' from the test set to the training set, a source of inflation in performance measures (Table 3.4).

Leakage handled appropriately?	Studies	Models
Yes/Probably Yes	7	44 (57%)
No/Probably No	7	32 (42%)
Predictor selection performed prior to cross-validation	2	7
Predictor transformed prior to cross-validation[1]	4	22
Prior knowledge in predictors generated from test set	1	4
DEV and VAL sets overlap	1	4
HP chosen by test-set/split performance	4	22
GRN from whole dataset used to set NN architecture	1	6
Unclear[2]	1	1 (1%)

TABLE 3.4: handling of information "leaks" during training. Where studies have multiple reasons for suspected leakage, each of these is counted separately. If predictors were reduced to a set number before cross-validation was described, or a transformation was not reported as having been done within a pipeline or for each fold of cross-validation, this is recorded as 'probably no'. [1]Transform includes anything that summarises information from the test set, such as the mean of the whole sample in a z-transformation. [2]Predictor handling implied, as scikit-learn is listed for pre-processing and preparation, but no pre-processing steps are given (Ghafouri-Fard et al., 2019). DEV: development, VAL: validation, HP: hyperparameter, GRN: gene regulatory network, NN: neural network.

3.3.8 Sample size

Total sample size was generally low where a single sample had been used, but higher if genotypes from publicly-available amalgamated datasets used in a GWAS had been downloaded (median 3486, range 40-11853) (Table A.10). Number of events in development followed a similar pattern (median 1341, range 20-5554) as class imbalance was minimal (median 1, range 0.65-2.93, calculated as non-events over events). Around half of studies gave sufficient information to calculate events per variable (EPV) (median 0.69, range 0.00063-74.6). It could not be calculated where the number of candidate predictors where not reported for models in 2 studies (Pirooznia et al., 2012; Wang et al., 2018); approximations are given in appendix A where reporting was unclear in a further 5 studies (Guo et al., 2015; Chen et al., 2018; Trakadis et al., 2019; Aguiar-Pulido et al., 2010; Aguiar-Pulido et al., 2013; Laksshman et al., 2017) (Table A.10).

3.3.9 Hyperparameter search

Hyperparameter search was mostly unreported or unclear (41 models, 9 studies), with some models reported as having been used with default settings. Ambiguous reporting resulted from description of search and tuning for a specific model, with no clarity as to whether these conditions applied to other models in the study. Only 19% of models clearly reported attempting different hyperparameters for the extracted models (Table 3.5). Studies also report non-standard final hyperparameters, such as uneven batch size in neural networks, or showed good accuracy for a model which is highly sensitive to tuning of crucial hyperparameters, yet few reported tuning (Table A.14). It is therefore likely that most studies evaluated several hyperparameter choices but did not report this.

Search method for hyperparameters	Studies	Models
Search method reported	4	15 (19%)
Grid	1	1
Random	1	8
Manual	2	12
Bias variance decomposition	1	2
Default hyperparameters	1	16 (21%)
Search method unclear/unreported	9	46 (60%)
Not clearly reported[1]	2	8
Not reported	7	38

TABLE 3.5: hyperparameter search technique. [1]Methods reported clearly for other models in publications, but not made clear that the same methods apply to extracted models. One publication (Vivian-Griffiths et al., 2019) used both manual and random elements for search, and is counted in both categories. Manual tuning by Chen et al. (2019) is implied through

TABLE 3.5: reported values which were attempted for hyperparameters, but not explicitly stated (Chen et al., 2018). Hyperparameters searched systematically using a given set of values are denoted as grid search. If authors report attempting various hyperparameter choices but give no indication of systematic search or value choices, this is recorded as manual. Two studies (12 models) reported hyperparameters that were tuned but gave no indication of how this was done (Laksshman et al., 2017; Wang et al., 2018). A study (1 model) reported search methodology, but not what hyperparameters were tuned (Yang et al., 2010a).

3.4 Discussion

3.4.1 Predictive ability in schizophrenia

All studies displayed high risk of bias in model development and validation with infrequent reporting of standard modelling steps. Performance measures consequently demonstrated a wide range of abilities to discriminate between cases and controls (0.48-0.95 AUC) (Figure 3.3). These are likely optimistic owing to the high risk of bias identified through PROBAST and unaddressed sample overlap and population structure, as two studies showing the highest AUCs left these issues unresolved (Chen et al., 2018; Trakadis et al., 2019). Interpretation is further muddied by low effective sample size: 13 studies studies were included, where at least half had clear sample overlap, and not all studies reported a measure of discrimination. Ideally, use of a ROB assessment tool enables identification and meta-analysis of a subgroup with little or no suspected inflation of performance measures. This was not possible, as all studies were identified as high ROB. For schizophrenia, removing the studies where inflation of performance measures was most clear (Chen et al., 2018; Trakadis et al., 2019) reduces the range of discrimination measures from 0.54-0.95 AUC to 0.54-0.66 AUC. While this may give a more feasible range for predictive ability for schizophrenia, it is still highly likely to be inflated, and is based on just two studies (Li et al., 2014; Vivian-Griffiths et al., 2019).

Relating discrimination for schizophrenia and other psychiatric disorders to machine learning efforts in genetics more generally is difficult. Very few systematic reviews exist, and assessment of risk of bias, which is critical for interpreting machine learning and other predictive models, is also uncommon. Beyond psychiatry, however, broad discrimination has also been observed for machine learning studies in cancer genomics (Patil et al., 2019), though this constituted just 7 studies and did not include an assessment of ROB. More established fields with clearer predictor-response relationships, such as medical imaging, have much more consistent values for measures of predictive ability across studies (for instance (Islam et al., 2020)).

3.4.2 Comparison with logistic regression and polygenic risk scores

Similar factors constrain comparison of machine learning methods to logistic regression and polygenic risk scores (Figure 3.4). Within each study, machine learning appears to show similar performance to logistic regression, in line with previous findings (Christodoulou et

al., 2019). As machine learning approaches require tuning of hyperparameters, a common source of inflation in performance measures, such a comparison may give an overly-favourable impression of ML. Where PRS appears to give better discrimination than ML, this may be due to the value of incorporating external weights from much larger sample sizes or the assumption of additivity within and between loci.

3.4.3 Within-study risk of bias

Despite difficulty with interpretation of predictive performance, low standards of model development, validation and reporting are a clear and consistent theme throughout all studies. Issues relating to ROB often rest on distinctions in methodology between clinical prediction modelling, machine learning and genetic association studies. For instance, genetic studies most commonly employ a case-control design. Such studies are extremely useful for identifying genetic risk factors for rare outcomes, but are considered inadequate for prediction modelling as absolute risks cannot be estimated; instead, case-cohort, nested case-control, or prospective cohort designs are preferred (Moons et al., 2012). Case-cohort and nested case-control designs involve sampling from an existing cohort and can be used for prediction models if the sampling fraction in controls is accounted for in analysis (Biesheuvel et al., 2008). To project the prediction to the whole population in case-control studies, positive and negative predictive values should be corrected in accordance with the disease prevalence in the population and ratio of cases and controls in the sample (Kallner, 2018). Similarly, univariable tests of association are applied routinely in GWAS, and are often used in selection of predictors for genetic prediction models. Their application in prediction modelling though is usually discouraged, as predictors may differ in their importance when evaluated in isolation as compared to when considered concurrently with other variables (Sun, Shook, and Kay, 1996).

Lack of adherence to appropriate procedures for machine learning are also a common cause of a model being assessed as at high risk of bias. Standard model validation procedures were followed by some researchers; however, many 'leaked' information between training and testing sets by not restricting predictor manipulation and selection to the training set/fold, or by using the testing set/fold to adjust model hyperparameters, which can impose significant bias on estimates of prediction performance (Vabalas et al., 2019). Most studies provided a measure of classification or discrimination for each model; none reported a measure of calibration. Model calibration compares observed and predicted probabilities of the outcome occurring, and is a crucial part of model development (Steyerberg et al., 2019) which has been noted for its absence in genetic prediction literature (Janssens et al., 2011). Authors reporting only classification measures, such as accuracy, sensitivity or specificity, should also note that measures of discrimination are preferred as they use all the information over predicted probabilities and delay any thresholding of risks to a more appropriate time. Of discrimination measures, the AUC is the most widely used in both machine learning and genetics (Bradley, 1997; Wray et al., 2010).

Hyperparameter optimisation is an essential part of developing machine learning models as it determines how they navigate the bias-variance trade-off and learn from data (James et al., 2013). It is therefore surprising that it was so often unreported or subject to a small number of manual experiments. Hyperparameters should be systematically searched to ensure a model is not over or under-fit. Randomised search has been shown to be more effective than grid search where two or more such parameters require tuning (Bergstra and Bengio, 2012), though grid search is also recommended by practitioners for SVMs, often with an initial 'coarse' search followed by a more thorough exploration of a finer grid of values (Hsu, Chang, Lin, et al., 2003; Ben-Hur and Weston, 2010). The importance of search is particularly relevant in domains where there are a small number of events per candidate predictor (Pavlou et al., 2015), such as genomics, as appropriate hyperparameter choices can reduce overfitting.

Split-sample approaches were used by several studies, but should be avoided in favour of resampling methods such as bootstrapping or k-fold cross-validation (Steyerberg et al., 2001). The latter is an appropriate form of internal validation for traditional statistical methods; however, estimated prediction accuracies become overly-optimistic if done repeatedly, as when used for hyperparameter tuning through repeated rounds of CV. Nested cross-validation, where hyperparameters are optimised in an inner-fold and evaluated in the outer-fold, has been shown to give more realistic estimates (Varma and Simon, 2006; Vabalas et al., 2019) but was not used in any studies. A single study presented both internal and external validation of models (Chen et al., 2018), for which a large drop in performance is seen upon replication. Though partly due to sample overlap between the development set and the summary statistics used for generating a PRS, difficulty with replication is a wider issue in polygenic risk prediction. Risk scores for psychiatric disorders typically explain a small proportion of variance in a trait (Lee et al., 2013), with generalisation issues compounded by variants with small effect sizes and different allele frequencies between populations. Risk scores generated through machine learning methods have the potential to be more affected by these issues if appropriate modelling procedures are not followed.

3.4.4 Population structure

A source of bias not explicitly covered in PROBAST is population structure. Genetic ancestry has the potential to bias both associations (Marchini et al., 2004; Price et al., 2006) and predictions (Belgard et al., 2014; Martin et al., 2017) from genetic data. Supervised machine learning methods have proved particularly sensitive in detecting aspects of population structure (Bridges et al., 2011; Schrider and Kern, 2018; Flagel, Brandvain, and Schrider, 2019). Few researchers discussed visualising ancestry or reported exclusions, and none reported modelling adjustments, even when previous association studies on the same datasets had demonstrated stratification and included principal components as covariates. The extent of the bias introduced in these studies is not clear: evidence mostly relates to deliberately predicting populations in humans using ML or looking at bias in complex trait

prediction from PRS. Furthermore, work using neural networks has shown equivalent performance in genetic prediction of populations when non-linear activation functions are removed and networks are restricted to learning only a linear model (Bridges et al., 2011). It may therefore be that the supervised nature of the learning problem drives stronger prediction of populations, rather than that flexible ML methods are better at detecting population structure, as the outcome or as a confounder of disease-status.

Despite this, the potential for population stratification to impact prediction in general is apparent, though the method for dealing with it when using machine learning methods is not. Several techniques have been proposed, including modifications to random forests (Stephan, Stegle, and Beyer, 2015); exclusions by, or inclusion of, principal components; and regressing-off the linear effects of principal components on SNPs before modelling (for example (Zhao et al., 2012; Zheutlin et al., 2018)). Whether any combination of these is sufficient to reduce the effects of population stratification in non-linear machine learning predictions has not been demonstrated.

3.4.5 Reporting

General reporting guidelines for machine learning prediction models are yet to be developed (Collins and Moons, 2019), though recommendations for undertaking (Teschendorff, 2019; Boulesteix et al., 2020) evaluating (Tandon and Tandon, 2019) or reporting (Luo et al., 2016) exist for machine learning in omics data, psychiatry and medicine respectively, in addition to reporting guidelines outside of machine learning (Cecile et al., 2011; Collins et al., 2015). Researchers applying machine learning approaches to genetic prediction should consult guidelines on best practices in model development, validation and reporting. Furthermore, methods should be reported alongside polygenic risk scores and a linear model, such as logistic regression, as standard baseline models for comparison.

3.5 Conclusion

The ability of machine learning methods to predict schizophrenia or other psychiatric disorders from genetics remains unclear. Attributes of studies which elevated risk of bias for analysis often relate to information leaking from the test set to the training set. Furthermore, comparison between machine learning, logistic regression and polygenic risk scores is hampered by low effective sample size. These limitations can be dealt with adequately by considering simulations. Here, for any given population parameters, a large external sample can be drawn after any models are developed on the training set, avoiding any possibility of information leaking. In addition, additivity of genetic effects, and deviations from this, can be investigated alongside polygenic risk scores with and without prior information. For these reasons, the next chapter turns to simulations.

Chapter 4

Simulation Study of Binary Classification of Complex Traits

4.1 Introduction

Machine learning (ML) methods are increasingly used in prediction modelling, with interest growing in psychiatric genetics. A systematic review of machine learning learning approaches found widespread risk of bias in analysis of models and a lack of comparison to polygenic risk scores (PRS) and logistic regression (LR). The discriminative ability of machine learning models for genetic data in psychiatric disorders remains unclear. Simulations allow for simplification of the intricate genetic landscape of complex traits, the ability to assess a wide array of potential scenarios, and separation of train and test data to avoid information leakage. Here they are applied to models of additive effects and interactions under a range of simulation parameters.

4.1.1 Main and interaction effects

The terms "epistasis" and "interaction effects" are both used for situations involving an interaction between genes. Epistasis was defined separately and differently by William Bateson (Bateson, 1909), referring to the masking of the effects of one locus by another (Figure 4.2 A), and Ronald Fisher (Fisher, 1918), describing a phenotype that deviates from the addition of the effects of two loci as exhibiting "epistacy". Fisher's definition is more expansive, and includes the situation described by Bateson and many others (Phillips, 1998). These uses of epistasis also differ in that the phenotype Fisher referred to was a quantitative trait or metric character, while Bateson described a generalisation of dominance at a single-locus to dominance between two loci for a Mendelian trait.

The definition of epistasis given by Fisher and used in quantitative genetics is taken forward here; however, the term "interaction" is used instead to avoid confusion, and refers to a deviation from a linear model that describes the effects of two or more loci on a phenotype. Absence of interaction for a binary trait on the penetrance scale is given by the model

$$p_{ij} = \alpha_i + \beta_j, \tag{4.1}$$

where α_i and β_j are the effects of genotype i at locus A and genotype j at locus B respectively, and p_{ij} is the penetrance (Cordell, 2002).

The effect of genotypes at a single locus may also differ from additivity if the effect of the heterozygote is not equal to the mean of the effects of the two homozygotes. This phenomenon, referred to a 'dominance deviation', violates linearity assumptions of typical logistic regression models. In extreme cases, the effect of the heterozygote may be greater than or less than the range of the effects of the homozygotes, termed 'overdominance' and 'underdominance' respectively. Models with monotonicity constraints, such as logistic regression, will be unable to correctly estimate such a function. The terms dominance deviation and interaction effects therefore describe within-SNP and between-SNP deviations from linearity, respectively.

A distinction is made between interaction effects and a "main effect", which is the effect of a single locus averaged over all other loci, regardless of whether it interacts with them or not (Frankel and Schork, 1996). Similarly, the term "marginal effects" is also used to refer to the effect of a genotype at one locus averaged across all other genotypes at a second locus when describing two-locus interaction models (Cordell, 2009), and the terms are used interchangeably here. The presence of interaction effects and use of the term here does not imply any underlying physical interaction between biological components, nor even that the effect is of any biological interest (Cordell, 2002).

		Genotype at locus B			
		bb	*bB*	*BB*	Marginal Effect
Genotype at locus A	*AA*	1	0	0	0.25
	aA	0	0.5	0	0.25
	aa	0	0	1	0.25
Marginal Effect		0.25	0.25	0.25	

FIGURE 4.1: A two-locus interaction model showing epistasis with no marginal effects, adapted from Frankel and Schork, 1996. Genotypic proportions are 0.25 and 0.5 for all homozygotes and heterozygotes respectively. The marginal effect for a genotype at one locus is the sum of the penetrance times the frequency across genotypes at the second locus.

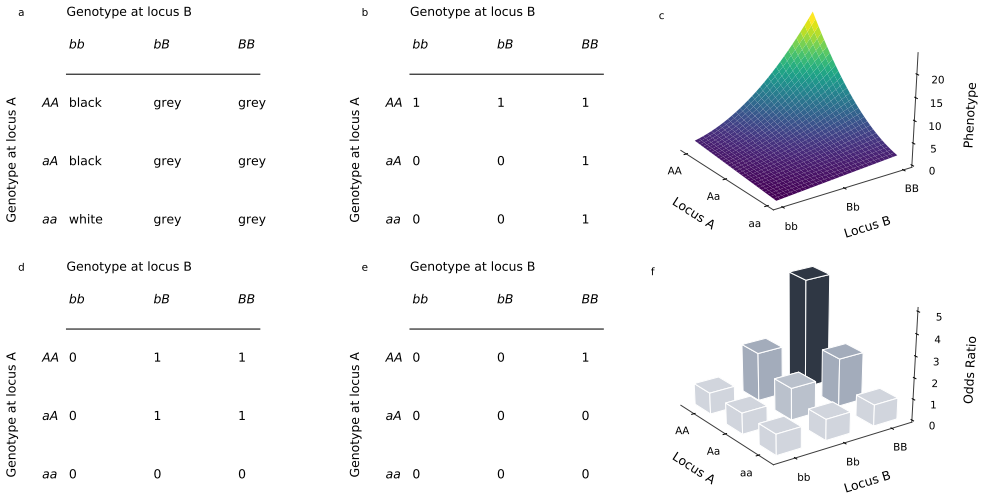


FIGURE 4.2: The many faces of epistasis. A shows Bateson's formulation of epistasis as a masking effect, where the presence of the B allele causes grey coat colour, and prevents us from seeing the white or black effect of the alleles at locus A. B and D show a heterogeneity model and a general model of epistasis (Cordell, 2002). The symmetry in D means we cannot tell which locus is "masking" which. Though B could be two loci acting independently through different mechanisms, when viewed as a recessive model it can be labelled as epistasis. E also falls under epistasis, showing a recessive effect between loci (Neuman, Rice, and Chakravarti, 1992). A, B, D and E show penetrance tables, where interactions exhibit complete penetrance, while C and F show multiplicative effects for two loci on a quantitative trait and a binary outcome with incomplete penetrance.

4.1.2 Models of two-locus interactions

Many interaction models have been described (Figure 4.2, Table 4.1). Although they are typically defined in tables on the penetrance scale (for example, Li and Reich, 2000), modern genetics studies are more commonly carried out as case-control studies, in which the main effects of a locus, and possible interaction models, are defined on the odds or log(odds) scale. The effects of two independent loci on the odds scale can be written as:

$$\frac{p_{ij}}{1-p_{ij}} = \alpha_i + \beta_j, \tag{4.2}$$

Departure from this additive model, for example as $\alpha_i + \beta_j + \alpha_i\beta_j$, implies statistical interaction. It is important to note, however, that this only suggests interaction on the odds scale; an additive effect defined on the odds scale is not additive on the penetrance scale, where interaction terms would be needed to describe the effect.

For many interaction models it is possible to devise a transformation which renders the relationship additive between loci. Such situations are included here as interaction models, provided they deviate from a linear model on the odds scale.

512 pair-wise SNP interactions for fully-penetrant binary phenotypes have been enumerated (Li and Reich, 2000), ranging from simple multiplicative interactions to complex exclusive-or

(XOR) models. Though described on the penetrance scale , these can easily be transported to an odds scale, and other models have been described on the odds scale directly (Figure 4.3).

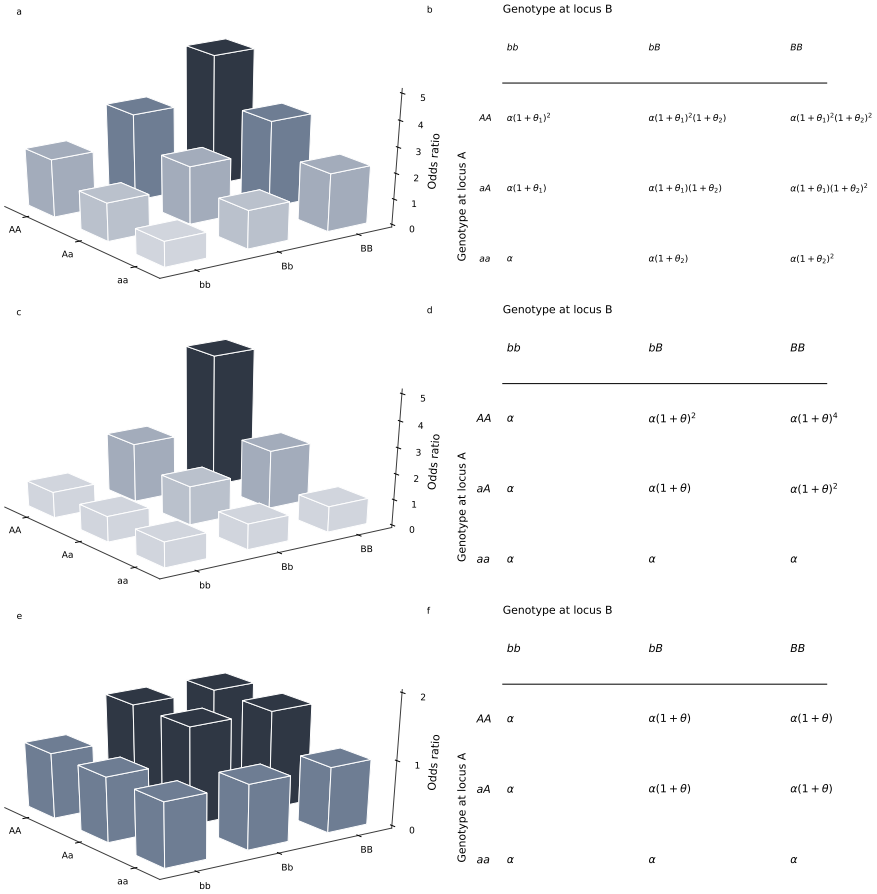


FIGURE 4.3: Interaction models defined on the Odds Scale, reproduced from Marchini, Donnelly, and Cardon, 2005. The top row shows an odds table, with the baseline effect α and interaction effect θ. A shows multiplicative odds within and between loci, and B requires at least one risk allele at both loci to show an effect. C is analogous to the fully-penetrant model in Figure 4.2, where risk hits some threshold and does not increase further. Multiplicity between loci, as shown in B, is taken forward as the multiplicative model.

4.1.3 Detection of interactions using machine learning

Many methods have been evaluated for detection of interaction effects using ML (Cordell, 2009; Koo et al., 2013; Niel et al., 2015; Uppu, Krishna, and Gopalan, 2016b). Random forests (RFs) have received particular attention. Modification of RFs to improve detection of interactions has been investigated (Pan et al., 2013; Yoshida and Koike, 2011; Li et al., 2016) along with their use upstream of statisical tests (Wei and Lu, 2014), though studies typically focus on the ability of variable importance measures (VIMs) in RFs, or novel variations of them, to screen for interactions (Lunetta et al., 2004; Wright, Ziegler, and König, 2016). RFs have also been adapted for high-dimensional data (Schwarz, König, and Ziegler, 2010) and applied for screening in rheumatoid arthritis (Liu, Ackerman, and Carulli, 2011) and

age-related macular degeneration (Jiang et al., 2009). Gradient boosting has been used in a pipeline to detect interacting SNPs (Behravan et al., 2018), and to detect interacting SNPs in schizophrenia (Andreasen et al., 2012), but has otherwise received much less attention for detection of interactions than random forests.

Support vector machines (SVMs) have been used to uncover interactions through combination with multifactor dimensionality reduction (MDR) (Fang and Chiu, 2012)) and were applied via L_1-penalised SVMs in Parkinson's disease (Shen, Liu, and Ott, 2010). They have also been assessed for detection of interactions in prostate cancer (Chen et al., 2008) and alongside neural networks to select interacting SNPs in simulations (Matchenko-Shimko and Dube, 2007). Deep learning has been assessed alone for prediction from genetic interactions in asthma (Tomita et al., 2004), and more recently for learning 2-SNP interaction models with reference to cancer genetics using multilayer feedforward networks (Uppu, Krishna, and Gopalan, 2016a), where the authors report a neural network with 3 hidden layers which outperforms logistic regression, random forests, naïve Bayes and gradient boosting machines when assessed by accuracy and area under the receiver operator characteristic curve (AUC). In addition, neural networks trained on 2-SNP interaction models have been found to give improved performance over logistic regression and MDR (Günther, Wawro, and Bammann, 2009), with optimisation of networks using genetic programming suggested to improve detection of interactions (Ritchie et al., 2003; Motsinger et al., 2006; Ritchie et al., 2007; Motsinger-Reif et al., 2008)

Machine learning approaches have often been evaluated on genetic simulations in order to assess a novel method; such studies may provide an optimistic view of predictive performance (Boulesteix, Lauer, and Eugster, 2013). Assessment of ML in psychiatric genetics has been less frequent. Comparison of ridge, least absolute shrinkage and selection operator (LASSO) and bivariate ridge regression was assessed in simulations of two binary outcomes which share a proportion of causal SNPs to inform results from real data on bivariate modelling of schizophrenia and bipolar disorder (Li et al., 2014).

While aspects of interactions and main effects have been considered for several ML models in different studies, a systematic assessment of supervised machine learning approaches with main and additive effects, along with the impact of unassociated SNPs, heritability, prevalence, effect size, minor allele frequency (MAF) and linkage disequilibrium (LD) has not been conducted. Furthermore, studies often do not overtly aim to maximise prediction through incorporation of interactions, but instead take an explanatory modelling approach in introducing new methods or comparing existing ones for detecting the presence of interactions (for example (Chatelain et al., 2018)), rather than predicting from them; here the focus is solely on prediction. In addition, no study has evaluated polygenic risk scores and logistic regression against machine learning classifiers on a range of 2-SNP interaction models for their effect on the decision boundary.

4.1.4 Aims and objectives

The aim of this chapter is to establish a baseline of what performance to expect for algorithms trained on SNPs where either main or interaction effects are present, and understand the effect of varying simulation parameters on predictive performance of linear models and flexible machine learning approaches. The objectives are:

- Evaluate which approaches are preferred for high or low polygenicity
- Compare performance of classifiers under $p > n$ and $p < n$ scenarios
- Understand the ability of linear and flexible approaches to fit pairwise interaction models
- Asses the effects of including SNPs which are unassociated, have varying minor allele frequency (MAF) or are in linkage disequilibrium (LD) with the causal variant

4.2 Methods

4.2.1 Main effects

4.2.1.1 Additive simulations

Additive simulations were created by combining the effects of SNPs on a phenotype through an additive model on the liability scale (Falconer and Mackay, 1996). This is the same underlying model of polygenic risk scores, that all allelic effects are additive within and between loci, and as such polygenic risk scores are expected to perform best. To allow for training of multiple machine learning models using cross-validation, and the evaluation of a range of parameters, simulations were restricted to 25 repeats. Additive simulations were performed with 100, 200, 500 1,000 or 2,000 SNPs and sample sizes of 500, 1,000, 1,500 or 2,000 with equal numbers of cases and controls. For each simulation, genotypes were assigned under Hardy-Weinberg equilibrium (HWE) using MAFs drawn from a uniform distribution between 0.05 and 0.5 for 500,000 observations.

Each causal variant was randomly assigned an effect size, u_j, drawn from a standard normal distribution, such that $u_j \sim \mathcal{N}(\mu, \sigma^2)$, using $\mu = 0$ and $\sigma = 1$, and $G_{ij} \in \{0, 1, 2\}$ for any sample i and causal SNP j, and $G \in \mathbb{R}^{n \times p}$ for n observations and p predictors. The dot product of the samples-by-genotypes matrix G and the effect sizes u gives the genotypic values of individuals, g, which is normally distributed.

The narrow-sense heritability of a trait, given by

$$h_2 = \frac{var(g)}{var(d)}, \tag{4.3}$$

is set, where d is the outcome on the liability scale, and environmental deviation, e, is drawn from a normal distribution with variance

$$var(g)(\frac{1}{h_2} - 1), \tag{4.4}$$

and mean 0, following (Yang et al., 2010b). The addition of an individual's genotypic value, g, and environmental deviation, e, gives their phenotypic value on the liability scale, d (Falconer and Mackay, 1996), i.e. $d = g + e$.

The result of this is a range of genotypic effects which together bestow some value on individuals, and environmental effects which cause a deviation from the genotypic values with a variance that gives the desired narrow-sense heritability of the binary trait on the liability scale (Falconer and Mackay, 1996). Narrow-sense heritability on the liability scale was fixed to 0.2 for most simulations, but is also shown for 0.15, 0.25 and 0.3. Proportion of genotypes which are causal, m, was varied for $p < n$, where $m \in \{0.05, 0.25, 0.5, 0.75, 1\}$ and $p > n$ scenarios, where $m \in \{0.05, 0.1, 0.25, 0.5\}$.

A threshold is applied to the phenotypic value on the liability scale to give a binary outcome. The prevalence of the population, K, was fixed to 0.0025 for the majority of simulations, but also evaluated at 0.0025, 0.005, 0.01 or 0.02. Phenotypic values on the liability scale, d, were standardised to have mean 0 and unit variance, with K used to obtain the threshold from the inverse cumulative distribution function; this was achieved with SciPy's percentage point function in 'stat.norm' to produce thresholds of around 2.81, 2.58, 2.33 and 2.05 for a prevalence of 0.0025, 0.005, 0.01 and 0.02 respectively. To mimic ascertainment in case-control studies and obtain a balanced sample, both classes were undersampled to bring the prevalence of the sample to 0.5; cases could not be oversampled, as is typical, as this would create overlap between training samples in cross-validation. For each simulation, a test set of 2.5 million of observations was created to ensure sufficient cases could be subsampled for the lowest prevalence; this was subsampled to 5,000 cases and 5,000 controls. The test set was created with a separate random seed independently of any training simulations.

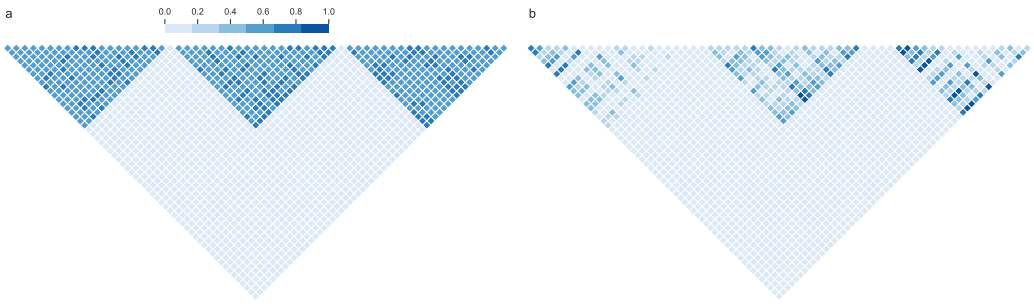


FIGURE 4.4: 20 SNP LD blocks. LD was either kept constant, shown left with $r^2 = 0.8$, or varied by drawing r^2 values uniformly between 0 and 1 (right). All causal SNPs were replaced by LD blocks.

The proportion of SNPs which are causal, m, or unassociated was varied, with unassociated SNPs drawn from a binomial distribution with two trials and chance of success (MAF)

drawn from a uniform distribution between 0.05 and 0.5. SNPs in LD with the causal SNPs were created in block structures using fixed or varying LD (Figure 4.4). LD was simulated by shuffling a proportion of observations for the LD-SNP to give the desired r^2 with the causal variant. Fixed LD was assigned a single r^2 value for all LD-SNPs to give homogenous LD blocks (Figure 4.4a). Varying LD simulations took r^2 values drawn from a uniform distribution between 0 and 1 to give a more complex structure closer to that observed for real data (Figure 4.4b). In all simulations the causal SNP was replaced by the variants in LD. Blocks were set at 20 SNPs. LD was considered for the $p > n$ scenario, where $p = 1,000$ and $n = 500$, and LD SNPs replaced all variants in the dataset, including those unassociated with the outcome. As p was also fixed, where the proportion of causal variants m was limited to 5% of SNPs, replacing these by 20-SNP LD blocks kept the original value for p. However, when $m > 0.05$ the dimensions of the LD-SNPs exceeded p. To keep p consistent, LD-SNPs were randomly subsampled down to give a sparse LD structure. LD simulations, where m is 0.05, 0.1, 0.25 or 0.5 therefore give contrasting polygenicity and LD structure (Figure 4.5).

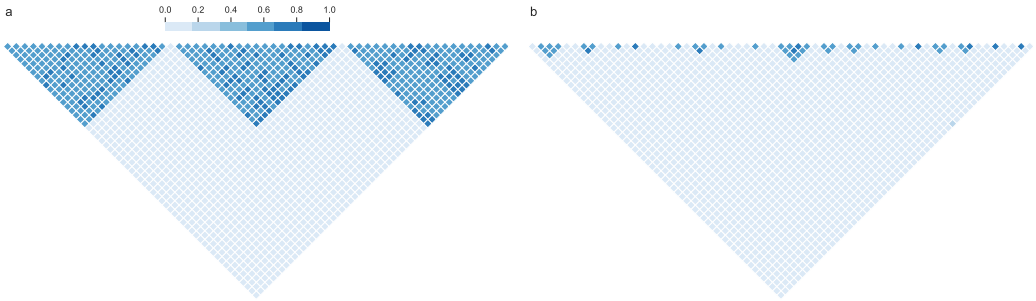


FIGURE 4.5: LD structures vary with the number of causal variants, m. For fixed LD block size of 20, $n = 500$ and $p = 1,000$, m was taken as 0.05, 0.1, 0.25 or 0.5. For $m = 0.05$, the dimensions of the LD-dataset equal the original dataset, so predictors show strong LD structures between a few causal loci (left). For $m = 0.5$, p increases to 10,000 when replacing causal variants with 20 SNP LD blocks. Predictors are then randomly subsampled to the original dimensions to give a more sparse LD structure. Code for simulations and plotting is given at https://github.com/seafloor/simulations.

4.2.1.2 Independent simulations

The simulations in section 4.2.1.1 combine effects across loci in an additive manner to create a genotypic value for each individual, assigning case-control status from the combination of this and noise. An alternative is to assign case-control status beforehand and simulate genotypes separately. This method is labelled 'independent' simulations to distinguish it from the 'additive' simulations described in the previous section, though both describe additive effects.

Genotyping data was simulated for independent SNPs with main effects only according to HWE. Genotypes were coded as number of risk alleles, $G_{ij} \in \{0, 1, 2\}$, for any individual i and SNP j. MAF for each SNP were drawn from a uniform distribution with range 0.05 to 0.5, as for additive simulations. Genotypes were then assigned to that SNP for controls

using HWE proportions, and MAF and genotypes in cases were calculated according to odds ratios (OR) chosen for each dataset. Odds ratios are established based on two-by-two tables, where presence of a risk allele is considered 'exposed'.

Number of samples (500, 1,000, 1,500 or 2,000) and SNPs (100, 200, 500, 1,000 or 2,000) were varied in each dataset as done for additive simulations, along with the proportion of SNPs which were associated with the outcome, $m \in \{0.05, 0.25, 0.5, 0.75, 1\}$ for $p < n$ and $m \in \{0.05, 0.1, 0.25, 0.5\}$ for $p > n$, and the OR of the association. The latter was either held fixed for all causal variants in a simulation, with OR taking values of either 1.1, 1.2, 1.3, 1.4 or 1.5 for all SNPs, or drawn from a distribution. Given the bimodal distributions for OR of significant SNPs from GWAS, kernel density estimation (KDE) was used to estimate a probability density function (PDF) from the empirical distribution of odds ratios for genome-wide significant SNPs reported by the largest schizophrenia GWAS available when simulations were generated (Pardiñas et al., 2018). The KDE was tuned to give a close fit to the empirical distribution (Figure B.1). Odds ratios of causal SNPs were drawn randomly from the estimated PDF for simulations. Independent simulations were also evaluated in the presence of unassociated SNPs, with OR set to 1, or where causal SNPs were replaced by LD blocks. The chosen MAF range is a reasonable match to the MAF distribution reported for genome-wide significant SNPs (Figure B.2). As with additive simulations, when one parameter was varied, all others were kept constant.

The fundamental difference between additive and independent simulations is that the former keeps heritability constant, while the latter sets odds ratios of individual SNPs to be constant. As a result, the methods differ in how they adapt as the proportion of causal variants, m, and number of total SNPs, p, change. Considering p set to a constant value, as additive simulations maintain a fixed heritability, where m is low they task a small fraction of SNPs with explaining variance on the liability scale, giving them larger effect sizes. By contrast independent simulations apply the same range of effect sizes to SNPs whether m is large or small. As m increases, independent simulations imbue a large number of SNPs with similar odds ratios, which is expected to cause prediction performance to greatly increase. Additive simulations explain the same proportion of variation using a larger number of causal variants, and so assign smaller effect sizes, so that prediction performance may instead decrease. As a consequence associated variants may be more distinguishable from noise when m is low under additive simulations (Figure B.3).

4.2.2 Interaction effects

Pairwise SNP-SNP interactions were simulated using a method similar to that described by Zhang and Liu, 2007 (Figure 4.6). Minor allele frequencies were set for each locus, and the probability of co-occurrence under random mating, calculated for controls. The element-wise product of the probability of co-occurrence and the odds model matrices gives the probability in cases. A weighted random choice sampling method, which allows for probabilities that do not sum to one, is then used to draw genotype combinations at the

desired frequencies for cases and controls. 25 simulations were created for each 2-SNP interaction model. Interaction models are parameterised by the baseline odds ratio α, set to 1 for all simulations, and θ which describes the interaction effect. Purely epistatic models that consistently exhibit no marginal effects, such as Table 4.1, were not considered; however, M170 XOR, M78 XOR and M68 interference models, described below, show weak or no marginal effects under some simulation parameters.

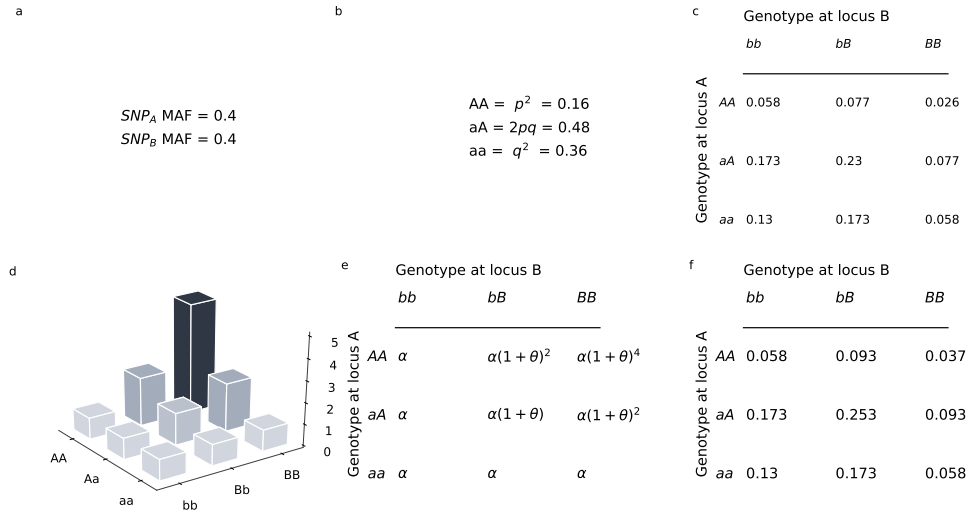
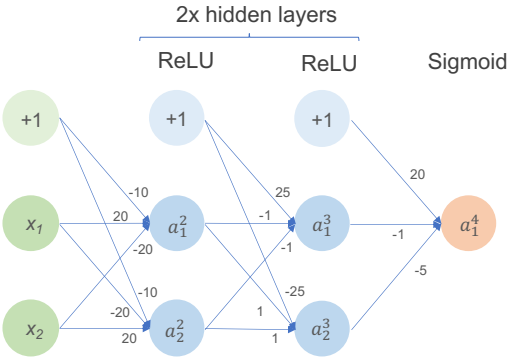


FIGURE 4.6: simulation of a 2-SNP multiplicative interaction model. Minor allele frequency is set to 0.4 for both SNPs (A), and Hardy-Weinberg Equilibrium is calculated (B). The product of HWE at both SNPs gives the frequency of each genotype combination between the two loci in controls (C). To set frequencies in cases, an interaction model (D) is first defined in terms of the baseline effect, α, and the interaction effect, θ (E), set to 1 and 0.5 here respectively. Finally, the product of the values in (C) and (D) give the frequencies in cases (F). Genotype combinations are then assigned in cases and controls according to these frequencies with some randomness. The same procedure was followed to produce 25 simulations of varying cases, controls, MAF, θ and LD for each interaction model.

Models were chosen to be consistent with previous literature on detecting epistatic effects, and to provide an assessment of simple and complex interactions. The 'multiplicative' and 'threshold' models are well-studied 2-loci interactions (Zhang and Liu, 2007; Marchini, Donnelly, and Cardon, 2005), while the XOR models show more complicated effects and have been previously described (Wan et al., 2010). Additionally, use of an XOR model here matches a history in computer science of using binary XOR problems to demonstrate an ability to learn complex functions, for example the requirement of a hidden layer in neural networks to fit a binary XOR model. The three-levels of SNPs, however, brings an increase in complexity to XOR problems; two hidden layers are sufficient to represent such functions, shown by the table in Figure 4.7b, which gives the prediction and calculations at each hidden unit for a combination of inputs from two loci and the manually chosen weights in 4.7a. Tracing each pair of inputs through the network in 4.7a produces predictions in 4.7b which correctly match the model for M170 XOR.



(A) Network architecture

Input		Hidden layer 1		Hidden layer 2		Output
x_1	x_2	a_1^2	a_2^2	a_1^3	a_2^3	a_1^4
0	0	0	0	25	0	$g(-5) \approx 0$
0	1	0	10	15	0	$g(5) \approx 1$
0	2	0	30	0	5	$g(-5) \approx 0$
1	0	10	0	15	0	$g(5) \approx 1$
1	1	0	0	25	0	$g(-5) \approx 0$
1	2	0	10	15	0	$g(5) \approx 1$
2	0	30	0	0	5	$g(-5) \approx 0$
2	1	10	0	15	0	$g(5) \approx 1$
2	2	0	0	25	0	$g(-5) \approx 0$

(B) Node outputs

FIGURE 4.7: Neural networks are capable of representing M170 XOR models using two hidden layers. A hypothetical deep neural network with two hidden layers is presented with weights chosen to produce the desired classification (A). Circles represent neurons, or nodes. The first layer is the input layer; in every subsequent layer, each neuron is a computation unit where inputs are combined using manually-chosen weights and the activation function for that layer. '+1' neurons are present in all layers, except the last, as bias terms. The final neuron uses a sigmoid function that gives an output between zero and one. The table shows the input values for an M170 XOR model, the results of calculations at each neuron and the final output of around 1 for a predicted case and zero for a predicted control (B). An output of 1 is correctly given for input combinations associated with higher risk of becoming a case. Outputs are all correctly predicted.

As multiple XOR and interference models have been proposed in genetics, those chosen have been attributed their codes M170, M78 and M68 from Li and Reich, 2000, where they are defined as fully-penetrant two-locus interactions on the penetrance scale; here they are transported to the odds scale (Figure 4.8). All methods were compared for interaction simulations, except the external PRS (detailed in section 4.2.3.1), as no marginal effects were set in simulations.

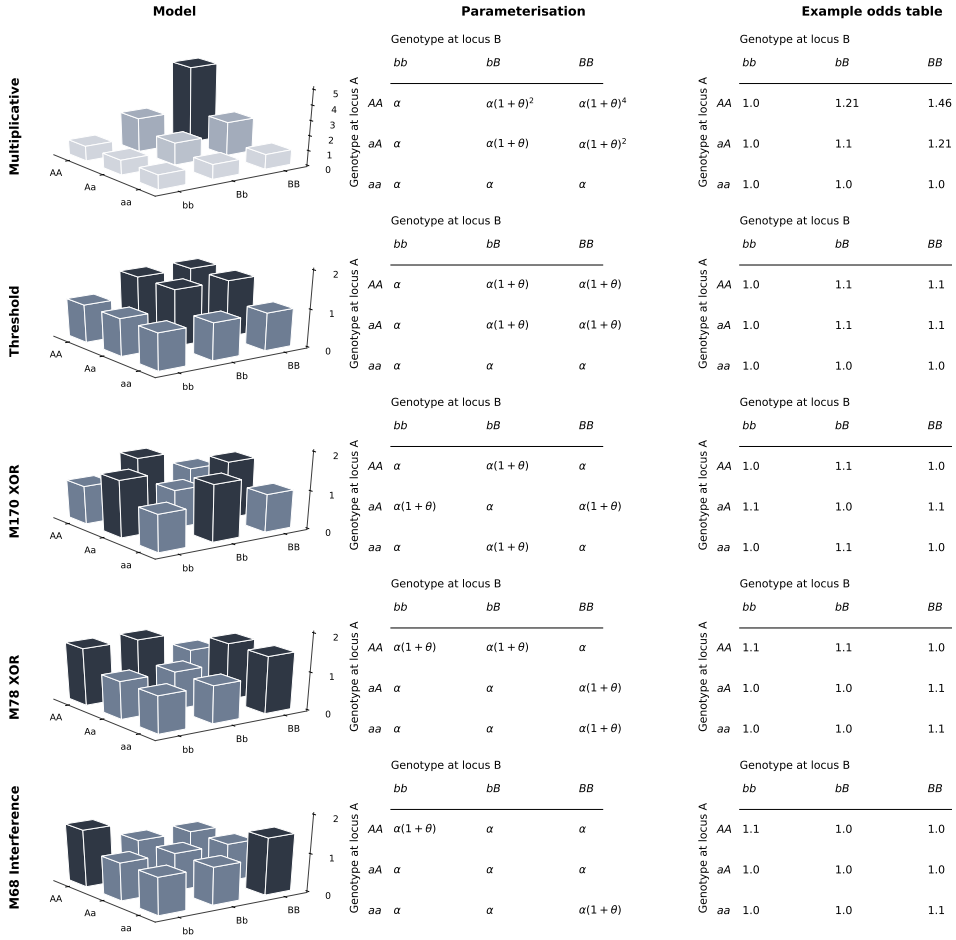


FIGURE 4.8: Interaction models used in simulations. The left column shows a 3D bar plot of the odds ratios for each two locus genotype with respect to the homozygous wild type, "aabb", using $\theta = 0.5$ for illustrative purposes. Odds ratio are given on the z axis (vertical). The central column gives the model in terms of the baseline effect α and the interaction effect θ. The right column gives an example of the expected odds table for a two-locus interaction with θ set to 0.1.

4.2.3 Statistical methods

4.2.3.1 Polygenic risk scores

Polygenic risk scores (PRS) are a simple weighted average of allelic effects across loci,

$$\sum_{j=1}^{p} \beta_j q_j, \tag{4.5}$$

where β_j is the effect size for SNP j, q_j is the number of risk alleles at SNP j, and p is the number of SNPs. They can be broken down into two components: the additive model and the weights. To separate these, two different versions of PRS are given, named here as external PRS (ePRS) and internal PRS (iPRS). ePRS is a traditional PRS with effect sizes taken from an external source; here β are the effects in the population set during simulations. For additive and independent simulations ePRS can be interpreted as the best

possible prediction, as it contains the true effect sizes in the population; it is therefore a useful benchmark but not a realistic comparison. In contrast, β in iPRS is an estimate of the effect sizes from fitting a univariable logistic regression model in the training set for each SNP. All PRS are taken as the sole predictor in a logistic regression.

4.2.3.2 Machine learning algorithms

Ridge, LASSO, linear support vector machines (SVMs), radial basis function kernel (RBF) SVMs, random forests, extreme gradient boosting (XGBoost) and neural networks were evaluated as the most widely reported methods in psychiatric genetics (described in chapter 3 and compared to unpenalised logistic regression, iPRS and ePRS.

Ridge (Hoerl and Kennard, 1970) and LASSO (Tibshirani, 1996) logistic regression apply L_2 and L_1 norms to the negative log-likelihood to give a penalised objective function. Ridge and LASSO provide dense (all coefficients are non-zero) and sparse (some coefficients are shrunk to zero) solutions respectively; LASSO therefore performs predictor selection. Values for C, which controls the strength of the penalty term in the objective function, were drawn from the exponential distribution with $\lambda = 1$.

SVMs approximate the target function geometrically and use the side of the hyperplane on which a test observation falls as a means of direct class assignment (Cortes and Vapnik, 1995). Linear SVMs can only learn linear functions, while RBF-kernel SVMs are flexible models capable of learning non-linear decisions boundaries. The distribution for C is the same as for penalised regression; γ, which controls how flexible the decision boundary is, was drawn from the exponential distribution with $\lambda = 10$.

Random forests ensemble CARTs, a form of decision tree which takes a greedy approach to partition the predictor space, to reduce the variance of the predictions from low-bias high-variance trees (Breiman, 2001). Maximum depth of trees was drawn from a shifted binomial distribution with two trials and a 0.45 chance of success per trial, such that values were generally less than 10 with a minimum of 2; minimum samples per split was drawn from a similar distribution but shifted to values closer to 2. Max features was searched between $\sqrt{p}$ and $\frac{p}{3}$.

XGBoost implements an efficient, heavily-regularised gradient boosting algorithm which sequentially adds weak base learners to build a strong model (Chen and Guestrin, 2016). Learning rate was drawn from the log-uniform distribution between 0.001 and 0.25. Proportion of columns and rows were taken uniformly from 0.4 to 1 and 0.5 to 1 respectively. Maximum depth of trees was taken from a shifted binomial distribution with two trials, 0.5 chance of success and minimum value of 2, favouring a slightly smaller tree than random forests. L_2 penalty was drawn from a shifted exponential distribution with $\lambda = 0.16$ and a minimum value of 1.

Neural networks combine units of computation through a network of weights which are updated through backpropagation (LeCun, Bengio, and Hinton, 2015). Deep learning models consisted of feed-forward multilayer perceptrons with ReLU activation functions in all hidden units and a sigmoidal activation function in the output neuron. Overfitting was controlled through the L_2 penalty. Learning rate with momentum and weight decay were both drawn from the log-uniform distribution between 0.0001 and 0.1, with the momentum coefficient set to 0.9. Hidden layers were fixed to be 1 or 2, with number of hidden units chosen from between $\frac{1}{2}p$ and $\frac{3}{2}p$. All simulations used 15 epochs, a batch size of 32, batch norm (Ioffe and Szegedy, 2015) and He initialisation of the weights (He et al., 2015).

Hyperparameter tuning differed slightly between main and interaction effect simulations for some models. Ridge, LASSO, linear SVM, RBF SVM and random forests kept the same distributions for both types of simulation. However, XGBoost made no use of L_2 penalisation or column sampling for interactions. Neural networks also allowed for a lower learning rate (uniformly drawn from 0.00001 to 0.1 on the log scale) and L_2 weight decay (0.00001 to 0.01, also from the log-uniform distribution) for interactions. In addition, architecture was searched for 3-6 hidden layers, each having between 2 and 8 hidden neurons, as smaller networks following the design used for main effects struggled to learn interaction models.

Unpenalised multivariable logistic regression was only run for simulations where $p < n$, as a unique solution for coefficient estimates cannot be obtained when the number of predictors outstrips the number of observations. ePRS was not compared when causal SNPs were replaced with LD blocks in simulations of main or interaction effects, as these do not directly assign population parameters to each SNP in the simulated dataset. As LASSO, random forests and XGBoost are capable of predictor selection, they are referred to as "sparse" methods, while remaining approaches are labelled "dense".

4.2.4 Model development

Separate train and test simulations were drawn for developing and validating models. Training sample sizes varied; test sets for main effects had a sample size of 10,000, while interactions had a sample size of 100,000. Models were trained using 5-fold cross-validation with Monte Carlo random hyperparameter search, as described in chapter 2, with the best model refit on the entire training set before predicting on the test set.

To ensure many models could be trained, 30 iterations of random search were performed for machine learning models. As all predictors are SNPs with values in the set $\{0, 1, 2\}$, no scaling was performed within cross-validation. All PRS were z-transformed within cross-validation using a pipeline.

4.2.5 Model evaluation

For simulations with interaction effects, visualisations of the decision boundary were computed by predicting the probability of the positive class (cases) for all combinations of a

continuous range of values between -0.5 and 2.5 at each locus. The result of this is visualised as a contour plot. The boundary is an estimation of the function learned on the training set, not a calculation of the exact decision boundary, nor a representation of the algorithm's predictive ability on a new dataset. All methods were compared for discrimination using the area under the receiver operating characteristic (AUC) in the test set.

4.2.6 Code

4.2.6.1 Simulations

All simulations were run in Python 3 using core packages from Python's scientific computing stack: SciPy (Virtanen et al., 2020), NumPy (Walt, Colbert, and Varoquaux, 2011) and Pandas (McKinney et al., 2010). All images were generated using seaborn 0.10.1 (Waskom et al., 2020) and matplotlib 3.2.1 (Hunter, 2007). Separate classes were defined for independent, additive and interaction simulations, each with a 'simulate' method which can be called repeatedly to generate new simulations with an incrementing random seed. Test sets were created by including an option to maintain simulation parameters, such as LD structure, MAF, effect sizes, heritability or location of causal SNPs, but supply a new seed and sample size before drawing genotypes from the binomial distributrion. All classes inherit from a base class which has methods for adding simple or complex LD as SNPs or blocks of SNPs, to combine with or replace the causal SNP, and methods for plotting LD, adding noise SNPs and checking effect sizes. All simulations and analyses were run on the Cardiff Hawk supercomputer, part of the Supercomputing Wales project. Code for simulations is available at https://github.com/seafloor/simulations.

4.2.6.2 Machine learning

Neural networks were implemented in PyTorch v1.5. A Python wrapper was written to ensure cross-validation and any pipeline-dependent operations such as scaling or generating an iPRS could be carried out in the same manner for PyTorch tensor's and NumPy arrays. iPRS and ePRS were calculated by creating a custom transformer in scikit-learn that can accept effect sizes or derive them from the training fold, and can be incorporated into pipelines to ensure it is correctly refit in each training fold for cross-validation. All cross-validation used the same random splits across Scikit-learn and PyTorch. The Python API was used for XGBoost (Chen and Guestrin, 2016), and all other algorithms were done in Scikit-learn (Pedregosa et al., 2011).

4.3 Results

4.3.1 Main effects

Comprehensive simulations of main and interaction effects were undertaken (Figure 4.9).

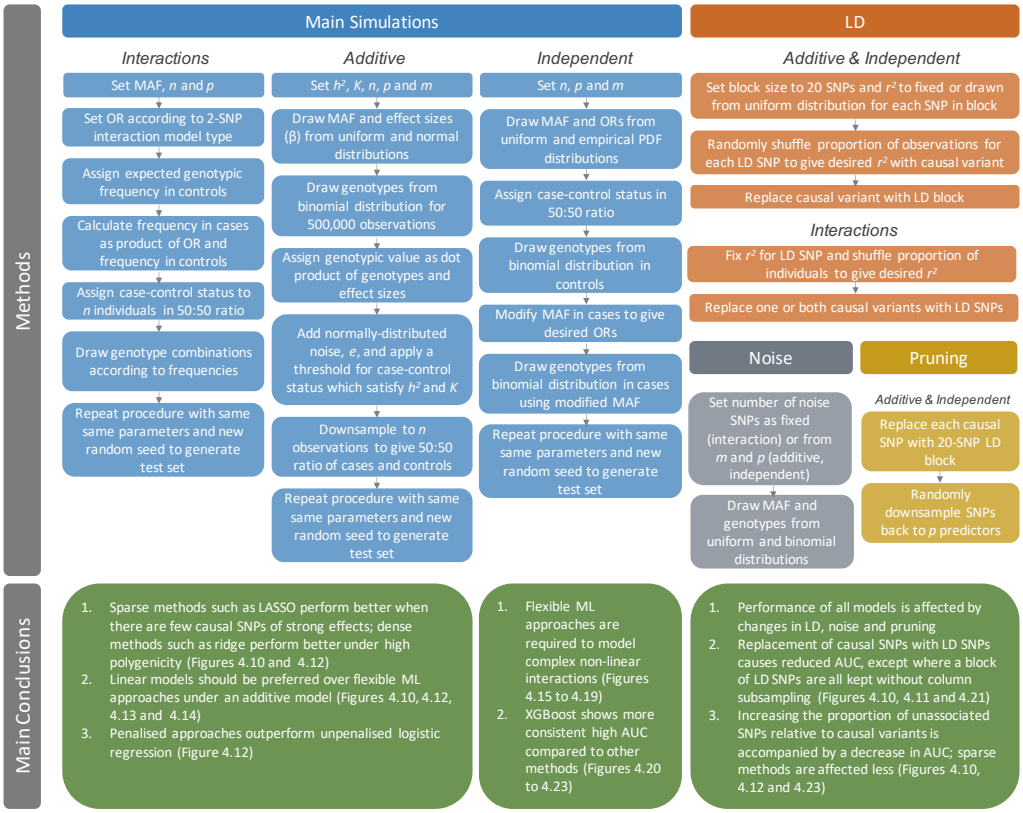


FIGURE 4.9: Workflow and key conclusions for additive, independent and interaction simulations. h^2: narrow-sense heritability, K: prevalence, m: proportion of SNPs which are causal, n: number of observations, p: number of SNPs, AUC: area under the receiver operator characteristic curve, LASSO: least absolute shrinkage and selection operator, LD: linkage disequilibrium, MAF: minor allele frequency, ML: machine learning, OR: odd ratio, PDF: probability density function, SNP: single nucleotide polymorphism, XGBoost: extreme gradient boosting.

For main effects, the proportion of causal SNPs, m, in a simulation was evaluated at $p < n$ with 1,000 observations and 200 SNPs, where the remaining non-causal SNPs are drawn randomly from the binomial distribution (Figure 4.10). Additive (Figure 4.10a) and independent (Figure 4.10b) simulations show contrasting results. While independent simulations showed increasing test set AUC for all classifiers as m increased, additive simulations demonstrated roughly constant AUC for iPRS, logistic regression, ridge regression, SVMs and neural networks, but falling AUC for models which perform embedded predictor selection: random forest, XGBoost and LASSO. Under additive simulations, such sparse approaches all showed greater AUC than dense models where $m = 0.05$. This approached the maximum AUC given by ePRS, calculated using the population effect sizes set during simulations, as m decreases, but equal or worse discrimination when $m = 1$. Sparse machine learning approaches therefore show greater discrimination than dense methods when a few variants of large effects are simulated, but worse discrimination for a large number of SNPs of smaller effect.

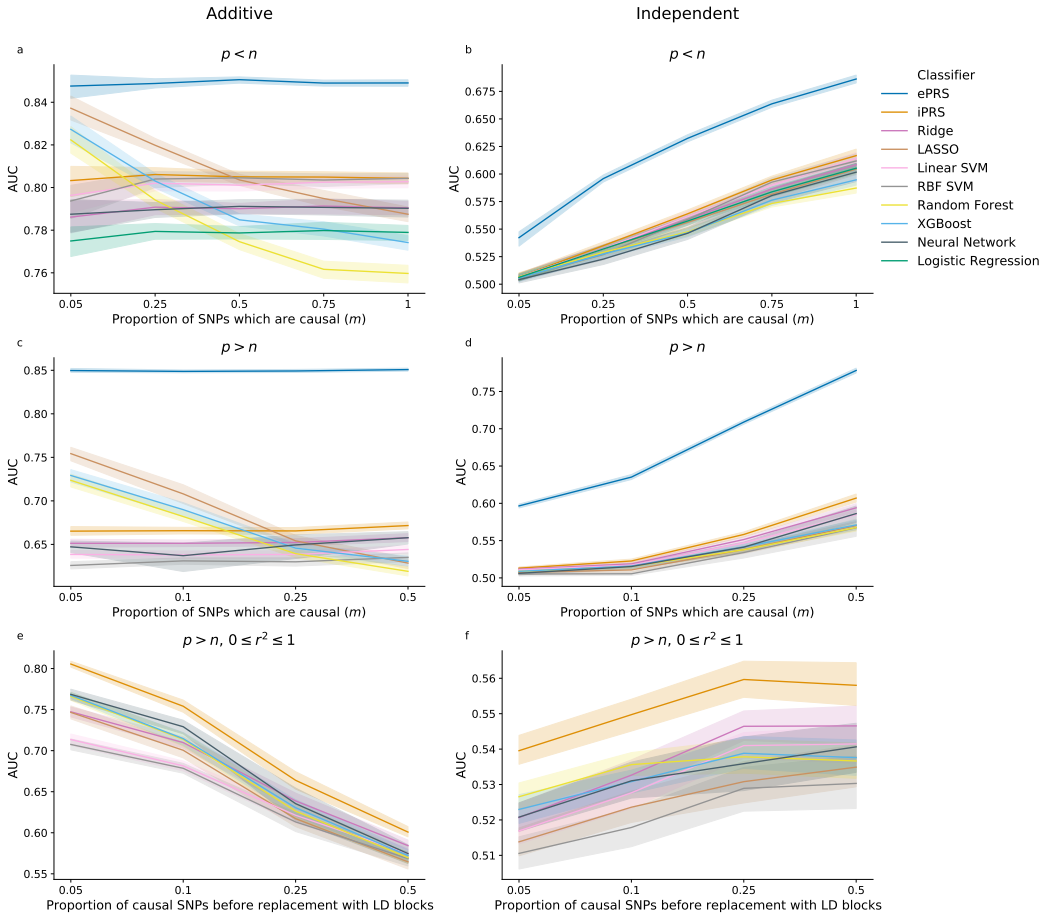


FIGURE 4.10: Discrimination of independent and additive simulations with varying proportion of causal SNPs, m. Additive simulations (a, c, e) are contrasted with independent simulations (b, d, f). $p < n$ scenarios set $p = 200$ and $n = 1000$, while $p > n$ used $p = 1000$ and $n = 500$. Simulations with LD (e, f) create 20-SNP LD blocks where r^2 with the causal SNP is drawn uniformly between 0 and 1; these replace all original SNPs, including those unassociated with the outcome. All independent simulations used empirical PDF-derived values; all additive simulations use $k = 0.0025$ and $h^2 = 0.2$. ePRS was not calculated in e and f as population effect sizes are not set for LD SNPs.

Discrimination was also examined when $p > n$, but only a proportion of SNPs are causal, as is common in genetics of complex traits (Figure 4.10c and d). Unlike $p < n$ simulations, m was kept less than 0.5 to focus on the more likely scenario in high dimensions where a smaller fraction of SNPs show association with the outcome. Results demonstrate the same relationships as when $p < n$. By contrast, replacing all causal SNPs with LD blocks where r^2 is sampled uniformly from between 0 and 1 (Figure 4.10e and f) shows decreasing discrimination for all models as m increases under additive simulations, but increasing discrimination for independent simulations. While m is the initial proportion of causal SNPs chosen in simulations, the LD-blocks replace all SNPs in the original dataset to create simulations which vary in the strength of LD block structure; the majority of SNPs in the final dataset will therefore show some association with the outcome. Where $n = 500$, $p = 1000$ and $m \in \{0.05, 0.1, 0.25, 0.5\}$, original simulations of 50, 100, 250 or 500 causal SNPs result in around 1000, 2000, 5000 and 10000 associated SNPs after replacement with 20-SNP

LD blocks; the exact number will be lower on average as LD may be set to 0. Replacing $m = 0.05$ simulations with 20-SNP LD blocks recreates a dataset of $p = 1000$, but with strong correlation through intact block structure. At $m = 0.5$, a $p = 10000$ dataset is made which is subsampled back to $p = 1000$, creating more sparse LD structure. As a consequence, the lowest values of m result in a larger number of SNPs with weaker effect sizes than in $p > n$ simulations with no LD, so that the contrasting relationship between sparse and dense methods is no longer observed. Independent simulations with LD show the same increase in AUC with m as seen in simulations without LD, but with a dampened increase in discrimination. Results from additive simulations with LD illustrate that discrimination when $m = 0.5$ is low relative to when $m = 0.05$, which can be ascribed to the assignment of smaller effect sizes to a larger number of SNPs when $m = 0.5$, combined with the less direct predictor-response relationship than that simulated in datasets without LD.

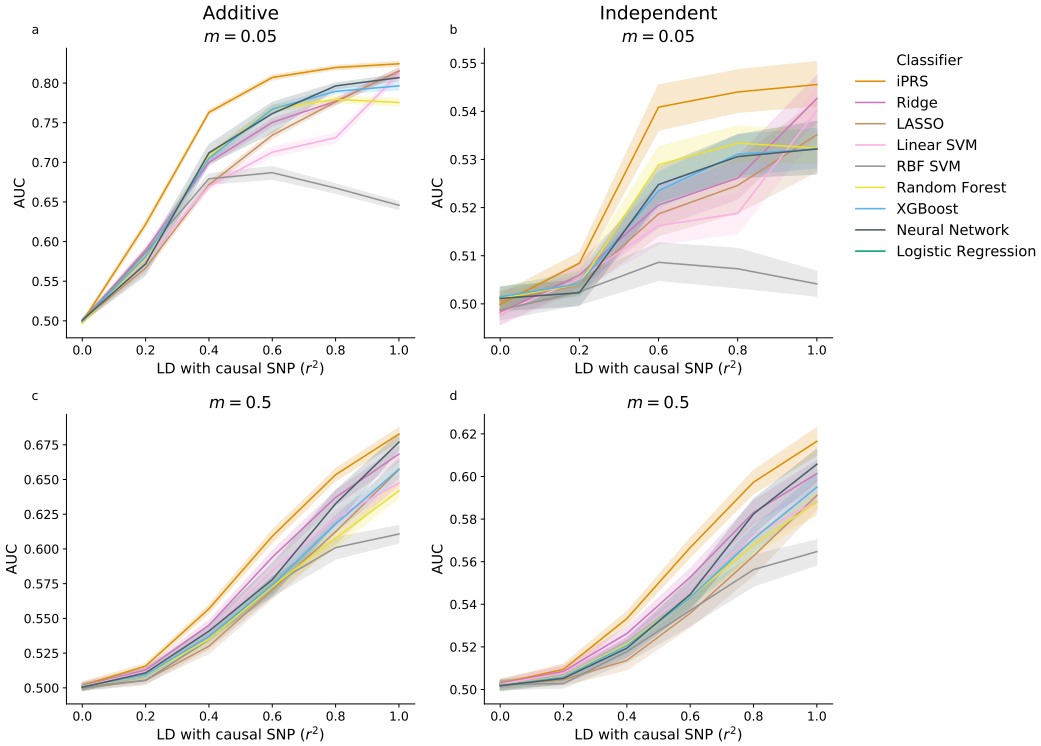


FIGURE 4.11: Varying r^2 in LD blocks when $p > n$ for $m = 0.05$ and $m = 0.5$. All independent simulations used empirical PDF-derived values; all additive simulations use $k = 0.0025$ and $h^2 = 0.2$. ePRS was not calculated as population effect sizes are not set for LD SNPs.

A cleaner but less realistic view of the effect of LD can be seen by setting LD to constant values when $p > n$ (Figure 4.11). For both simulations of main effects, AUC shows a slight sigmoidal relationship with r^2 for fixed $n = 500$ and $p = 1000$ when $m = 0.5$. At $m = 0.05$, all SNPs are part of strong uniform LD blocks in association with the outcome, as illustrated in Figure 4.5a; as such performance saturates as LD approaches 1. At $m = 0.5$, replacement of causal variants with 20-SNP LD blocks creates 10000-dimension datasets

which were subsampled back to $p = 1000$, creating the more sparse LD structure in Figure 4.5b. Performance does not saturate as not all variants are present.

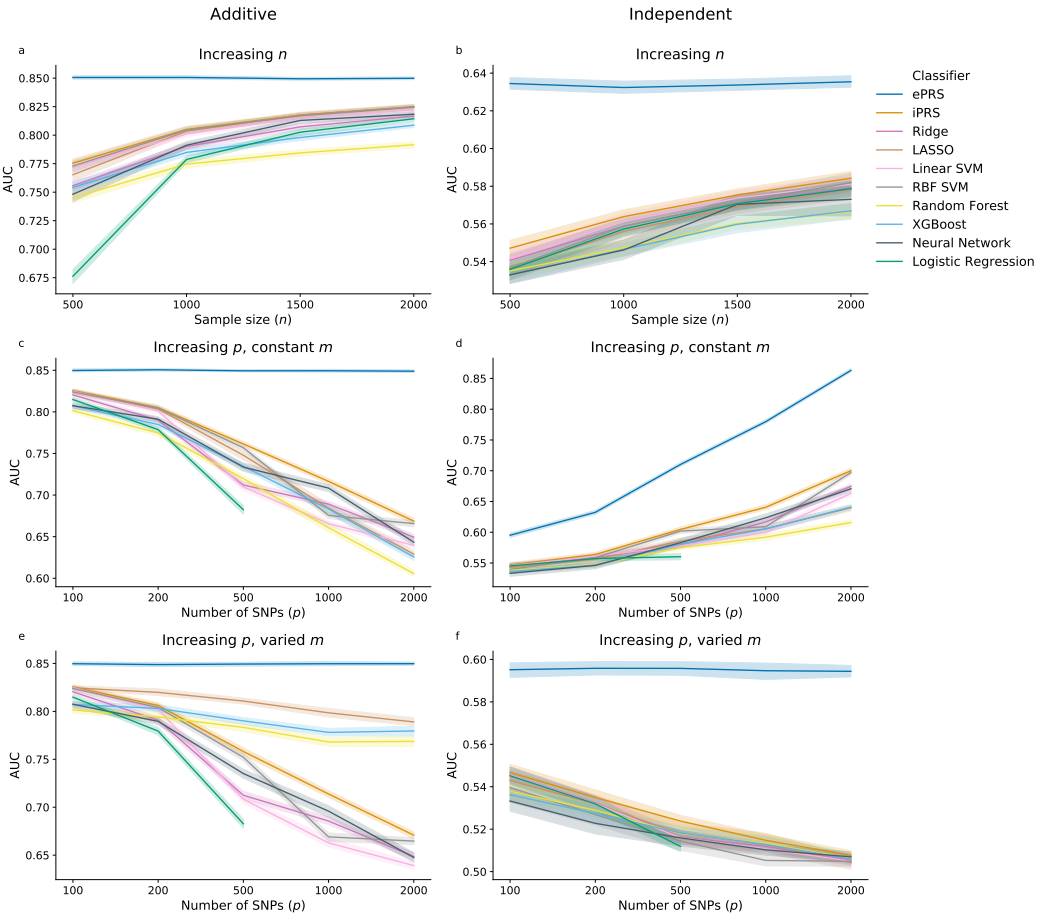


FIGURE 4.12: Varying sample size and number of predictors in simulations of main effects. For different values of n (a, b), $p = 200$ and $m = 0.5$. For simulations altering p with fixed m (c, d), $n = 1000$ and $m = 0.5$. Simulations increasing p, while also varying m to maintain a constant number of causal variants (e, f), set $n = 1000$ and $m \in \{0.5, 0.25, 0.1, 0.05, 0.025\}$.

Increasing sample size for all simulations of main effects results in improved discrimination (Figure 4.12a and b). Increases from baseline (500 observations) to the maximum (2000 observations) causes around 2-5% increase in AUC for most models in both simulations. Increasing the number of predictors, p, while holding m constant, again shows characteristic negative and positive relationships between AUC and p for additive and independent simulations respectively (Figure 4.12c and d). As the proportion of causal SNPs is kept constant, the actual number of causal variants increases with p, and so discrimination also rises with p for independent simulations. This represents a scenario where additional SNPs include some associated variants so that the proportion associated stays constant. Under an alternative scenario, the number of causal variants was kept constant at 50 by setting $m = \frac{50}{p}$ for increasing values of p (Figure 4.12e and f). Here, the additional SNPs do not contain more associated variants, and so only dimensionality and noise are increased. Discrimination decreases with p for both additive and independent simulations in this situation.

All additive simulations set prevalence, K, to 0.0025, and heritability, h^2, to 0.2. Though these may be considered low compared to typical estimates for schizophrenia, assessing discrimination for different values of K and h^2 demonstrated the same relationship between models (Figure B.4). Discrimination shows a positive relationship with heritability and a negative relationship with prevalence (Figure 4.13). As such, raising h^2 to 0.25 increased discrimination, while setting K as 1% decreased discrimination. With the exception of additive simulations where m is low, iPRS performed as well or better than logistic regression or machine learning approaches.

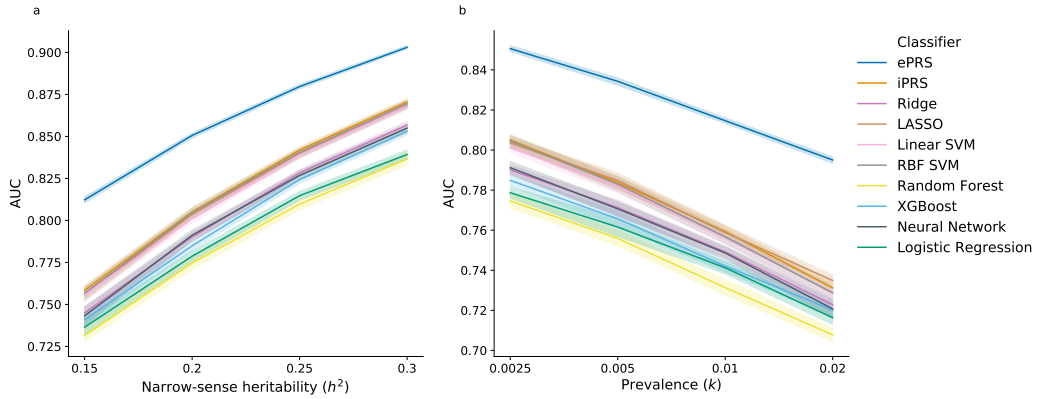


FIGURE 4.13: Evaluating discrimination for additive simulations under alternative values for narrow-sense heritability (a) and prevalence (b). AUC shows an approximately linear relationship with both, where simulations are fixed to $n = 1000$, $p = 200$ and $m = 0.5$ to be comparable with previous $p < n$ simulations.

Independent simulations set effect sizes of causal variants to all be drawn from an estimated PDF fit to the distribution of odds ratios for genome-wide significant SNPs (Pardiñas et al., 2018). Odds ratios were also assessed when set to a constant value for all variants in a dataset. Comparing the empirical odds ratios against $OR \in \{1.1, 1.2, 1.3, 1.4, 1.5\}$ shows a strong linear relationship between AUC and effect sizes (Figure 4.14). Discrimination for the empirical PDF odds ratios are similar to those when all ORs are set to 1.1.

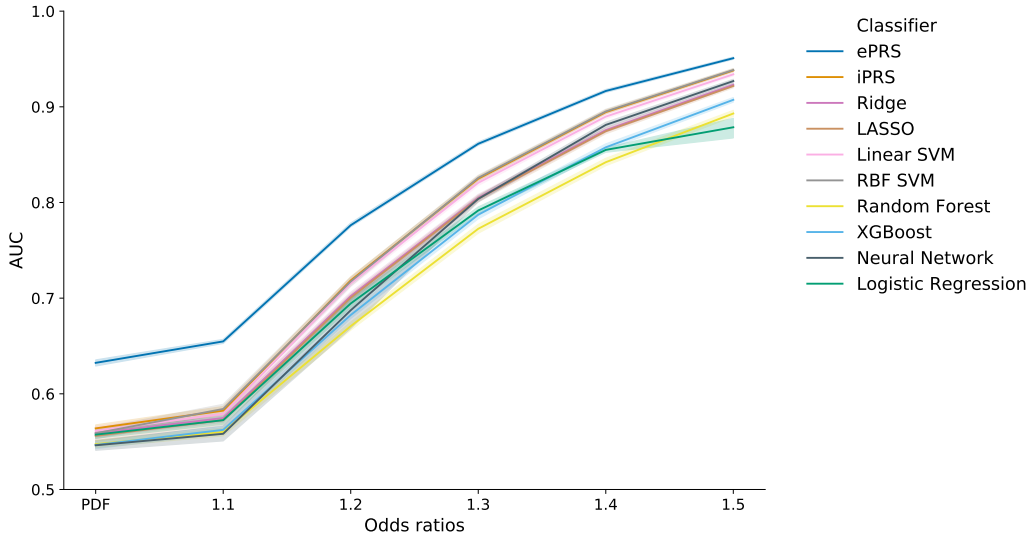


FIGURE 4.14: Discrimination for independent simulations with odds ratios set to constant values or drawn from an estimated PDF from GWS SNPs in Pardiñas et al., 2018. For all simulations $p = 200$, $n = 1000$ and $m = 0.5$.

4.3.2 Interaction effects

Interactions were simulated under 5 different 2-SNP interaction models, with varying effect sizes (parameterised by θ), MAF, sample size, LD and number of additional unassociated SNPs. Holding number of observations fixed at $n = 2000$, discrimination from 2-SNP interactions with no additional SNPs simulated were strongly influenced by MAF and θ for all models (Figure 4.15), with AUC either at or slightly above chance for all models when both θ and MAF are set to 0.1. Increasing MAF for a fixed θ, or vice versa, caused small improvements in discrimination, with average AUC across machine learning approaches highest when $\theta = 0.5$ and MAF $= 0.5$.

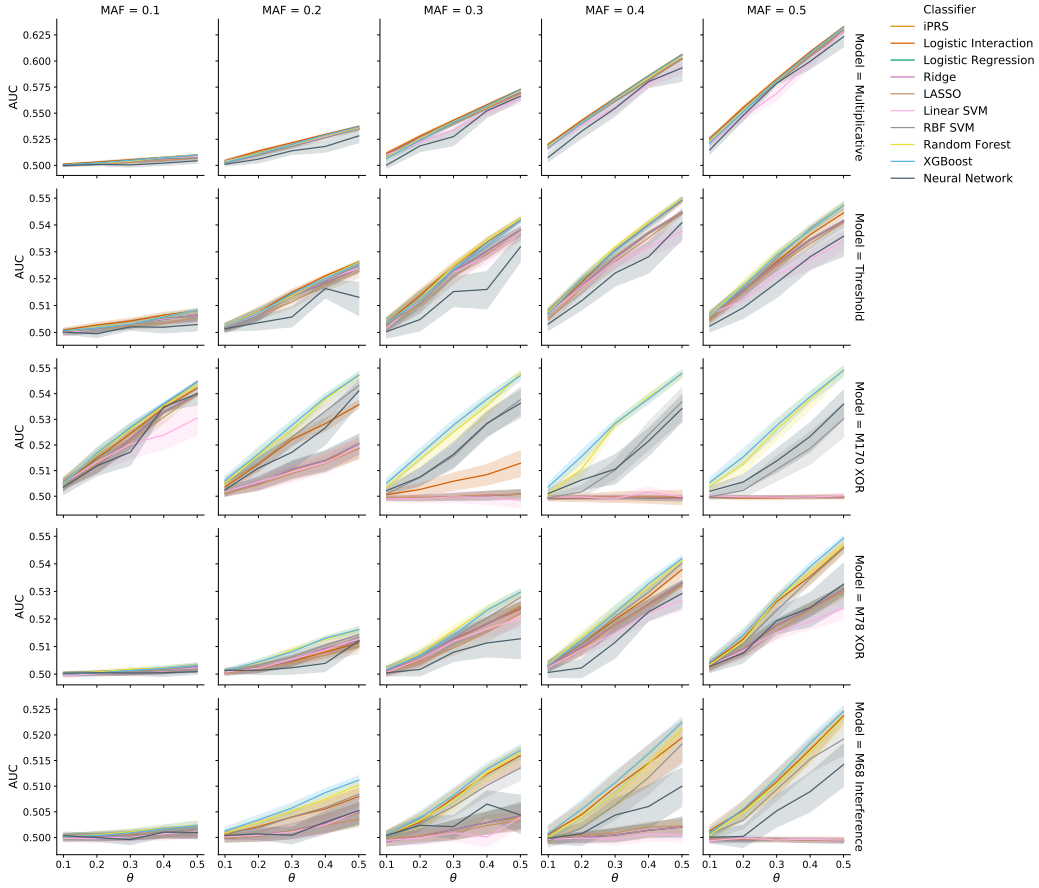


FIGURE 4.15: Discrimination of classifiers for simulations of 2-SNP interaction effects at $p = 2$, $n = 2000$, $MAF \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\theta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Interaction models which can be mostly separated by a linear decision boundary (multiplicative and threshold) show similar performance between classifiers. XOR and interference models show differences between linear and non-linear models, particularly for high MAF and θ. A θ of 0.1 corresponds to an odds ratio of 1.1 when the interaction between genotypes is given by $\alpha(1+\theta)$ and baseline odds ratio α is set to 1; multiplicative models give higher odds ratios of 1.21 and 1.46 for genotype interactions of $\alpha(1+\theta)^2$ and $\alpha(1+\theta)^4$ respectively when $\theta = 0.1$. Figure 4.8 describes model parameterisation further. Y axis limits differ between rows of subplots to allow for trends to be clear for all models.

Interactions represent a deviation from a linear model, where interaction terms or other methods are required to correctly model the effects. This situation is a natural fit for flexible methods that can learn complex patterns. Visualisations of the decision boundary confirm expectations that flexible ML methods are capable of fitting simple and complex interaction models (Figures 4.16, 4.17, 4.18). Linear statistical and ML methods gave reasonable approximations of multiplicative and threshold models, but were unable to approximate the more complicated simulations of XOR and interference models. The ability to learn complex models was not uniform across non-linear machine learning algorithms, however. RBF SVMs, random forests and XGBoost generally gave more consistent results than neural networks (Figure 4.15), and were more able to learn both linear and non-linear models well.

A notable exception to the rule that low MAF results in AUC around 0.5 is the M170 XOR

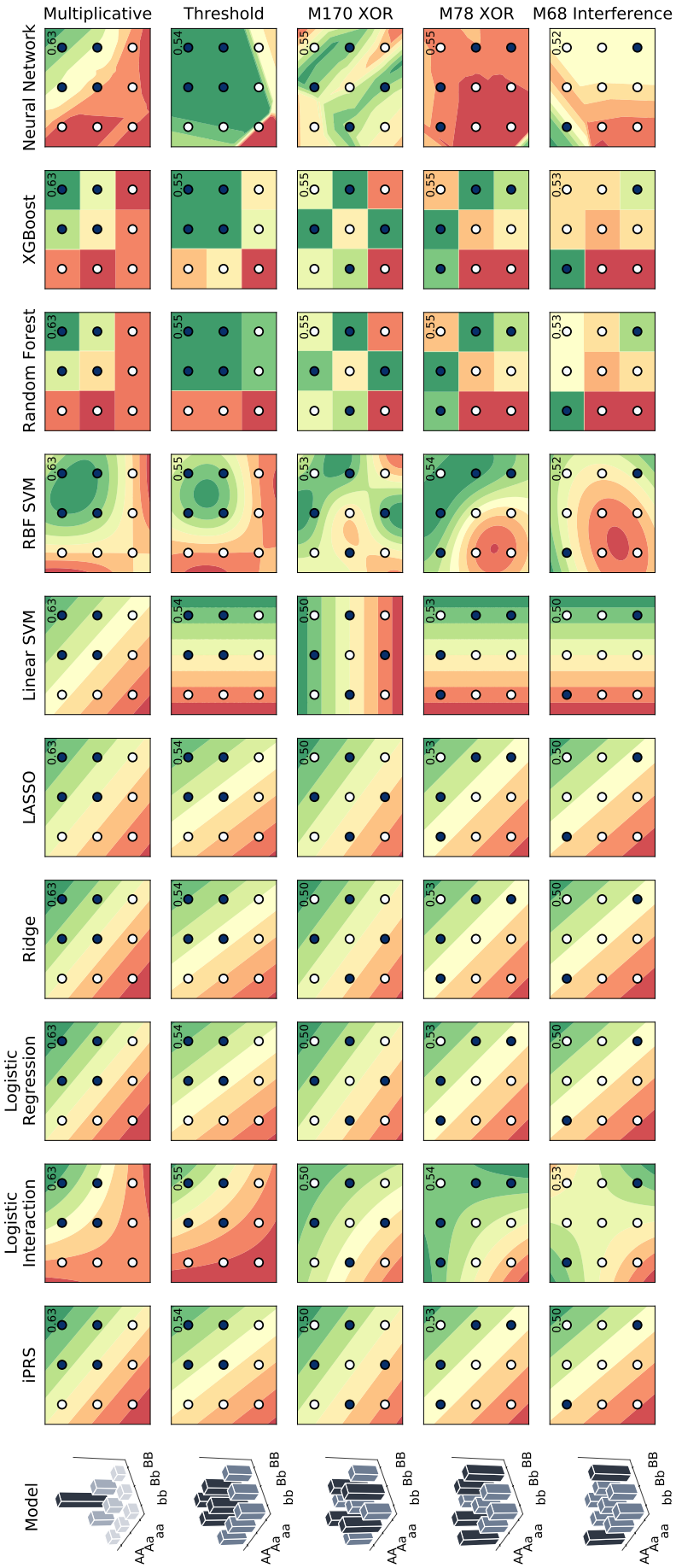


FIGURE 4.16: Decision boundaries for all classifiers under 2-SNP interaction models when $\theta = 0.5$ and MAF $= 0.5$. Boundaries are displayed by contour plots. Green regions are on the positive side of the decision boundary and red the negative. X and Y axes indicate the two loci, with dark points highlighting genotype combinations which increase risk. Effective classifiers should highlight dark points in green and white points in red. AUC, annotated on the top right of each subplot, does not use a single threshold, and so classifiers may have high AUC but still assign both light and dark points to the negative class.

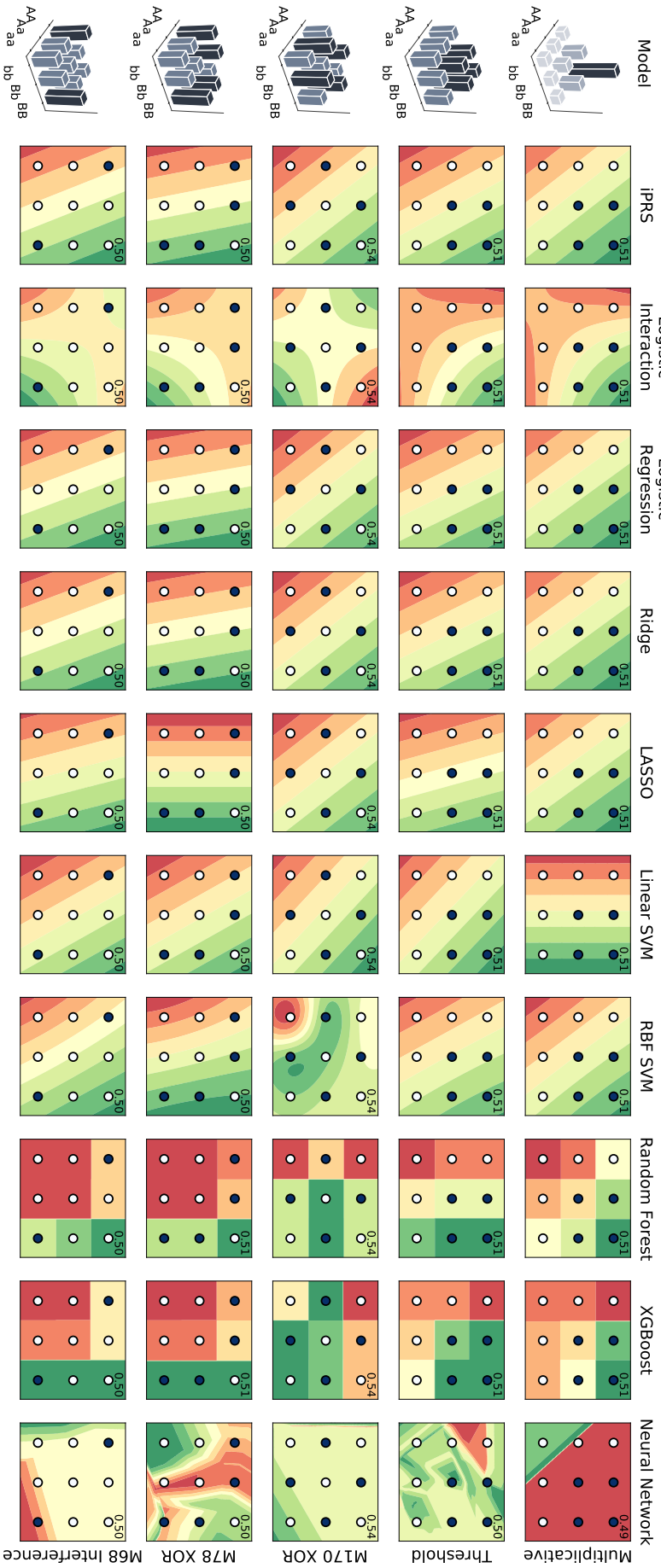


FIGURE 4.17: Decision boundaries for all classifiers under 2-SNP interaction models when $\theta = 0.5$ and MAF $= 0.1$.

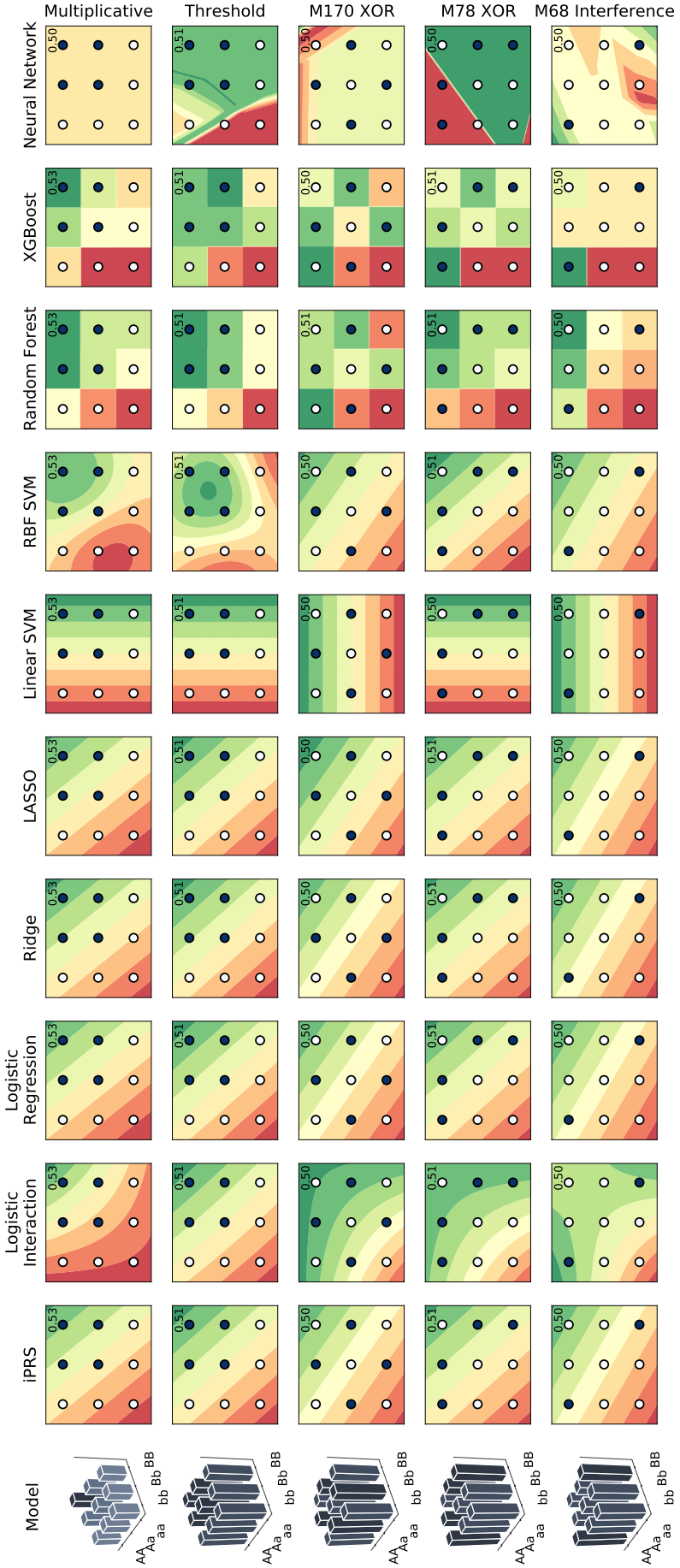
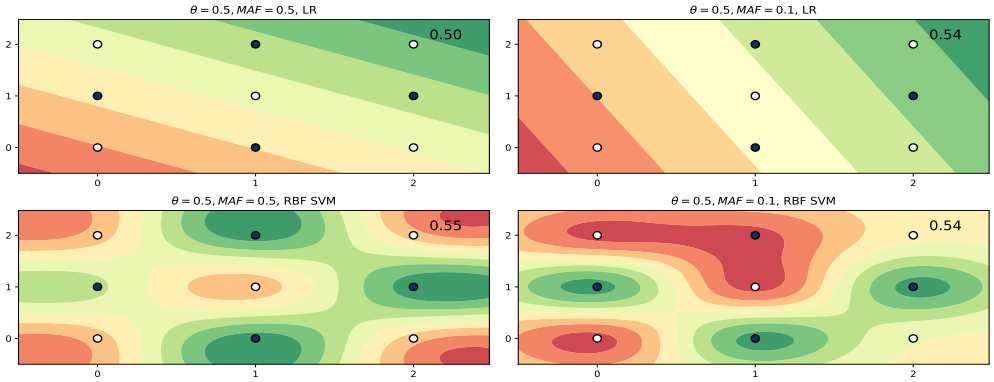


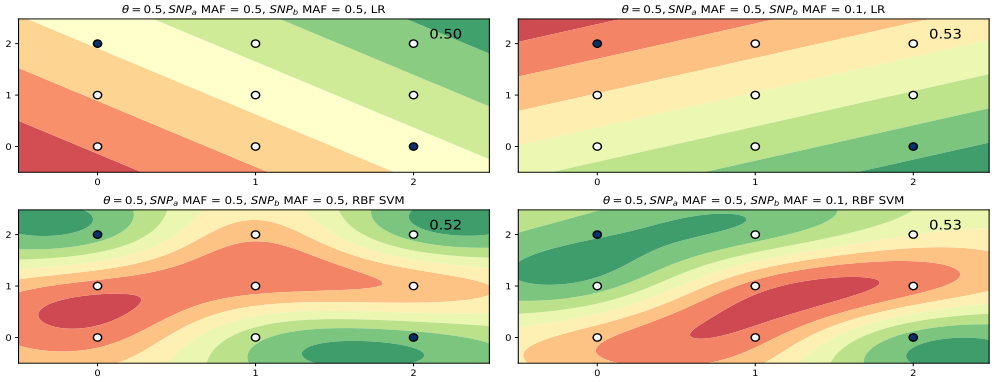
FIGURE 4.18: Decision boundaries for all classifiers under 2-SNP interaction models when $\theta = 0.1$ and MAF $= 0.5$.

model, which showed higher discrimination than other interaction models and a strong positive relationship between AUC and θ when MAF $= 0.1$ (Figure 4.15). This is due to M170 XOR's use of wild type homozygote-heterozygote combinations, aAbb or aabB, where A and B are risk alleles. All other models depend on double heterozygotes (aAbB), double homozygotes for the risk allele (AABB), risk-wild type homozygotes (AAbb, aaBB) or risk homozygote-heterozygote combinations (AAbB, aABB), as illustrated in the parameterisation column of Figure 4.8.

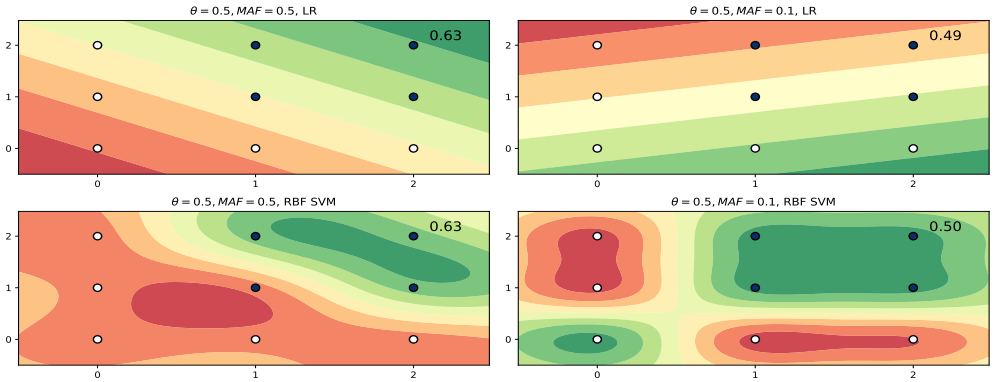
As MAF is assigned to the risk allele and $MAF \leq 0.5$, the wild type homozygote-heterozygote combinations (aAbb, aabB) used in the M170 XOR model are more likely to be observed in the training set for low n and MAF than the risk homozygote-heterzygote combinations (AAbB, aABB). The double homozygote (AABB), is not expected to be observed in the training data when $n = 2000$ and $MAF = 0.1$ (Table B.3); it is expected to be present at $MAF = 0.5$. When MAF is high, the full XOR model is simulated; it is unlearnable by linear models and noise dictates the decision boundary. At low MAF, genotypes in the upper right in Figure 4.19a influence the model less, causing linear models to base the decision boundary on the aAbb and aabB combinations, with the result that AUC increases for linear models at lower MAF (Figure 4.19a). A related scenario occurs for decision boundaries in M68 interference models (Figure 4.19b) where a drop in MAF at a single locus causes an increase in discrimination for linear models (Figure 4.22). These contrast with multiplicative models, where AUC increases with θ as expected for higher MAF (Figure 4.19c). Expected number of observations for each AABB genotype combination when $n = 2000$ are given for each interaction model in the appendix (Tables B.1, B.2, B.3).



(A) M170 XOR



(B) M68 Interference



(C) Multiplicative

FIGURE 4.19: Decision boundaries for logistic regression (LR) and RBF SVM trained on M170 XOR (A), M68 interference (B) and multiplicative (C) models. Presence of genotypes AABB, AAbB or aABB confound linear models when the classification task cannot be solved by a linear model. At low MAF, these risk genotypes are less likely to be observed in training data at $n = 2000$ and linear models can achieve better discrimination (A). The opposite situation occurs for multiplicative models (C). Decision boundaries for M68 interference are shown where MAF at only one SNP is dropped (B), while M170 XOR and multiplicative show the effect of a drop in MAF at both loci. LR and RBF SVM illustrate typical decision boundaries for linear and non-linear models. AUC in the test set is annotated in the top right of each plot.

To examine the sample size at which discrimination from interactions moves above chance, machine learning approaches were assessed for three scenarios. As stronger effects are generally more common at lower MAF, the first simulated high MAF (0.5) and low effect sizes

($\theta = 0.1$), the second low MAF (0.1) and high effect sizes ($\theta = 0.5$), and the third with both MAF and θ set to 0.5, each with increasing sample size $n \in \{1000, 2000, 5000, 10000\}$. Interactions models which had previously shown discrimination around chance for all classifiers at $MAF = 0.1$ or $\theta = 0.1$ showed minor improvements in AUC (typically less than 1%) when $n = 10000$ (Figure 4.20). Similarly, machine learning and statistical approaches previously demonstrated AUC between around 0.5 and 0.6, with classifiers using multiplicative interaction simulations showing the highest AUC. Discrimination for these increased with n by around 1-3%.

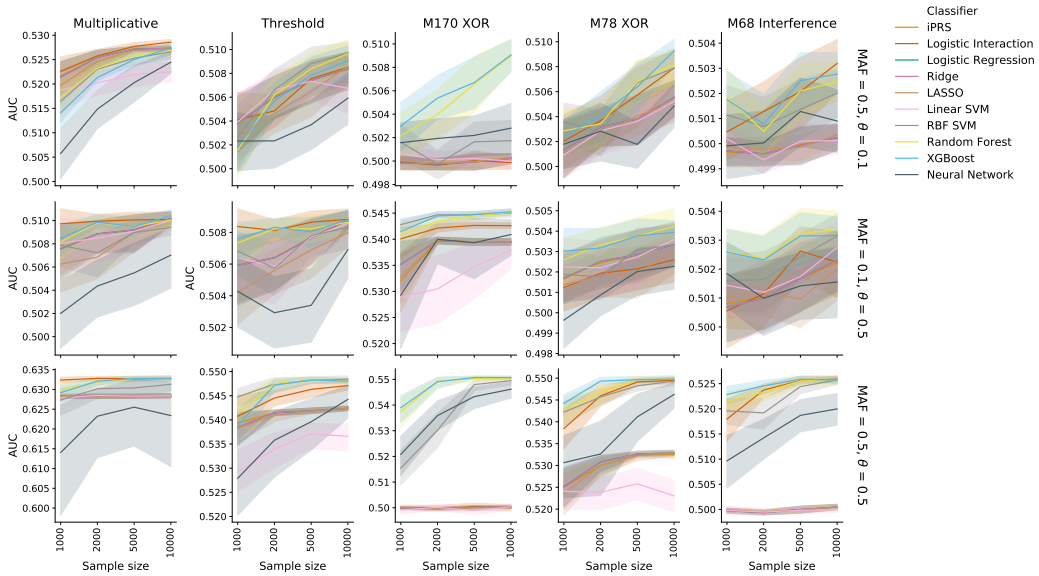


FIGURE 4.20: Discrimination increases on average with sample size. Varying sample size was assessed for common variants with small effects (top), less common variants with strong effect (middle) and common variants with strong effects (bottom). Y axis limits differ between subplots to allow for trends to be clear for all models.

Results in Figure 4.15 represent an ideal scenario of tagging both causal SNPs in a 2-SNP interaction and learning when dimensionality is very low ($p << n$). Three scenarios assessed how discrimination degraded in response to decreasing LD with the causal SNP at one or both loci (Figure 4.21), decreasing MAF in one or both loci (Figure 4.22), and addition of noise through unassociated SNPs (Figure 4.23). To ensure initial AUC was high enough to observe any decrease, all simulations fixed $\theta = 0.5$ and MAF 0.5, with $n = 2000$.

The effects of LD on 2-SNP interactions were assessed by varying r^2 at one or both loci, with $r^2 \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Where $r^2 = 0$ at SNP_b, only the marginal effect of SNP_a is present; $r^2 = 1$ at both loci gives the original simulation in Figure 4.15.

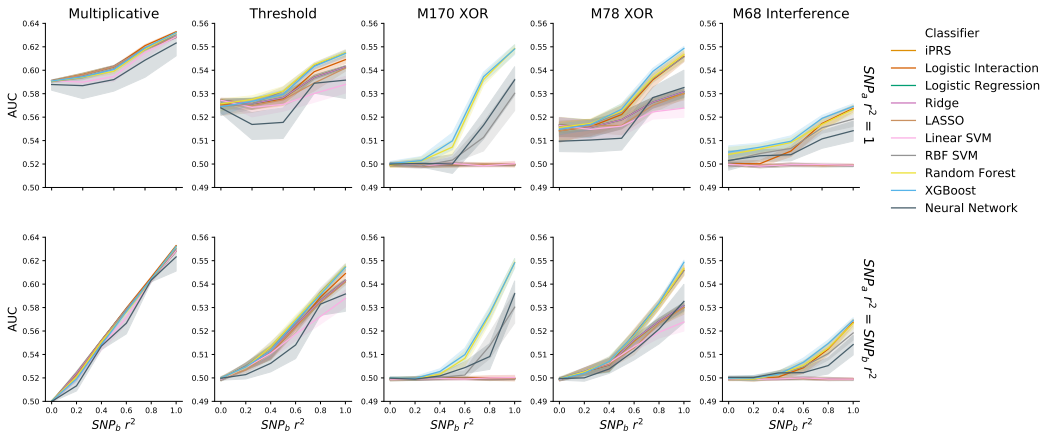


FIGURE 4.21: Degradation of classifier performance on 2-SNP interaction models when LD is introduced at one (top row) or both (bottom row) loci. All simulations fixed $n = 2000$, $\theta = 0.5$ and MAF 0.5. Y axis limits differ between subplots to allow for trends to be clear for all models.

Where LD between a single SNP and the causal variant is reduced from 1 to 0 (Figure 4.21, top row), discrimination of classifiers on multiplicative and threshold models decreases by around 2% but AUC is still maintained above 0.58 and 0.52 respectively. XOR models show a decrease of 4-5% AUC, dropping to around 0.5 and 0.51-0.52 for M170 and M78. Interference models similarly drop around 2% AUC to just above chance. All curves for classifiers show a sigmoidal relationship when decreasing LD at a single locus, so that when $r^2 = 0.8$ predictive performance is largely maintained, but at $r^2 = 0.6$ a large decrease in AUC occurs. By contrast, introducing LD at both loci results in a sharp linear decrease in AUC from between $r^2 = 0.6$ and $r^2 = 1$, with AUC for all models reduced to 0.5 when $r^2 = 0$. Linear models show AUC of 0.5 for all values of r^2 under M170 XOR and M68 interference models.

Increasing MAF at both loci is generally followed by an increase in discrimination, as shown in Figure 4.15, with the previously-noted exceptions of M170 XOR and M68 interference interaction models (Figure 4.22, bottom row).

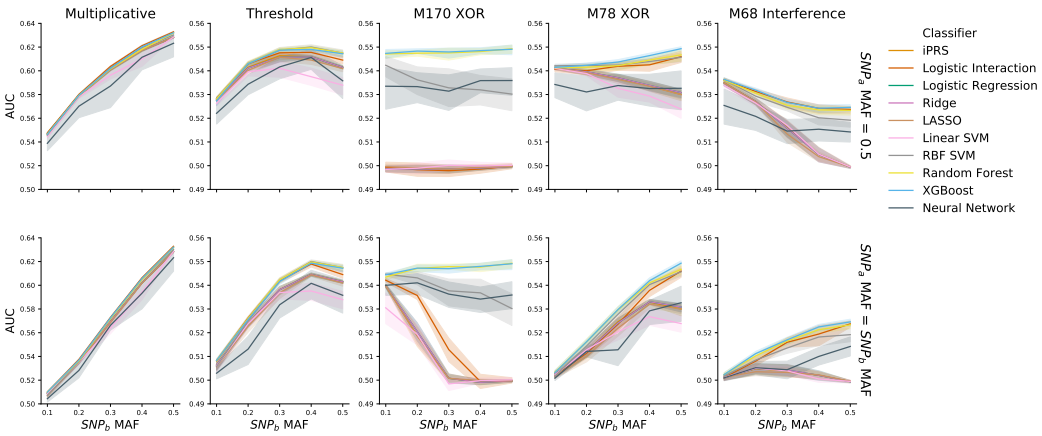


FIGURE 4.22: Reduction of classifier performance in response to decreasing MAF at one or both loci in 2-SNP interaction models. MAF at SNP_a was held at 0.5 when MAF at SNP_b was varied. All simulations fixed $n = 2000$, $\theta = 0.5$ and did not include LD. Y axis limits differ between subplots to allow for trends to be clear for all models; multiplicative models show high AUC on average.

Reduction of MAF at a single locus shows a similar decrease in performance for classifiers under multiplicative and threshold models, but with a smaller drop in AUC with decreasing MAF (Figure 4.22, top row); however, classifiers trained on non-linear M170 XOR, M78 XOR and M68 interference models show reasonably consistent performance despite changing MAF at one locus, and even decreasing AUC with increasing MAF under an interference model, previously visualised in Figure 4.19b.

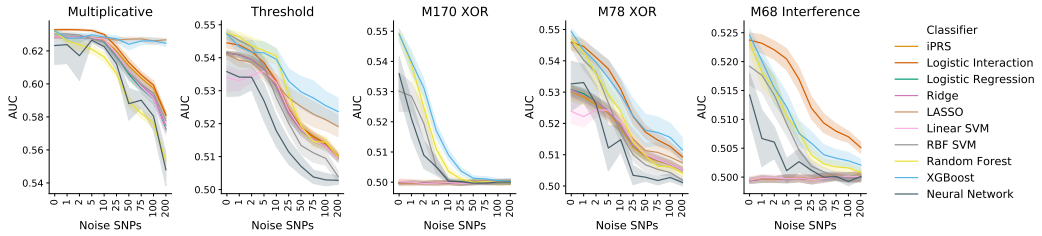


FIGURE 4.23: Decrease in discrimination for classifiers trained on 2-SNP interactions models with an increasing number of unassociated SNPs. Simulations set $n = 2000$, $\theta = 0.5$ and $MAF = 0.5$ with no LD. Unassociated SNPs are drawn randomly from the binomial distribution with two trials and chance of success equal to MAF, which is taken uniformly between 0.05 and 0.5. Y axis limits differ between subplots to allow for trends to be clear for all models.

Addition of noise to simulations through unassociated SNPs also causes discrimination to decay. Under multiplicative models, all classifiers except XGBoost and LASSO, which enforce sparsity, show decreased AUC with increasing number of noise SNPs. Addition of even a single unassociated SNP causes a decrease in AUC for all interaction models except multiplicative; performance of classifiers on M170 XOR was most affected by noise, dropping to around chance after the addition of 10-50 noise SNPs. XGBoost is generally the least affected by noise, but its performance is similar to logistic regression with an interaction term for M78 XOR and is exceeded by it for interference models. However, logistic regression

models with an interaction term give an unfair comparison as they only include an interaction term for the two associated SNPs; its use presupposes the interacting loci are known.

Across all simulations of interaction effects, discrimination remains low, with AUC typically below 0.6 and often closer to 0.5 despite accurate estimates of the decision boundary for some flexible models. This is expected given the use of only two SNPs showing small effect sizes. Logistic regression with an interaction term demonstrates competitive predictive performance under multiplicative and threshold models, but chance discrimination under M170 XOR models; iPRS does not show superior performance under any 2-SNP interaction models. Evaluation of decreasing test set size for all 3 simulation types indicate it does not contribute strongly to variance in discrimination (Figure B.5).

4.3.3 Neural networks

Computation time for training varied significantly between models, with neural networks displaying the longest training time. Larger neural network architectures were required to adequately learn interactions in a small number of epochs for the given learning rate and weight decay distributions used in hyperparameter tuning. Small architectures showed worse discrimination in the test set and less ability to fit approximate flexible functions (Figure 4.24). This effect was replicated across different simulation runs and interaction models (Figures B.6, B.7, B.8, B.9). Networks employing larger architectures and ReLU activation functions in hidden layers showed angular decision boundaries. Though smoother functions were learned with tanh activation functions, ReLU allows for faster computation of the gradients, and no significant corresponding improvement in AUC in the test set was observed when using tanh. Even with a modified architecture, neural networks show a higher variance that most other models, with occasional drops in average performance for a group of simulations due to low AUC in 1 or 2 iterations. This reflects their requirement for more careful hands-on tuning of network architecture and hyperparameters.

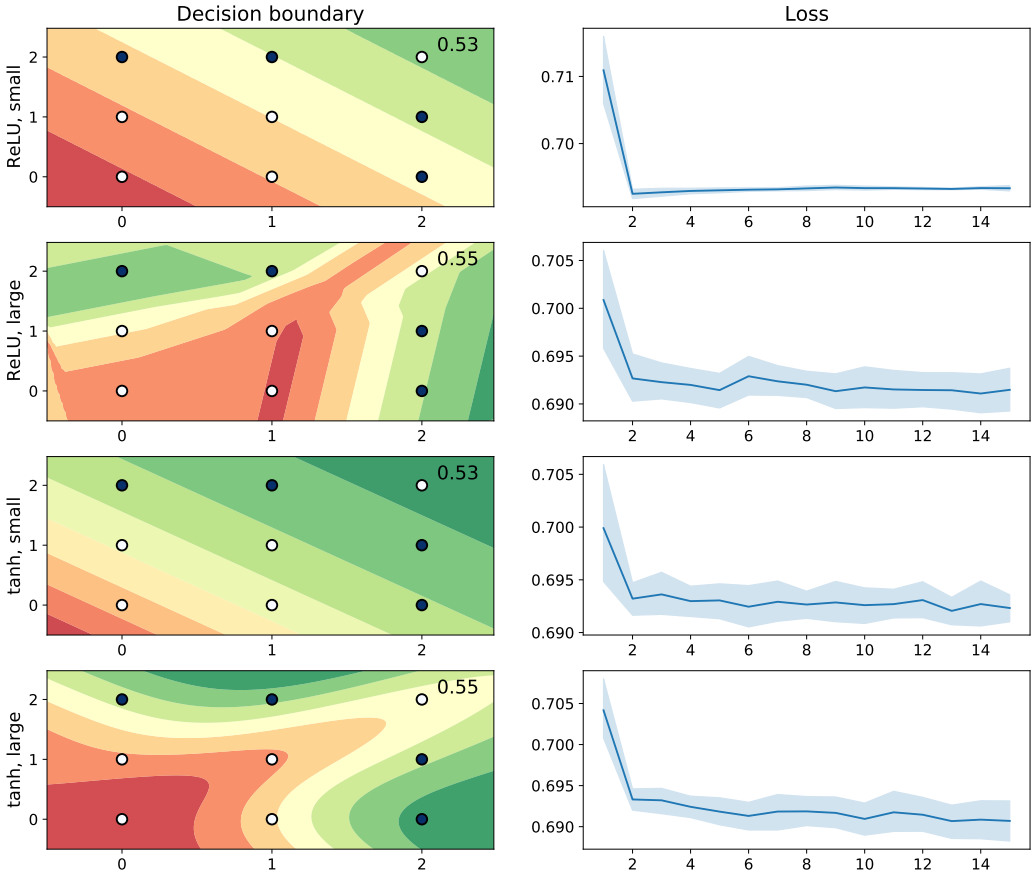


FIGURE 4.24: Examination of different neural network architectures on the decision boundary for M78 XOR models. The left-hand column gives the contour plots for the decision boundaries using either ReLU or tanh activation functions, and the small or large architectures given in section 4.2.3.2. The right-hand column shows the loss function, which should converge during the 15 epochs.

4.4 Discussion

4.4.1 Sparsity and flexibility

Two of the most fundamental machine learning principles separating results from classifiers are sparsity and flexibility. LASSO and ridge are illustrative of the former, as they apply either the L_1 or L_2 penalty to the same logistic regression model and so enforce a sparse or dense model respectively. Their results diverge most clearly in when a small number of associated variants are in the presence of a large number of unassociated SNPs, as seen for additive simulations when $m = 0.5$ (Figure 4.10), independent and additive simulations with increasing p for a fixed number of causal variants (Figure 4.12), and 2-SNP interaction simulations with increasing noise SNPs (Figure 4.15). These additive and independent simulations also describe the only scenarios in which machine learning models improved upon polygenic risk scores in the presence of additive effects, where LASSO, random forest and XGBoost all outperformed internal PRS (iPRS). This suggests that, under the assumption of additive effects, sparse machine learning methods may perform best in diseases with low

polygenicity and stronger effect sizes, while highly polygenic diseases where common variant associations are driven by many associations of small effect may be more favourable toward polygenic risk scores. External PRS (ePRS) represents the maximum achievable AUC and is not a fair comparison to machine learning approaches. However, the higher discrimination indicates the improvements in prediction that more accurate estimates of the true effect sizes may give.

The second main separating factor is the bias-variance trade-off, which is most easily viewed through the lens of model flexibility. The ability of models to learn flexible functions provided no advantage under additive simulations. This is clear through the comparison of LASSO with random forests and gradient boosting. All 3 machine learning approaches enforce sparsity, yet superior discrimination was consistently observed for LASSO in the presence of an additive model. By contrast, flexible approaches showed greater consistency in achieving higher discrimination in 2-SNP interaction models, with linear models failing to achieve AUC above 0.5 for M170 XOR and interference simulations. XGBoost and random forests also typically outperformed LASSO in the presence of noise.

4.4.2 Differences between classifiers

The effects of the bias-variance trade-off are also seen among linear models; logistic regression was often outperformed by other models which are capable of balancing bias and variance through choice of hyperparameters, including both linear and flexible models, particularly as p approached n. This confirms the importance of machine learning not just for $p > n$ scenarios, but also when $p < n$. Performance of logistic regression degraded when events per variable (EPV) fell below 2.5, with an almost 10% AUC decrease in performance compared to the best performing classifiers when EPV decreased to 1. Performance of logistic regression also contrasts with that from polygenic risk scores. Predictors in additive and independent simulations of main effects are truly independent, and any correlation between them incidental. It is therefore expected that iPRS scores, which include an additive combination of independently-estimated effect sizes, should slightly outperform multivariable logistic regression.

Ridge and linear SVM learn penalised linear models, but show differences in predictive performance. As detailed in chapter 2, the methods are connected: the hinge loss function used in SVMs is related to the logistic loss function (James et al., 2013), with both applying an L_2 penalty. Both methods also give greater priority to instances closer to the decision boundary, though SVMs enforce sparsity on observations through only using instances which prop-up the hyperplane margin. Differences in AUC between models may be due to sub-optimal hyperparameter tuning.

Random forests and gradient boosting generally perform worse than other models in independent and additive simulations, particularly where embedded predictor selection does not confer an advantage ($m \geq 0.5$). This confirms expectations that tree-based models will

perform poorly under additive effects (Hastie, Tibshirani, and Friedman, 2009). Though rare variants were not simulated, tree-based approaches would be expected to deviate from other models. For any rare variant V_j associated with the outcome where $V_j \in \{0, 1, 2\}$ and the risk allele homozygote (coded as 2) is not observed in the training data, a single cut-point will be learned at 0.5. Tree methods will therefore learn a dominant model. By contrast, a logistic regression assumes a linear predictor-response relationship and will learn an additive model whether the risk homozygote is observed or not.

4.4.3 Interaction and main effects

Flexible machine learning approaches are capable of learning complex 2-SNP interactions. This ability is largely unaffected by change in MAF at one locus for XOR and interference models. Only small reductions in performance are seen when r^2 at one locus is set to 0.8. However, decreasing MAF at both loci, or introducing more extreme LD at one or both SNPs causes a decrease in discrimination for most classifiers such that almost all show AUC of around 0.5 when MAF or LD drop to their lowest values in simulations, with the exception of multiplicative and threshold models, wherein classifiers show a more modest decrease in AUC in response to introduction of LD at one locus. The results imply that applying machine learning models to variants with low MAF does not necessarily mean interactions will not be incorporated into the model. Furthermore, selection of the causal variant or a tagging SNP in high LD is important for building prediction models with interaction effects, particularly when marginal effects at a locus are weak. In the presence of only additive effects, inclusion of multiple SNPs in LD with the target variant may improve prediction, with the caveat that this risks greatly increasing p for low r^2 thresholds.

Pairwise simulations alone are limited by the extent to which results can be applied to complex diseases, as intuitions around how methods behave may not apply in higher dimensions; findings from 2-SNP simulations do not necessarily generalise to datasets with a much larger number of predictors. Unassociated SNPs were introduced to evaluate the extent to which an increase in dimensionality and noise would drown-out signal from interactions. While classifiers trained on multiplicative and threshold models showed some resistance to noise, almost all learners showed AUC approaching chance as the number of noise SNPs increased. A potential criticism of using random forests in the presence of interactions is that they favour learning of marginal effects in high dimensions (Winham et al., 2012); however, all models here are impacted by high dimensions, with random forests among the best performing classifiers. Assessment of the decrease in discrimination with noise indicates that increasing the number of SNPs in a machine learning model without prior expectation of association may decrease predictive performance under both additive and interaction effects, and that the sparse approaches of LASSO and XGBoost should be used if such a scenario is unavoidable.

Though it is clear neural networks can compute complex interactions, for example, an M170 XOR model can be solved by hand as in Figure 4.7, representable is not synonymous with

learnable (Domingos, 2012). A large number of arbitrary functions can be represented by a neural network with a single hidden layer (Hornik, 1991) through choice of appropriate weights and architecture to compute the expected output. This does not imply that the corresponding weights and architecture can be learned from a real dataset. This is particularly true where the effect sizes of the predictors in a dataset are low. As such, application of a small network with 2 hidden layers and 2 hidden units often resulted in a linear or piece-wise linear function which failed to adequately learn XOR and interference models for the given learning rate and number of epochs. Deeper or wider models may be necessary to more easily learn complex interactions with neural networks, and more advanced hyperparameter tuning methods such as genetic programming, as proposed by Ritchie et al., 2003, may be helpful.

Classifiers were compared to logistic regression with an interaction term, which showed competitive performance under most scenarios except for M170 XOR models. However, though a useful comparison in assessing the flexibility required to learn an interaction, using multivariable prediction models with interaction terms for genetic datasets with a large number of predictors is not a practical solution. In simulations on adding noise SNPs to 2-SNP interactions, only a single interaction term was included. It is not feasible to include all pairwise interaction terms in a multivariable model for all unassociated variants when $n = 2000$; at 50 noise SNPs, 1,225 interaction terms are required, exceeding the number of cases, and at 100 noise SNPs 4,950 terms are required, far exceeding the number of observations. Further examining alternative models with higher-order interactions, or polynomial terms for dominance deviation at a single locus, exacerbates this issue (Frankel and Schork, 1996). Computational burden is also an issue when running pairwise tests for interactions under an explanatory modelling paradigm. Although strides have been made in increasing computational speed for this (Yung et al., 2011), typically only a multiplicative model is tested for. Prediction by a flexible machine learning model offers a more practical solution that avoids systematically considering all possible pairwise interactions and focuses on generalisation. XGBoost shows the strongest ability to detect interactions across models, MAF, LD, and in the presence of unassociated SNPs.

4.4.4 Randomness

Noise across simulations of training data comes from several sources. In all simulations genotypes are random draws from the binomial distribution. While MAF and effect size are fixed for interactions, they are randomly drawn in each iteration for additive and independent simulations. After selecting parameters for a simulation round, genotypes were assigned to create the sample directly. In real data individuals might be considered to be drawn from an underlying data generating process to create a population of individuals from which a sample is taken for study. This therefore includes an additional round of sampling, and so includes a further source of randomness not covered in independent or interaction simulations. In addition, simulations represent an ideal scenario of a single homogeneous population with a

single true effect size for each variant from which both training and test data are drawn. In reality, samples are heterogeneous. They may comprise several underlying populations with differing population parameters, with train and test data also possibly drawn from separate populations. Simulations may give an overly-optimistic view of the discrimination that is achievable. However, though they may simplify the true complexity of disorders such as schizophrenia, simulations allow for observations which would otherwise likely be obscured by the heterogeneity of real data.

Four main sources of randomness have been highlighted in evaluation of machine learning models (Dietterich, 1998). First, variation between different random draws of the test set, has been mitigated through simulation of independent test sets with large enough sample size that discrimination converges on the true AUC. Use of a separate simulation for test data also avoids any leakage of information that may occur in poorly devised resampling approaches, as observed in chapter 3. The second, variation between training sets, described previously, is one of the main causes of variance across simulations. Third, randomness in the machine learning model, is due here to both internal aspects of the algorithm - such as out-of-bag sampling in random forests, or weight initialisation in neural networks - and model selection through Monte Carlo random search. This is the second source of randomness in evaluation observed across simulations. The fourth, random mislabelling of observations which sets an upper limit on possible classification error, was not simulated here. In real learning problems for schizophrenia, it is expected that test error will also affect variance of predictions, particularly for rare variants and interactions, and that mislabelling between cases and controls, or between psychiatric disorders, may be present and further limit discrimination.

4.4.5 Simulation parameters

The study aimed to provide general guidance for common variants in psychiatric disorders. An important question is whether parameters are appropriate for schizophrenia in particular. In independent simulations, effects sizes and MAF mimic those observed for genome-wide significant SNPs in real data. In additive simulations, effect sizes varied with m, for fixed p and h^2. At $m = 0.05$, odds ratios on the observed scale for a small number of variants are large and deviate from empirical results, with odds ratios approaching more realistic values as m increases (Figure B.3). While additive simulations at $m = 0.05$ do not yield realistic effect sizes for common variants in schizophrenia, they provide an important explanation of why some machine learning methods are reported to do particularly well in other fields, or in diseases for which stronger effect sizes have been reported. Values for narrow-sense heritability on the liability scale ($h^2 = 0.2$) and prevalence ($k = 0.0025$) were fixed at low values for many simulations; estimates of h^2 for schizophrenia are closer to 0.26 (Anttila et al., 2018). Median point prevalence, period prevalence and lifetime morbid risk for schizophrenia have been estimated as 0.46% (0.19% - 1%), 0.33% (0.13% - 0.82%) and 0.72% (0.31% - 2.71%), with 10th and 90th percentiles shown in brackets (McGrath et al.,

2008). A recent estimate places point prevalence at 0.28% (0.24 – 0.31 95% CI) (Charlson et al., 2018). The value of $k = 0.0025$ used for most simulations, a prevalence of 0.25%, is within reasonable limits for point and period prevalence. Altering these parameters, for $h^2 \in \{0.15, 0.2, 0.25, 0.3\}$ and $k \in \{0.0025, 0.005, 0.01, 0.02\}$, showed patterns replicate across a range of heritabilities and prevalences. Though $h^2 = 0.2$ is below the 0.26 estimate for schizophrenia, it is above the variance explained for only genome-wide significant SNPs. PRS analysis using an independent PGC sample as the training set and CLOZUK data as the target using GWS SNPs found an AUC of 0.57 and R^2 on the liability scale of 0.011 (Pardiñas et al., 2018). Effect sizes in independent simulations are reasonable when p is low, but assigning the odds ratios found in genome-wide significant SNPs to all associated SNPs as p increases is likely to mean AUC is an overestimate when p is high. The most realistic simulations are therefore when p is high for additive simulations, where AUC is 0.606-0.669 for 1000 causal SNPs, and when p is low for independent simulations, achieving 0.546-0.563 AUC for 100 causal variants.

Simulations took a simplified approach to linkage disequilibrium, contrasting extremes. While those at $m = 0.05$ represent complete LD blocks, SNPs generated under higher values of m have more diffuse correlation. Simulations of LD with r^2 drawn uniformly from 0 to 1 and $m = 0.5$, where blocks were subsampled back to maintain p, may come closest to representing situations where genotyped SNPs are in LD with the causal SNP and correlated SNPs have been pruned. Classifiers in this scenario achieve lower AUC, with the majority of classifiers obtaining AUC between 0.565 and 0.6. Under additive simulations, performance is similar to creating LD blocks which are in constant moderate LD ($r^2 = 0.6$) with the causal SNP at $m = 0.5$.

Interactions set odds ratios for 2-loci combinations to between 1.1 and 1.5, except for the presence of risk alleles at both loci in multiplicative models. The extent to which interactions are present in schizophrenia is unclear, as testing for interactions requires large computational power and sample size and tests may assume only a single interaction model. Testing of GWS SNPs from PGC found no evidence of interaction effects (Ripke et al., 2014), but interactions with weak marginal effects may be missed by stringent p-value cut-offs in GWAS. Furthermore, evidence for interactions in schizophrenia has been found by several studies (Andreasen et al., 2012; Nicodemus et al., 2014; Guan et al., 2016). Of these, Guan et al., 2016, report odds ratios for schizophrenia risk for interactions at 0.47 (0.12-0.96), 2.63 (2.37-3) and 3.56 (3.27-3.98) for different genotype combinations between two SNPs with respect to the baseline in a Han Chinese population; 95% confidence intervals are given in brackets. Though it is unclear if these are typical values for interaction effects in schizophrenia, simulated effects span from 1.1, using $\alpha(1 + \theta)$ at $\theta = 0.1$, to 5.06, using $\alpha(1 + \theta)^4$ when $\theta = 0.5$, with the majority of interactions between 1.1 and 1.5 across simulations, indicating values chosen in simulations are not unreasonable. It should also be noted that the largest simulated odds ratios for double homozygotes for the risk allele (AABB) in multiplicative models are less likely to be observed in smaller samples

with low minor allele frequencies at both loci (Table B.1), and so are unlikely to greatly influence the decision boundary for multiplicative and threshold models. The issue of rarely observing combinations of genotypes with low expected frequency extends to other risk allele combinations under 3 or 4-loci interactions, which would require even larger sample sizes, or higher MAF, to reliably observe simulated odds ratios in both train and test sets, as well as within cross-validation folds in the training set.

4.4.6 Hyperparamter tuning

Hyperparameter tuning affected differences between models. More flexible models may benefit from a greater number of iterations of random search; low AUC seen for RBF SVMs (Figure 4.11, for instance) or large variance in AUC for neural networks (such as Figure 4.15) may resolve to a level similar to other flexible approaches with more tailored hyperparameter tuning. It is challenging to choose ranges or distributions for hyperparameters to enable fair comparison of flexible classifiers across datasets with different dimensionality and main or interaction effects. Such large-scale comparisons on heterogeneous data are less favourable toward flexible models. The chosen distributions and number of iterations produce machine learning classifiers which are good enough to detect patterns of behaviour in types of approach across simulations, but small differences, for example between ridge regression and linear SVM under an additive model, are likely to decrease with improved search.

4.5 Conclusion

This chapter turned to simulations for a fundamental understanding of the behaviour of machine learning approaches in the presence of main and interaction effects. This allowed investigation of changes in sample size, effect size, heritability, prevalence, and number of associated predictors while controlling the type of association. In addition, it enabled assessing of machine learning methods where bias was avoided by independently sampling test data from the same population parameters; it therefore compares PRS and ML without the suspected high risk of bias seen in the previous chapter, and addresses the lack of fair comparisons in the literature by applying PRS, LR and ML to all simulations. It demonstrated that machine learning methods are capable of learning complex genetic interactions, but that performance from such models is liable to decay under imperfect conditions. It also highlighted the expected dominance of PRS under additive effects, which is rarely challenged, and their poor performance on interactions. However, simulations are partly limited in that ePRS represents the unrealistic scenario of knowing the true effects in the population. Furthermore, AUCs of below 0.7 were typically achieved for more realistic simulations. The next chapter turns to real data from the UK Biobank to address these challenges, and whether combining genetic predictors with widely-available demographic factors can further improve prediction in machine learning models.

Chapter 5

Multivariable machine learning models of schizophrenia in UK Biobank

5.1 Introduction

Simulations provide a useful simplification, but real data may show more complicated relationships between predictors and outcome. The aetiology of schizophrenia in particular is complex, involving contributions from both environmental and genetic risk factors (Owen, Sawa, and Mortensen, 2016). As noted in chapter 1, many predictors in early life have been long-established (McDonald and Murray, 2000), while large strides in identifying common genetic variants associated with schizophrenia have been made through genome-wide association study (GWAS) consortia (for example, Pardiñas et al., 2018). Despite these advances, prediction from common variants through a polygenic risk score typically explains only a small proportion of the variance in liability to a psychiatric disorder. Narrow-sense (SNP-based) heritability for schizophrenia is estimated at around 0.26 (Anttila et al., 2018). For prediction in a new sample (CLOZUK) using a polygenic risk score (PRS) from an independent training set (psychiatric genetics consortium; PGC2), variance explained on the liability scale is just under 6% (Pardiñas et al., 2018); this amounts to a Nagelkerke's pseudo-R^2 of 0.12, and an area under the receiver operating characteristic (AUC) of 0.68. This supports the predictive ability in PGC2 (Ripke et al., 2014), for which an R^2 of around 0.18 and variance on the liability scale of about 7% were reported. Ripke et al., 2014, also provide results for a 40-fold leave-out-one PRS analysis of each of the datasets in the meta-analysis (mean 0.70 AUC, range 0.62-0.81). While an AUC of 0.7 may be considered high for genetic prediction in psychiatry, it only gives moderate discriminative ability and is not high enough to be clinically-useful alone. Clinically prognostic models using only traditional factors in cardiovascular disease, for example, have AUCs of over 0.8 (Lewis and Vassos, 2020).

A potential solution to this is to combine PRS with other predictors. Multivariable prediction models can be used to combine risk factors of complex disorders into a single model. Use of PRS with other factors is motivated by the notion that while common variants are not clinically predictive alone, their incorporation into a model with non-genetic factors may be. Prediction models of complex disease have been shown to benefit from the combined

modelling of genetics as PRS and traditional factors, particularly in cardiovascular diseases. Inouye et al., 2018, demonstrated a model where PRS alone (0.62 AUC) was combined with traditional risk factors (0.67 AUC) to give improved discriminative ability (AUC 0.70). Results in cancer studies have also suggested the combination of PRS and traditional factors may be predictive, though by using PRS with fewer single nucleotide polymorphisms (SNPs) (Fung et al., 2019), and a similarly PRS-augmented traditional model may also improve psychosis prediction in those deemed clinical high risk (Perkins et al., 2020).

Multivariable modelling efforts which are applied in this context typically make use of a linear model. However, alternative approaches using flexible machine learning models have the potential to detect more complicated relationships than a linear additive model, build models when the number of predictors outstrips the number of observations, and may enhance discriminative ability by optimising the trade-off between bias and variance for improved prediction. As with traditional statistical models, use of machine learning for multivariable modelling of risk factors has been applied to a range of diseases and disorders.

For schizophrenia and related outcomes, demographic, environmental and genetic predictors have been used in prediction studies applying machine learning (ML). The earliest employed a neural network to predict response to clozapine using SNPs alongside demographic and physical data (Lin et al., 2008); the authors report that a neural network (0.82 AUC) outperformed logistic regression (0.58 AUC) in this study. Demographic information, SNPs, predictors covering physical characteristics and lifestyle factors such as smoking were also used to predict weight change in response to medication using a neuro-fuzzy model (Lan et al., 2008) which combines neural networks with human-like fuzzy logic. The authors report the ML model outperforming logistic regression for sensitivity but not specificity. Both studies evaluated a multivariable model without comparison to models with a single data type (genetic-only or environmental-only), making it unclear whether genetic information improved prediction.

Later efforts have focused more on neuroimaging. Genetic, imaging and cognitive predictors showed variable accuracy when used by a support vector machine (SVM) to differentiate healthy, ultra-high risk and first episode psychosis individuals (Pettersson-Yeo et al., 2013), but for which the added value of data types was also not examined. Machine learning using random forests found no added value for combining PRS with cognitive and neuroimaging data, despite modest prediction from PRS alone (Doan et al., 2017). Several methodological studies have focused on development of sparse or hybrid models for functional magnetic resonance imaging (fMRI) and SNP data, where their combination was either not compared (Li et al., 2020) or was shown to improve prediction (Yang et al., 2010a; Cao et al., 2013). Prediction of schizophrenia-related traits has also been evaluated by using schizophrenia-associated SNPs to predict cognitive measures with random forests (Zheutlin et al., 2018).

Chapter 3 reviewed genetic prediction of psychiatric disorders, identifying variable predictive abilities of machine learning approaches in schizophrenia (0.54-0.95 AUC) and demonstrating

high within-study risk of bias (ROB). ROB has not been similarly assessed for the models combining genetic and other predictors described above, but given widespread sub-optimal practices, caution should be taken when interpreting the measures of discrimination and classification reported in such studies. Further scepticism should also be brought to studies reporting novel methods, such as those developed to combine fMRI and SNP data, as these may be more favourable toward the newly-developed technique than objective comparison studies (Boulesteix, Lauer, and Eugster, 2013).

Chapter 3 revealed that reporting results alongside logistic regression and PRS, or evaluation of models with SNPs as independent predictors or combined in a PRS, have been largely absent from the literature. Furthermore, it highlighted a dearth of investigation of models for potential confounders, with no recommended strategy available for machine learning in psychiatric genetics.

5.1.1 Aims and objectives

The aim of this chapter is to report the development of multivariable machine learning models for the prediction of schizophrenia using data in the UK Biobank. The objectives are:

- Evaluate statistical and machine learning approaches for their discrimination and calibration
- Compare prediction using genetic, demographic or combined predictors
- Investigate model predictions for potential confounders and generalisable associations
- Assess subsampling approaches to class imbalance in a large dataset

Linear and flexible modelling approaches were compared in a nested case-control design in the UK Biobank using common variants and demographic predictors.

5.2 Methods

5.2.1 Participants

The UK Biobank is a large prospective cohort of around 500,000 individuals aged 40 to 69 that were recruited between 2006 and 2010 and assessed at 22 centres across the UK. Participants gave consent, provided physical and cognitive measures, and completed a touch-screen questionnaire with inputs checked by a nurse. They supplied blood, saliva and urine samples (Sudlow et al., 2015). This study is part of UK Biobank project 15175. Ethical approval was given by the North-West Multi-Centre Research Ethics Committee.

5.2.2 Genetic quality control

Participants were restricted to unrelated individuals with a kinship coefficient of less than 0.2; one individual was randomly removed from each related pair, with preference given to keeping affected participants. The sample was further restricted to those who self-reported as white British or Irish (field 21000). Individuals were removed if consent was withdrawn as of 17/03/20. Genetic data was imputed by the UK Biobank (Bycroft et al., 2018). Imputed SNPs from the Haplotype Reference Consortium (HRC) were kept if they had Hardy-Weinberg equilibrium less than 10e-6, minor allele frequency greater than 1%, imputation quality score greater than 0.4 and posterior probability greater than 10e-4. Genotyping and imputation have previously been described by the UK Biobank (Bycroft et al., 2018).

5.2.3 Outcome

Schizophrenia was defined by the presence of international classification of diseases (ICD)-10 codes for schizophrenia (codes F200-F209) or schizoaffective disorder (codes F250-F259) in primary or secondary hospital records (fields 41202 and 41204) or death records (fields 40001 and 40002), or if self-reported (field 1289) in any of the first three timepoints (initial assessment, first repeat assessment and imaging visit). Participants with no occurrence of these were considered as controls, excluding those with other psychoses (codes F21-23, F28, F29) in hospital and death records or bipolar disorder in hospital and death (codes F30-31) or self-report (code 1291) fields. A total of 807 cases and 370,468 controls remained before subsampling.

5.2.4 Predictors

5.2.4.1 Genetic predictors

31,603 nominally significant and 116 genome-wide significant (GWS) SNPs (Pardiñas et al., 2018) were used as dosages from imputed genotypes after pruning for linkage disequilibrium (SNPs with strongest evidence of association preferentially kept, $r^2 = 0.2$, distance 1Mb; Figure C.3). The 0.05 threshold is the most predictive for schizophrenia (Ripke et al., 2014; Pardiñas et al., 2018; Ripke et al., 2020), while GWS SNPs give contrasting dimensionality, and consequently faster computation, yet still account for around half the variance explained on the liability scale to schizophrenia in a PRS (Ripke et al., 2014; Ripke et al., 2020).

For both the nominally significant and genome-wide significant SNPs, two types of PRS were created to evaluate the role of prior information. PRS were either generated using weights from summary statistics of an external dataset (Pardiñas et al., 2018), known here as external-PRS (ePRS), or from univariable association tests in the training fold which were performed in cross-validation, termed internal-PRS (iPRS). Where an ePRS was used, genotypes were thresholded (imputation quality > 0.9) and combined into a PRS, with any subsequent adjustments for transformations performed within cross-validation, to match the

methodology used for individual SNPs. ePRS were computed using PLINK (Purcell et al., 2007).

PRS using weights generated within each training fold of cross-validation (iPRS) were derived by performing univariable linear association tests for each SNP in the training fold, then taking the dot product of the resulting effect sizes and the genotypes in the training and test folds to produce a PRS without prior (external) weights. A scikit-learn-compatible transformer was created to ensure coefficients from the training fold were stored and applied to the test fold correctly; this allowed the transformer to be incorporated into a standard pipeline. PRS using external weights of all nominally-significant SNPs are referred to as 0.05 ePRS, while those using only genome-wide significant SNPs with external weights are labelled GWS ePRS. To ensure experiments were able to complete, given the large number of models and training time needed for gradient boosting and neural networks, comparison of SNPs, iPRS and ePRS was limited to GWS SNPs only. Nominally significant SNPs were only considered as an ePRS.

5.2.4.2 Deconfounding

To account for the linear effects of genetic ancestry and genotyping array, SNPs used as independent predictors were regressed against 40 principal components, provided by UK Biobank, and array type, with residuals z-transformed. Several procedures are available to apply a linear adjustment for genetic confounders in this way during model development (Figure 5.1) or after (Dinga et al., 2020), in addition to more complex modified neural network architectures to select features which predict the outcome independently of confounders (Zhao, Adeli, and Pohl, 2020). So-called "deconfounding" may be performed before cross-validation; this has previously been applied in genetic prediction models (Zhao et al., 2012; Zheutlin et al., 2018). A recommended alternative, outlined by Chyzhyk et al., 2018, is to apply deconfounding within cross-validation in a prediction modelling paradigm, such that a separate fit on both training and test folds is avoided as potentially causing overly pessimistic results and requiring new observations to always appear in groups, instead favouring a single fit on the training fold and prediction in the test fold, shown in Figure 5.1c. This approach is used here and applied in all main analyses. Labels of prior, double and single are used to refer to deconfounding performed before cross-validation (Figure 5.1a), or using two separate fits (coefficients estimated separately in training and testing folds; Figure 5.1b) or one fit during cross-validation (coefficients estimated once in training fold; Figure 5.1c) respectively.

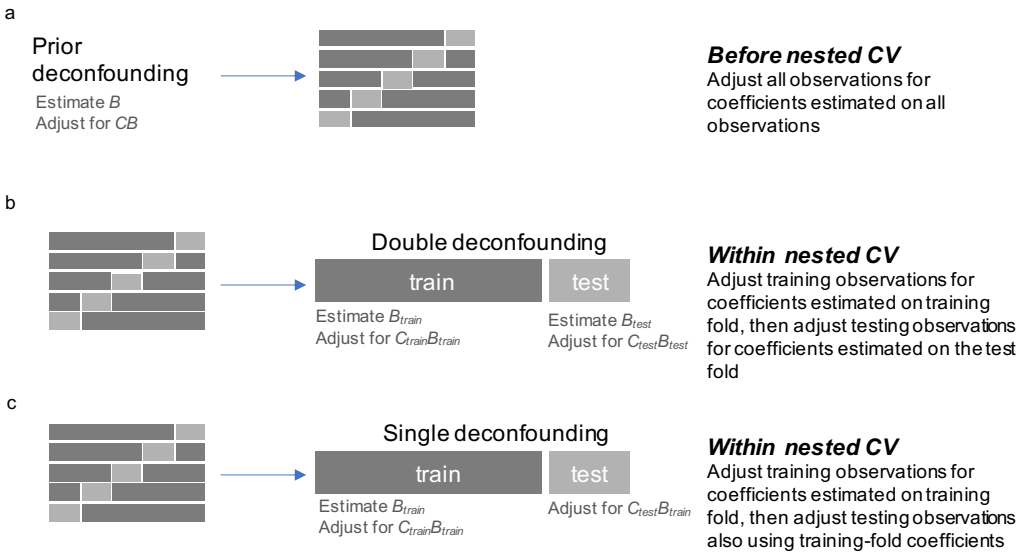


FIGURE 5.1: Deconfounding procedures. A common approach is to use a linear regression to remove the linear effects of confounding. Often this is done on data before cross-validation (a). However, it is preferred to keep the step within cross-validation to avoid inadvertently using information from the test folds to modify predictors in the training fold, and to keep train and test folds independent. This may be done by creating two separate regression models for train and test folds for each predictor in a train-test pair within cross-validation, leading to two sets of coefficients for each predictor (b). The alternative, which is used here as it sticks more closely to cross-validation principles, is to create a single regression model for each predictor in the training fold, and use the same coefficients for the test fold, for each train-test fold pair within cross-validation (c).

For any train-test fold pair within cross-validation, an ordinary least squares (OLS) linear regression is built for each SNP or PRS in the training set, as

$$X_{train_j} = \hat{\beta}_{train_{0j}} + \hat{\beta}_{train_{1j}} C_{train_1} + \cdots + \hat{\beta}_{train_{kj}} C_{train_k} \tag{5.1}$$

for k confounders and j predictors, where C_{train} is an $n_{train} \times k$ matrix of confounders in the training fold and X_{train} is an $n_{train} \times j$ matrix of genetic predictors for observation in the training fold. All coefficients except the intercept for all OLS models are then stacked into a matrix B_{train}, so that $B_{train} \in \mathbb{R}^{k \times j}$, and is stored for use in both training and test folds. To apply deconfounding to the training fold, the dot product of C_{train} and B_{train} is then taken, so that a new $n_{train} \times j$ matrix of residuals is defined as

$$R_{train} = X_{train} - C_{train} \cdot B_{train}. \tag{5.2}$$

This is the new matrix of residuals for the training fold. The same procedure is applied to the test fold, using the stored coefficient matrix B_{train} from the training fold, as

$$R_{test} = X_{test} - C_{test} \cdot B_{train}, \tag{5.3}$$

where R_{test} and X_{test} are $n_{test} \times j$ matrices and C_{test} is the $n_{test} \times k$ matrix of confounders for the test fold. Residuals are then z-transformed in each train-test fold pair using statistics from the training fold. A scikit-learn compatible transformer was created to incorporate this method into pipelines. Methods essential to scikit-learn's RandomizedSearchCV class and the cross-validate function were also modified to allow for setting and checking the correct confounder index for train and test folds within cross-validation. As the main analysis was performed using self-reported ancestry, results for deconfounding were also re-analysed after reducing participants to a homogeneous cluster of genetic ancestry using principal components (field 22006 from UK Biobank).

5.2.4.3 Demographic predictors

Participants self-reported answers to a touchscreen questionnaire at assessment centres. Predictive models included 6 variables derived from information reported at baseline. Number of siblings (UK Biobank fields 1883 and 1873), handedness (field 1707), severe parental depression (fields 20107 and 20110), educational attainment (field 6138), season of birth (derived from field 52) and sex (field 31) were used in prediction models. Demographic predictors were selected if they showed evidence for prior association in the literature and mostly complete records. Factors associated with schizophrenia surrounding obstetric complications and trauma were absent in too many participants to be included, while measures such as deprivation, body mass index (BMI), smoking and cognitive measures are partly a consequence of schizophrenia and its treatment, and so were not included. Included demographic predictors benefit from being easily-collected variables for events which largely occur before disorder onset.

Lower educational attainment and cognitive impairment are associated with schizophrenia (Rajji et al., 2013), with education often taken as a proxy phenotype for cognition. Though published associations with schizophrenia in UK Biobank were not used to select predictors, as this would introduce bias in validation, educational attainment has also been associated with genetic liability to schizophrenia within the UK Biobank (Escott-Price et al., 2020) and shared loci for education and schizophrenia have been identified in PGC2 (Le Hellard et al., 2017). Association between non-right handedness and schizophrenia has been reported by many, and is suggested to occur via abnormal brain lateralisation. Two early meta-analyses confirmed an excess of mixed handedness in schizophrenia (Sommer et al., 2001; Dragovic and Hammond, 2005), with a more recent meta-analysis offering support that the relationship is genuine (Hirnstein and Hugdahl, 2014).

Sex differences are robustly associated with schizophrenia, affecting many aspects of the disorder, including prevalence, onset and response to treatment (Aleman, Kahn, and Selten, 2003; McGrath et al., 2008). Maternal depression during pregnancy has been associated with schizophrenia (Jones et al., 1998), possibly through obstetric complications and post-natal depression (Verdoux and Sutter, 2002), though more recent analysis in the same data reports elevated risk for schizophrenia only in the presence of genetic risk for psychosis

(Mäki et al., 2010). Weaker relationships have also been reported for number of siblings with schizophrenia (Wahlbeck et al., 2001) and schizophrenia-associated outcomes (Üçok and Bıkmaz, 2007). Season of birth is a well-studied association with schizophrenia whereby increased risk is observed in those born during winter in both hemispheres; stronger effects are seen in the Northern hemisphere (McGrath and Welham, 1999; Davies et al., 2003).

Importantly, such factors are typically assessed for their association with schizophrenia under an explanatory modelling paradigm, which does not guarantee generalisation (Bzdok, Engemann, and Thirion, 2020); here, they are assessed only for prediction.

Number of siblings was counted by summing number of full brothers (field 1873) and number of full sisters (field 1883), curtailed at 10 and log transformed. Handedness was binary-coded as 1 for right-handed and 0 for left-handed or ambidextrous. Additive coding was used for severe parental depression, with 1 for a single affected parent and 2 if both parents were affected. Educational attainment is coded as 1 for GCSE or higher, and 0 for lower. Season of birth is coded as 1 for winter birth (December to February inclusive) and 0 otherwise. Sex was kept binary, as coded by UK Biobank. All items selected by participants as 'Prefer not to say' were coded as missing.

5.2.5 Study design

5.2.5.1 Model development

Genetic and demographic variables were considered both individually and together for each modelling method. Analysis followed a nested case-control design, where a ratio of 1:5 of cases to controls was chosen for computational efficiency, in keeping with recommendations in epidemiology (Biesheuvel et al., 2008). Participants missing demographic predictors in each dataset were excluded prior to randomly subsampling controls. Models were trained and evaluated using 10-fold nested cross-validation (CV), an approach involving hyperparameter tuning on inner folds before evaluating on outer folds using the best-performing hyperparameters, as this has been shown to approximate the true error rate and give more accurate estimates of prediction performance when compared to split-sample or non-nested cross-validation (Varma and Simon, 2006; Vabalas et al., 2019). k was set to 10 for both inner and outer folds. Identical training folds were maintained across classifiers by passing the same random seed to scikit-learn's StratifiedKFold class, ensuring all methods performed nested CV on the same data splits. For all methods, hyperparameters were tuned using Monte Carlo randomised search, which is more efficient than grid search when many hyperparameters require tuning (Bergstra and Bengio, 2012). 100 values of hyperparameters were randomly drawn from the chosen distributions for each hyperparameter to ensure thorough hyperparameter search (Figure C.1). Deconfounding and transformation of predictors were performed within each fold of cross-validation, with deconfounding following the procedure in Figure 5.1c, using a scikit-learn pipeline to avoid 'leakage' of information from the test fold. The most predictive models were evaluated for predictor importance and generalisable

associations. In evaluating 9 classifiers on 9 datasets, using 10x nested CV and 100 iterations of random search, 810,000 models were trained for the main analysis alone.

5.2.5.2 Discrimination

Discrimination was evaluated for all models using the area under the receiver operating characteristic curve. The median AUC across CV folds is presented alongside the distribution for each model. Machine learning classifiers underwent pairwise comparisons with logistic regression using AUCs from the outer round of nested CV. As detailed in chapter 2, though the paired *t*-test is still commonly used, it displays a greater chance of type I error (Dietterich, 1998); instead, the non-parametric Wilcoxon signed-rank test is used for all pairwise comparisons (Demšar, 2006). All tests were conducted with a two-sided alternative hypothesis. Due to correlation between tests, multiple testing was accounted for using the Benjamini-Hochberg false discovery rate (FDR) at 0.1.

5.2.5.3 Calibration

Models were re-calibrated to allow for fair comparison of predicted probabilities, as SVM output is distance from the hyperplane, and the structure of trees forces random forests and gradient boosting to push predicted probabilities toward 0.5 (Niculescu-Mizil and Caruana, 2005); such max-margin-based methods typically produce sigmoidal-shaped calibration plots which can be appropriately corrected by re-calibration using a logistic regression. To ensure bias was not introduced in calibration, Platt scaling was used in cross-validation (Platt et al., 1999), where the test-set predictions of the base model from the inner-fold of cross-validation for the best performing hyperparameters is used to fit a logistic regression, which then takes predicted outer-fold test set probabilities from the base model as the input to give re-calibrated probabilities. Recalibration changes the predicted probabilities but will not change AUC, as this is a rank-based measure and calibration here applies a monotonic transformation.

Calibration was assessed graphically using loess smoothers (Austin and Steyerberg, 2014), which were extended to be "validation plots" by including the distributions of predicted probabilities (Steyerberg et al., 2010), and using Brier's score.

Calibration, predictor importance and confounding were further assessed in the best performing linear and non-linear modelling approaches across datasets. To rigorously assess calibration, plots were also generated by predictor and confounder subgroups where categorical variables were present (Steyerberg et al., 2019). To show calibration for models on the whole of UK Biobank, predictions for the unsubsampled controls not included in the nested design were generated by refitting models on the nested dataset before predicting on all remaining controls which were previously not included when subsampling. Predicted probabilities from the outer fold of nested CV were then combined with this so that all observations had a risk of schizophrenia assigned by each machine learning classifier.

Predicted probabilities were then adjusted for the sampling fraction using Elkan's transformation to alter probabilities where the base rate, or prevalence, differs between groups, as it does between the nested case-control sample and the full cohort (Elkan, 2001). Validation plots were previously unavailable in python and can be accessed through GitHub (https://github.com/seafloor/validation_plot).

5.2.5.4 Predictor importance

Predictor permutation importance (also called permutation feature importance) scores were used to estimate the relative contribution of each predictor to the model in a model-agnostic manor (Molnar, 2019) in the most performant models. Permutation importance expands on the idea developed by Breiman for random forests (Breiman, 2001) by applying permutations to the whole model, rather than to subgroups of trees. It was used in preference to built-in measures of predictors such as regression coefficients or hyperplane weights to enable consistent interpretation across different model types. Though training-set and test-set permutations may both be used (Molnar, 2019), here test-set permutations were chosen to focus on generalisable contributions to the model. For each predictor, and for each outer fold of nested CV, the test-set predictors were randomly shuffled and fed into the model 10 times. The drop in discrimination compared to the full (unshuffled) model is taken as the importance score, with the median of these for the 10 permutations taken forward. The final reported value is the median of these across the 10 outer-CV folds.

To give an estimate of the importance of groups of variables, the same method was applied using a predictor mask, where variables are assigned to different groups, and all predictors in a group are shuffled together. The subsequent drop in AUC gives an estimate of the relative importance of all the permuted predictors taken together. This technique is used in deep learning, for example (Kokhlikyan et al., 2020), and was implemented here by extending scikit-learn's "permutation_importance" function. A mask was applied to put genetic and demographic predictors into separate groups and estimate their relative importance.

5.2.5.5 Generalisable associations of model predictions

Predictions from the most performant models were further investigated for confounding and generalisable associations. Chapter 3 identified that no previous studies of ML prediction in psychiatric genetics had investigated models for population structure. Understanding model confounds is a complex issue; bias in machine learning is an ongoing area of research with no single solution (Mehrabi et al., 2019). However, as ML is focused on prediction over explanation, a recommended protocol for researchers is to investigate how well potential confounders predict either the outcome itself or the predictions from a model (Kohoutová et al., 2020). Assessment of models therefore used a regression model of UK Biobank variables to predict risk scores from machine learning models in 5-fold cross-validation. To ensure the bounded nature of predicted probabilities was correctly modelled, a beta regression was used. This assumes dependent variables are beta distributed, and is appropriate for use

where the response is between 0 and 1, such as with proportions, rates and fractions. As with generalised linear models, a link function is used, here to connect the linear predictor to the mean outcome. Models were fit using the response, y_i, as the risk for schizophrenia assigned by a machine learning model for the ith observation, and a matrix $C \in \mathbb{R}^{n \times p}$ of all p independent variables and n observations. C may be a matrix of confounders or other variables to be investigated for generalisation with the model. Principal components considered together, or combined with array type, were evaluated in unsubsampled controls.

Machine learning risk scores were also considered for how they were predicted by cognitive tests, neurological diseases, psychiatric disorders and widely-measured demographic variables in unsubsampled controls. Mean variance explained (R^2) was assessed between all test-fold predictions from a 5-fold cross-validated beta regression and the risk assigned by the model of interest to give an estimate of how well risk scores from machine leaning models could themselves be predicted using potential confounders and variables of interest. Cognitive, demographic and disease variables used in assessing model correlations are described in appendix C.1.2.

5.2.6 Statistical methods

5.2.6.1 Algorithms

Common linear and non-linear machine learning methods were compared for all datasets and predictors. Least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) and ridge regression (Hoerl and Kennard, 1970) are commonly-used penalised regression methods for genetic prediction. The LASSO (L_1) and ridge (L_2) penalties were applied to logistic regression.

Support vector machines (Cortes and Vapnik, 1995) take a kernel-based approach to learning, using similarity measures to fit a maximally-separating hyperplane to distinguish between classes (Noble, 2006). Linear SVMs maximise the margin either side of the hyperplane and control the porousness of the plane to violations using a hyperparameter, 'C'. Non-linear kernels are available which compute similarity between observations for higher-dimensional space without the need for transforming predictors. Radial basis function (RBF) SVMs are a popular choice, and include an additional tuning parameter, γ, for controlling how localised the influence from support vectors is on the hyperplane. Both linear and RBF-kernel SVMs were applied.

Ensemble methods, which combine multiple 'weak' learners together into a strong learner, are commonly used as they perform well in many domains. Random forests (Breiman, 2001) combine classification and regression trees which make recursive binary splits to partition observations (Breiman et al., 1984). Maximum depth of each tree, maximum number of predictors to search at each split, and minimum number of observations in the child nodes to allow splitting to occur were searched during hyperparameter tuning to reduce tree complexity and control overfitting. Number of trees was kept constant at 1000; the Gini

index was used as the splitting criterion. In contrast to random forests, gradient boosting machines (GBMs) sequentially optimise against the gradient of an objective function (Friedman, 2001). Extreme gradient boosting (XGBoost) is a highly-regularised implementation of gradient boosting which has been thoroughly optimised (Chen and Guestrin, 2016) and combines several useful features into a single package. XGBoost was assessed using decision trees as weak learners and a logistic objective function. Learning rate, maximum depth of trees, regularisation (L_2 penalty) and subsampling observations and columns (by tree) were searched during training. Number of trees was kept at constant at 1000.

Neural networks also make use of sequential optimisation of an objective function by applying a network of weights to the input which are iteratively updated through backpropagation to learn associations with the outcome (LeCun, Bengio, and Hinton, 2015). They show different abilities and properties depending on the architecture of the network. A fully-connected feed-forward network, or multi-layer perceptron (MLP), was used with batch norm, logistic loss and rectified linear unit (ReLU) activation function in all layers apart from the last, which uses a sigmoid function. Layer initialisations followed PyTorch defaults. Number of hidden layers and units per hidden layer were searched in training. Models were optimised using mini-batch gradient descent with momentum set to 0.9 and variable learning rate and weight decay (L_2 penalty). Batch size and number of epochs were kept constant at 32 and 15 respectively.

Stacking ensembles combine the output of several models by 'stacking' the output of base models together to create a new set of predictors (Wolpert, 1992). A meta model is then trained on these to decide the optimal weighting from each of the base models. A meta logistic regression was trained through stacking to combine the outputs of all models. All machine learning methods were also compared to an unpenalised logistic regression trained on the original predictors.

5.2.6.2 Software

Logistic regression, support vector machines and random forests were implemented in scikit-learn version 0.22.1 (Pedregosa et al., 2011). Unpenalised logistic regression for prediction modelling was implemented in scikit-learn using a newton solver with penalty set to 'none', so as to be equivalent to regression run in the R statistical programming language. Gradient boosting used the scikit-learn API for the python XGBoost package, version 1.1.1 (Chen and Guestrin, 2016). Neural networks were implemented in PyTorch version 1.5 via a custom wrapper which allowed scikit-learn pipelines to be applied. To speed computation and allow for a large randomised search, all analyses were performed on the Cardiff Hawk supercomputer, with neural networks run separately using Nvidia V100 and P100 graphical processing units (GPUs). Statsmodels 0.10.1 (Seabold and Perktold, 2010) and NumPy were used to adjust predictors for principal components and genotyping array, and to perform association tests between phenotypes and risk scores. PLINK version 1.9c3 was used to generate PRS (Purcell et al., 2007). Tests comparing machine learning models on the same

dataset were implemented in SciPy version 1.4.1 (Virtanen et al., 2020). Preprocessing and search used pandas 1.03 (McKinney, 2010; The pandas development team, 2020), NumPy 1.18.1 (Walt, Colbert, and Varoquaux, 2011) and SciPy. Adjustment for multiple testing using FDR was done with the fdrcorrection function in Statsmodels. Beta regression was run with the betareg R package version 3.1-3 (Cribari-Neto and Zeileis, 2009), which was called from Python using rpy2 version 3.1.0. Missingness plots were partly created using the missingno package (Bilogur, 2018). All other plots were created using matplotlib 3.2.1 (Hunter, 2007) and seaborn 0.10.1 (Waskom et al., 2020).

5.3 Results

5.3.1 Samples

371,275 individuals were retained for analysis after sample QC. (Figure 5.2). Controls were filtered for missingness and randomly subsampled in a nested case-control design of 1:5 cases to controls to give 738 cases and 3,690 controls (sampled from 341,822) for demographic data, corresponding to a sampling fraction of 0.01 for controls. Cases consisted of some self-report only, some hospital records only, and some participants with both hospital and self-reported schizophrenia (Figure C.2).

Comparison of demographic information in cases before and after excluding those with missing values (Figure C.4), or controls before and after excluding by missingness (Figure C.5) and subsampling (Figure C.6) indicate observations used in analysis are similar to the full UK Biobank cohort. By contrast, comparing the same characteristics between the analysed cases and controls shows differences for age, sex, BMI, deprivation and genotyping array (Figure C.7).

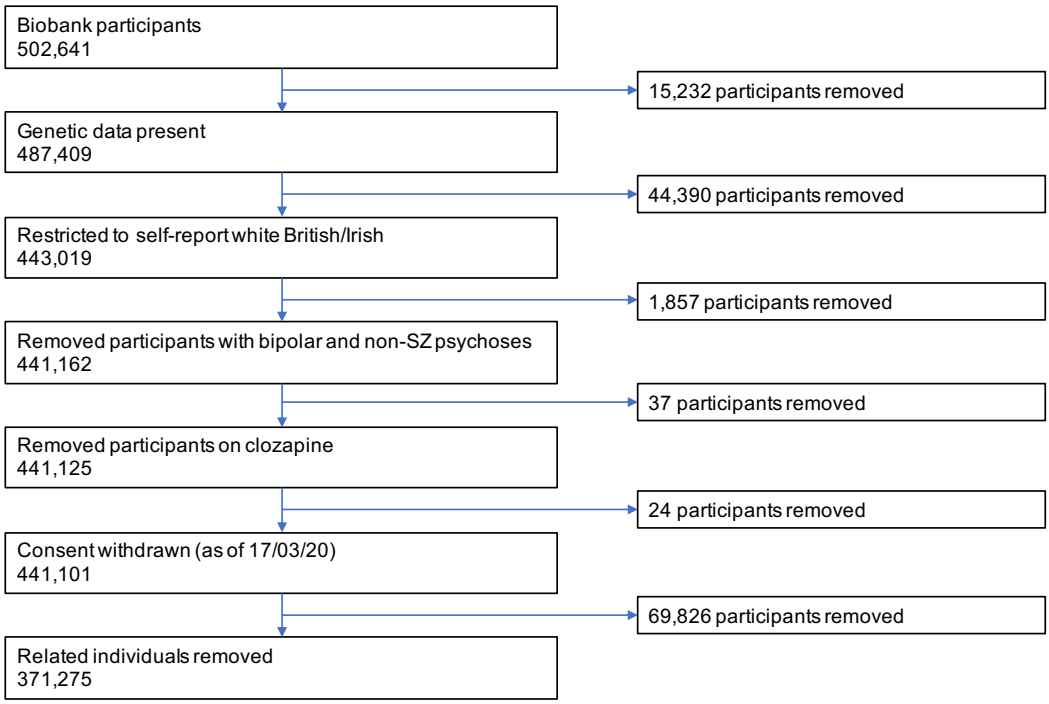


FIGURE 5.2: Workflow for UK Biobank participants before nested case-control and missingness filters. Number of participants removed reflects only that stage in the protocol; a larger number of participants may meet the criteria in the total sample. Individuals on clozapine were removed to reduce the likelihood of sample overlap with the CLOZUK dataset, used to produce the 0.05-threshold ePRS; overlapping controls may still be present.

Missingness was evaluated for demographic predictors only, as SNPs were imputed. Before subsampling, 39,788 participants contained missing values for at least one predictor, consisting of 39,675 controls and 113 cases, which were not used in later analyses. Tests were applied to each predictor using a binary coding for presence of missingness and a chi-squared or Fisher's exact test to evaluate association between missingness and case-control status. Two demographic predictors (number of siblings, parental depression) indicate strong evidence for association between missingness and case-control status (Table C.2), with presence of missingness also highly correlated between these predictors (Figure C.9). Remaining predictors show weak (education) or no evidence (sex, handedness, winter birth) for association between missingness and presence of schizophrenia (Table C.2). Further comparison of cases with and without missingness show almost identical distributions for age, BMI, deprivation, sex, array and the first two principal components (Figure C.4), indicating included cases are otherwise representative of all cases, after exclusions, in UK Biobank. Per-predictor missingness is given in appendix C (Figure C.8).

5.3.2 Predictive performance of machine learning methods

5.3.2.1 Discrimination by modelling approach

Methods were compared across all datasets for discrimination (Figure 5.3). Prediction using machine learning approaches showed small differences from logistic regression for some

models after correction for multiple testing (Table 5.1). However, no modelling approach showed consistently better results than any other. Tests could not be run for ridge, LASSO and linear SVM for PRS-only models as they returned identical AUCs to logistic regression. Models were also compared with and without educational attainment as a predictor, as education may not be completed before onset of schizophrenia in some individuals, but showed no difference in discrimination (Figure C.10).

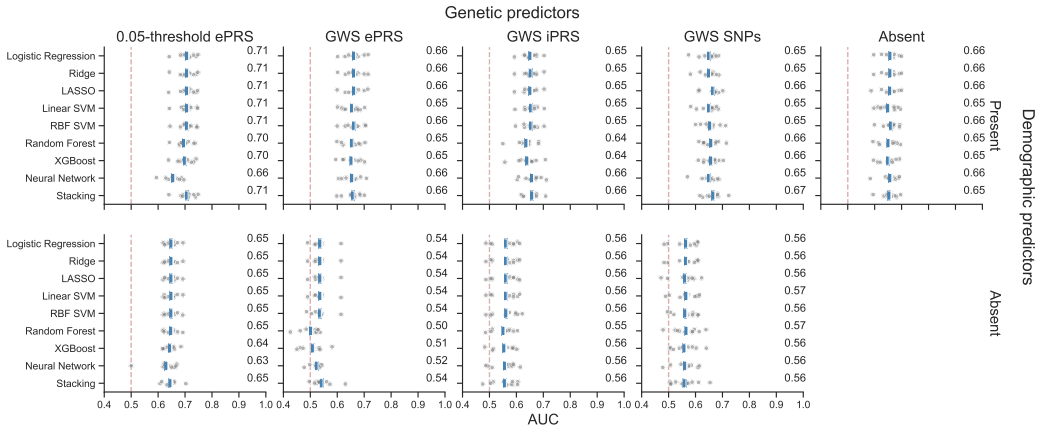


FIGURE 5.3: Discrimination for models across all datasets. 7 machine learning methods were compared to logistic regression. Models were trained with different types of genetic predictors, including or excluding demographic variables. Poor prediction is observed for neural networks for 0.05 ePRS, as models failed to converge within the specified number of epochs for some folds of cross-validation.

Despite moderate correlation between some model predictions (Figure C.11), a stacking-based meta-classifier trained on models built using the 0.05 ePRS did not improve upon other learning approaches.

Classifier	Genetic	Demographic	W	p	Adjusted p	Reject H_0
XGBoost	0.05-threshold ePRS	Present	0	0.005062	0.07463	True
Neural Network	0.05-threshold ePRS	Present	0	0.005062	0.07463	True
Random Forest	GWS ePRS	Absent	0	0.005062	0.07463	True
Random Forest	0.05-threshold ePRS	Present	1	0.00691	0.07463	True
XGBoost	None	Present	1	0.00691	0.07463	True
XGBoost	GWS iPRS	Present	3	0.01252	0.1126	False
XGBoost	GWS ePRS	Present	4	0.0166	0.1281	False
Random Forest	GWS iPRS	Present	5	0.02182	0.1309	False
Neural Network	GWS ePRS	Present	5	0.02182	0.1309	False
XGBoost	0.05-threshold ePRS	Absent	6	0.02842	0.1535	False
Random Forest	GWS iPRS	Absent	7	0.03666	0.18	False
Neural Network	GWS ePRS	Absent	2	0.04252	0.1914	False
Random Forest	0.05-threshold ePRS	Absent	8	0.04685	0.1946	False
Ridge	GWS SNPS	Present	9	0.05934	0.2003	False

Classifier	Genetic	Demographic	W	p	Adjusted p	Reject H_0
LASSO	GWS SNPS	Present	9	0.05934	0.2003	False
XGBoost	GWS ePRS	Absent	9	0.05934	0.2003	False
Random Forest	GWS ePRS	Present	10	0.07446	0.2234	False
Random Forest	None	Present	10	0.07446	0.2234	False
Ridge	GWS iPRS	Present	12	0.1139	0.3237	False
Ridge	None	Present	2	0.138	0.3726	False
RBF SVM	GWS SNPS	Present	14	0.1688	0.4143	False
Ridge	0.05-threshold ePRS	Present	14	0.1688	0.4143	False
Linear SVM	None	Present	15	0.2026	0.4757	False
Linear SVM	GWS ePRS	Present	16	0.2411	0.5425	False
XGBoost	GWS SNPS	Present	17	0.2845	0.5938	False
LASSO	GWS iPRS	Present	13.5	0.2859	0.5938	False
RBF SVM	GWS iPRS	Present	18	0.3329	0.6658	False
RBF SVM	GWS iPRS	Absent	15	0.3743	0.6717	False
Random Forest	GWS SNPS	Present	19	0.3863	0.6717	False
Linear SVM	GWS SNPS	Present	19	0.3863	0.6717	False
Neural Network	0.05-threshold ePRS	Absent	19	0.3863	0.6717	False
Neural Network	GWS iPRS	Absent	9	0.398	0.6717	False
RBF SVM	GWS ePRS	Present	20	0.4446	0.7061	False
Linear SVM	GWS iPRS	Present	20	0.4446	0.7061	False
Neural Network	GWS iPRS	Present	21	0.5076	0.7214	False
RBF SVM	None	Present	21	0.5076	0.7214	False
LASSO	GWS SNPS	Absent	21	0.5076	0.7214	False
XGBoost	GWS SNPS	Absent	21	0.5076	0.7214	False
Linear SVM	GWS SNPS	Absent	22	0.5751	0.7574	False
Linear SVM	0.05-threshold ePRS	Present	22	0.5751	0.7574	False
RBF SVM	0.05-threshold ePRS	Present	22	0.5751	0.7574	False
Ridge	GWS SNPS	Absent	18	0.594	0.7637	False
RBF SVM	GWS SNPS	Absent	23	0.6465	0.7934	False
XGBoost	GWS iPRS	Absent	23	0.6465	0.7934	False
RBF SVM	0.05-threshold ePRS	Absent	15	0.6744	0.8093	False
Neural Network	GWS SNPS	Absent	24	0.7213	0.8467	False
LASSO	GWS ePRS	Present	25	0.7989	0.9178	False
Ridge	GWS ePRS	Present	21	0.859	0.9488	False
Random Forest	GWS SNPS	Absent	26	0.8785	0.9488	False
Neural Network	None	Present	26	0.8785	0.9488	False
RBF SVM	GWS ePRS	Absent	10	0.9165	0.9541	False
LASSO	0.05-threshold ePRS	Present	26.5	0.9188	0.9541	False
LASSO	None	Present	27	0.9594	0.9594	False

Classifier	Genetic	Demographic	W	p	Adjusted p	Reject H_0
Neural Network	GWS SNPS	Present	27	0.9594	0.9594	False
Ridge	0.05-threshold ePRS	Absent			False	False
LASSO	0.05-threshold ePRS	Absent			False	False
Linear SVM	0.05-threshold ePRS	Absent			False	False
Ridge	GWS ePRS	Absent			False	False
LASSO	GWS ePRS	Absent			False	False
Linear SVM	GWS ePRS	Absent			False	False
Ridge	GWS iPRS	Absent			False	False
LASSO	GWS iPRS	Absent			False	False
Linear SVM	GWS iPRS	Absent			False	False

TABLE 5.1: Comparison between each classifier and logistic regression for all datasets using the Wilcoxon signed-rank test. Tests statistics, W, and p-values are missing where AUCs were identical between a model and logistic regression. p-values were adjusted for multiple testing using FDR at 0.1. "Reject H_0" gives a boolean of whether the null hypothesis of no association for the FDR-corrected p-value would be rejected. Stacking could not be included in comparisons as it includes the output of logistic regression in its stacked models. XGB: extreme gradient boosting, NN: neural network, RF: random forest, LASSO: least absolute shrinkage and selection operator, SVM: support vector machine, RBF: radial basis function, GWS: genome-wide significant, iPRS: internal polygenic risk score, ePRS: external polygenic risk score.

Though no model consistently out-performed all others, on average linear models showed slightly better prediction across all models (Figure 5.4a). In general, distributions became more diffuse with more flexible methods, though differences between modelling approaches and logistic regression are often small (<2% AUC), illustrated most clearly by comparing all linear models against all non-linear models (Figure 5.4b).

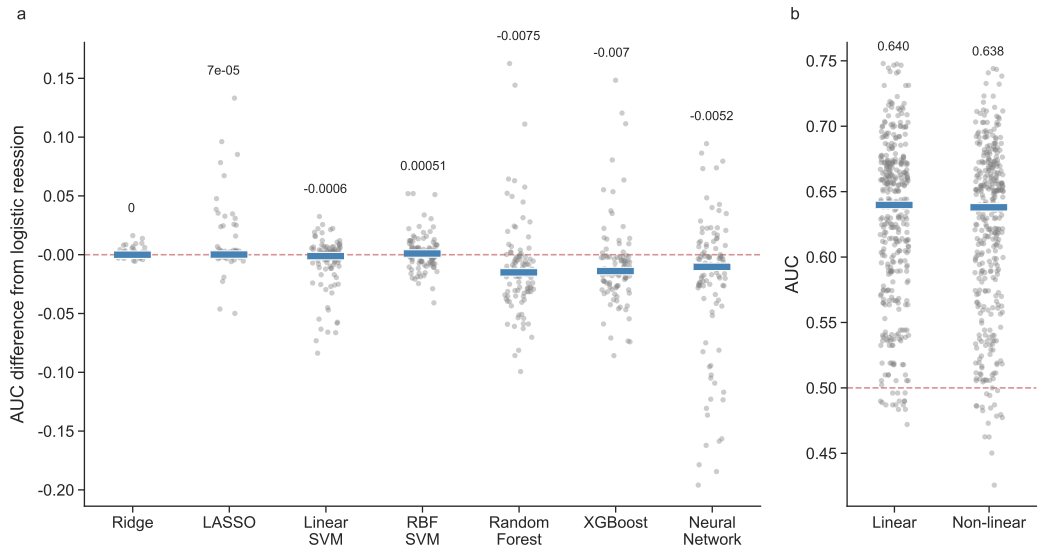


FIGURE 5.4: Comparison of modelling approaches. The outer-fold AUCs for each modelling approach were subtracted from the corresponding AUCs for logistic regression, to show the distributions of differences from logistic regression across all models (a). Models capable of linear (logistic regression, ridge, LASSO, linear SVM) and non-linear (RBF SVM, random forest, XGBoost, neural networks) mappings were binned together to show the overall difference between techniques (b). The median value is given by the blue bar and annotated on each strip.

5.3.2.2 Discrimination by predictor type

Discrimination between cases and controls (Figure 5.3) using genetic, demographic or combined predictors was compared for the datasets using the 0.05 ePRS, as this showed the highest discrimination of all genetic predictors considered. Tests for each classifier comparing combined data with either the 0.05 ePRS or demographic predictors showed strong evidence for differences for all approaches apart from neural networks (Table 5.2). Overall, models incorporating combined predictors improved discrimination by around 5-6% AUC compared to those using genetics or demographic factors alone. Combined predictors showed significantly higher AUC (median 0.71 AUC across methods) over prediction using only demographic predictors (median 0.66 AUC) or only 0.05 ePRS (median 0.65 AUC). By contrast, combination of demographic factors with genome-wide significant SNPs as individual predictors (median 0.66 AUC) or combined into a PRS with external (median 0.66 AUC) or internal (median 0.65 AUC) effect sizes showed equivalent prediction to demographic factors alone (median 0.66 AUC). This may be expected as prediction from each of the genetic predictors based on genome-wide significant SNPs alone is low (0.54-0.57 median AUC).

Classifier	Comparison	W	p	Adjusted p	Reject H_0
Logistic Regression	Combined vs. Genetic	0	0.0051	0.0058	True
Logistic Regression	Combined vs. Non-genetic	0	0.0051	0.0058	True
Ridge	Combined vs. Genetic	0	0.0051	0.0058	True
Ridge	Combined vs. Non-genetic	0	0.0051	0.0058	True

Classifier	Comparison	W	p	Adjusted p	Reject H_0
LASSO	Combined vs. Genetic	0	0.0051	0.0058	True
LASSO	Combined vs. Non-genetic	0	0.0051	0.0058	True
Linear SVM	Combined vs. Genetic	0	0.0051	0.0058	True
Linear SVM	Combined vs. Non-genetic	0	0.0051	0.0058	True
RBF SVM	Combined vs. Genetic	0	0.0051	0.0058	True
RBF SVM	Combined vs. Non-genetic	0	0.0051	0.0058	True
Random Forest	Combined vs. Genetic	0	0.0051	0.0058	True
Random Forest	Combined vs. Non-genetic	0	0.0051	0.0058	True
XGBoost	Combined vs. Genetic	0	0.0051	0.0058	True
XGBoost	Combined vs. Non-genetic	0	0.0051	0.0058	True
Neural Network	Combined vs. Genetic	18	0.33	0.36	False
Neural Network	Combined vs. Non-genetic	21	0.51	0.51	False

TABLE 5.2: Per-model comparison of combined versus genetic and demographic-only models using the Wilcoxon signed-rank test. Tests were carried out for models using demographic and 0.05 ePRS predictors. Demographic or ePRS models were compared to combined models in a pairwise manner for each classifier. Adjusted p-values were produced using FDR-correction at 0.1.

Models trained on genome-wide significant SNPs, without demographic variables, showed improved discrimination for iPRS or SNPs compared to ePRS (Figure 5.3). Pairwise comparisons of iPRS or SNPs to ePRS found weak evidence for a difference for the majority of classifiers (Table 5.3), which was not significant using an FDR of 0.1. As noted previously, to enable comparison across genetic and demographic datasets, all datasets had been reduced to the same sample of cases and controls by removing observations which showed missingness in demographic predictors. Re-analysis of ePRS-only models compared to iPRS-only and SNP-only models in the larger nested sample, which includes individuals previously removed due to missingness, found stronger evidence of a difference between cases and controls (Table C.3), with most classifiers reaching significance at FDR of 0.1.

Classifier	Comparison	W	p	Adjusted p	Reject H_0
Random Forest	ePRS vs. SNPs	3	0.013	0.13	False
Random Forest	ePRS vs. iPRS	4	0.017	0.13	False
XGBoost	ePRS vs. iPRS	7	0.037	0.17	False
XGBoost	ePRS vs. SNPs	8	0.047	0.17	False
Neural Network	ePRS vs. SNPs	9	0.059	0.17	False
Neural Network	ePRS vs. iPRS	10	0.074	0.17	False
Logistic Regression	ePRS vs. SNPs	13	0.14	0.17	False
Ridge	ePRS vs. SNPs	13	0.14	0.17	False
LASSO	ePRS vs. SNPs	13	0.14	0.17	False

Classifier	Comparison	W	p	Adjusted p	Reject H_0
Linear SVM	ePRS vs. SNPs	13	0.14	0.17	False
Logistic Regression	ePRS vs. iPRS	14	0.17	0.17	False
Ridge	ePRS vs. iPRS	14	0.17	0.17	False
LASSO	ePRS vs. iPRS	14	0.17	0.17	False
Linear SVM	ePRS vs. iPRS	14	0.17	0.17	False
RBF SVM	ePRS vs. SNPs	14	0.17	0.17	False
RBF SVM	ePRS vs. iPRS	14	0.17	0.17	False

TABLE 5.3: Per-model comparison of ePRS versus iPRS or SNP models of genome-wide significant SNPs only using the Wilcoxon signed-rank test. Each iPRS or SNP model underwent pairwise comparison with the corresponding ePRS model from the same classifier. Adjusted p-values were produced using FDR-correction at 0.1. iPRS: internal polygenic risk score, ePRS: external polygenic risk score, SNPs: single nucleotide polymorphisms.

5.3.3 Importance scores

Models using the 0.05 ePRS (which uses external estimates of effect sizes from Pardiñas et al., 2018) combined with all demographic predictors showed the highest discrimination on average; these were further evaluated for interpretation and calibration. To assess whether any difference was present in the weight given to different predictors, test set permutation importance was evaluated for the overall best-performing linear (LASSO) and non-linear (RBF SVM) models (Table C.4) for all outer CV folds. Models gave very similar weights to predictors despite different modelling approaches (Figure 5.5), ranking the 0.05 ePRS as the strongest predictor. Accuracy of importance scores was checked by inclusion of binary or normally-distributed random noise predictors in the model, which received the lowest importance ranking (Figure C.19).

A group-wise mask applied in permutation of predictors, however, showed similar importance was given to all demographic predictors taken together when compared to PRS.

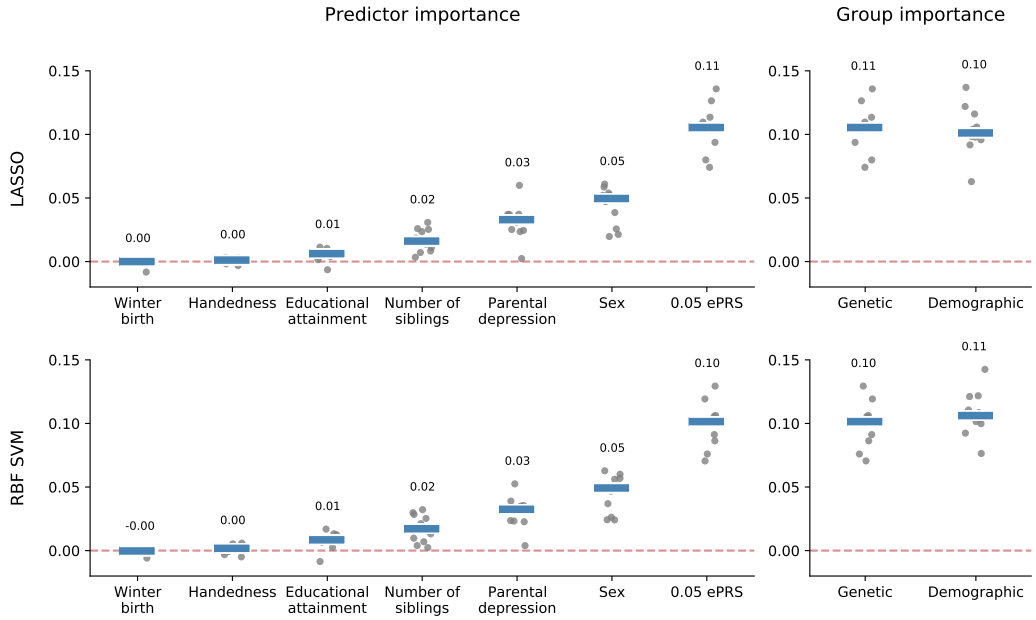


FIGURE 5.5: Permutation importance scores for LASSO and RBF SVM. Very similar values of relative importance are seen for both methods. Models were trained on the combined set of demographic and 0.05 ePRS predictors. Strip plots are annotated with the median value for each predictor across outer folds from nested cross-validation.

5.3.4 Calibration

Calibration was evaluated for all models. To allow for comparison, models were recalibrated in cross-validation. All models showed good calibration using graphical evaluation via loess smoothers (Figure 5.6), though neural networks still display a slight sigmoidal curve.

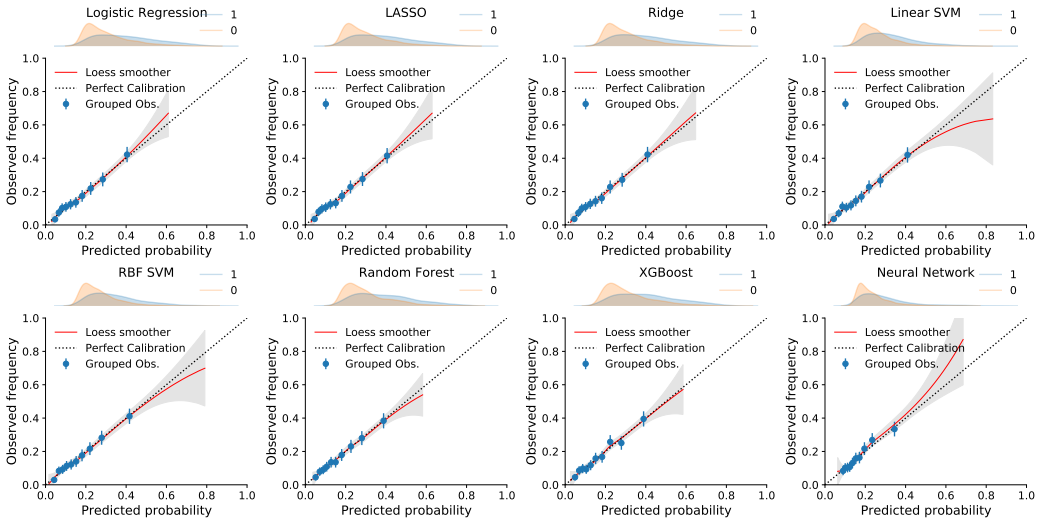


FIGURE 5.6: Calibration for all models in the nested case-control sample. Predicted probabilities have been adjusted through Platt scaling in cross-validation. Validation plots are shown for all models, displaying calibration via loess curves and discrimination through kernel density estimation. All models appear well-calibrated within the nested sample, with the exception of neural networks which display a slight sigmoidal shape. Validation plots are shown for the overall best linear and non-linear models in the combined demographic and 0.05 ePRS dataset.

Calibration in different predictor and confounder subgroups was investigated, as recommended for strong calibration (Steyerberg et al., 2019). LASSO and RBF SVM models showed slightly better calibration and discrimination on the Axiom array used in genotyping when compared to the BiLEVE array; however, perfect calibration is within the confidence intervals for the BiLEVE array and may converge toward that of Axiom with a larger sample size. (Figure 5.7). Calibration by principal components could not be assessed as it is only possible for categorical variables. Data from all folds are combined for validation plots; curves are similar when split by fold (Figure C.12).

Investigation by each categorical predictor shows similar calibration for subgroups. Deviation from the diagonal of perfect calibration occurred for handedness (Figure C.14), where models tended to give a lower predicted probability of schizophrenia occurring in those who are left-handed or ambidextrous than was actually observed. Simulations have shown that such deviations occur under small sample sizes with an AUC of 0.7 (Austin and Steyerberg, 2014), and that increased sample size and discrimination give more readable calibration curves. By contrast, sex (Figure C.15), education (Figure C.16), winter birth (Figure C.17) and parental depression (Figure C.18) all showed similar calibration across subgroups, with the exception of less common events (having severe depression in both parents) where confidence intervals can become particularly large, and interpretation of calibration plots is impaired.

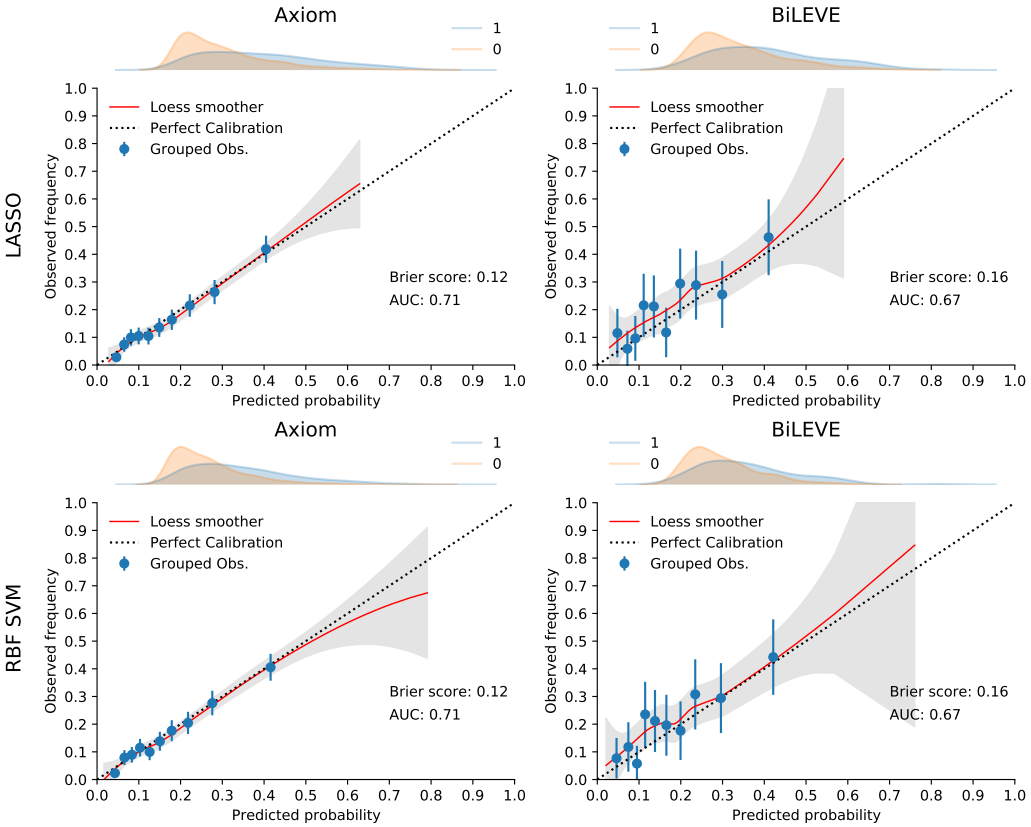


FIGURE 5.7: Validation plots by genotyping array. Calibration in categorical groups can reveal where a subgroup is not adequately handled by a model. Plots demonstrate that probabilities are not systematically over or underestimated for either array. Wider confidence intervals are present for the BiLEVE array, which has fewer observations. Validation plots are shown for the overall best linear and non-linear models in the combined demographic and 0.05 ePRS dataset.

Though overall calibration within the study sample is good, only the reduced, nested case-control dataset has been used. The 'true probability' in Figure 5.6 is therefore only the proportion in the nested case-control sample. Taking predicted probabilities for the whole UK Biobank cohort together, classifiers consistently overestimate the probability of schizophrenia occurring, as the sampling fraction of 0.01 means the prevalence of schizophrenia in the nested case-control sample and the whole UK Biobank cohort differ; adjusting for the sampling fraction ameliorates this issue but gives more variable predicted probabilities (Figure 5.8), which is expected when both sample size and prevalence are low (Austin and Steyerberg, 2014).

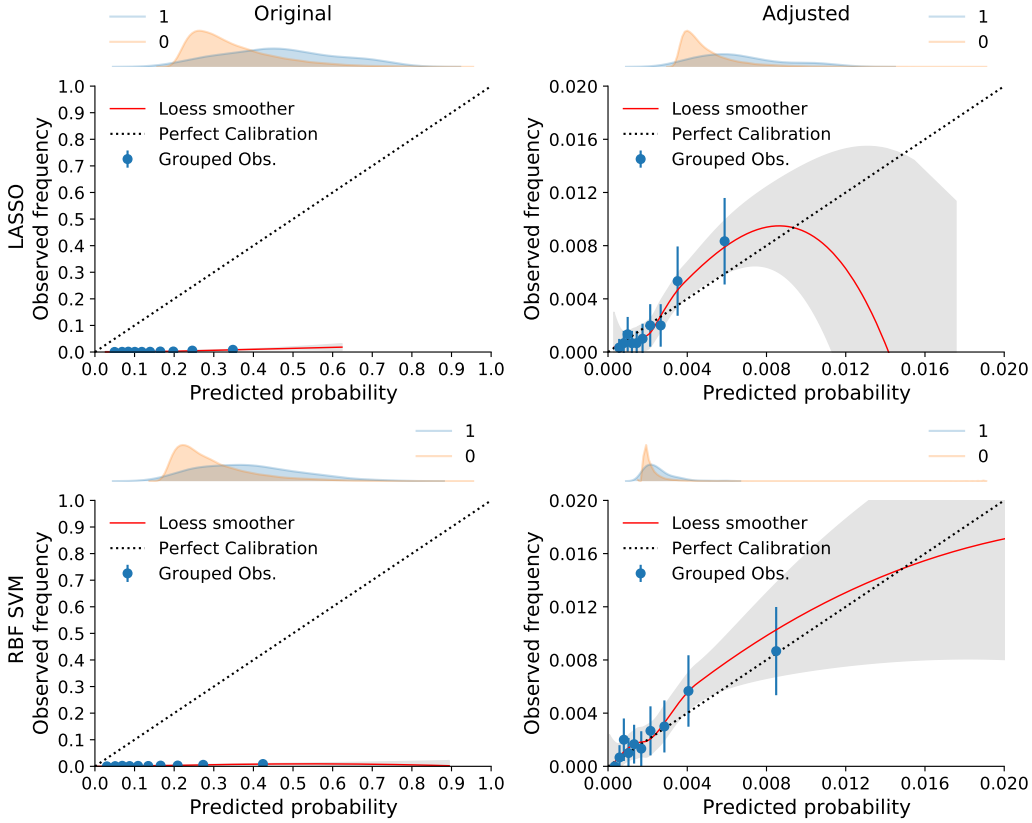


FIGURE 5.8: Calibration of models in the entire UK Biobank cohort. Predictions from cross-validation were combined with predictions in remaining unsubsampled controls in the cohort, generated through a refit on the nested case-control sample. Probabilities are consistently over-estimated due to the differing prevalences in the nested sample and the total cohort. Adjustment for the sampling fraction improves calibration but is difficult to assess due to variability. Validation plots are shown for the overall best linear and non-linear models in the combined demographic and 0.05 ePRS dataset. Due to in-memory limits of the loess algorithm in the scikit-misc package, a subsample of 30,000 participants is shown; regenerating validation plots with different random subsamples shows curves are representative (Figure C.13).

5.3.5 Generalisable association of predictions with cognitive, demographic, psychiatric and neurological outcomes

To further assess the prediction given by different models, predicted probabilities from LASSO and RBF SVMs in the remaining unsubsampled UK Biobank cohort were evaluated. The ability to predict calibrated risk scores from LASSO and RBF SVM models using cognitive and demographic variables was assessed through a cross-validated beta regression. Similar patterns were shown for both classifiers (Figure 5.9). Results are reported as heatmaps of mean test-set R^2 from 5-fold cross-validation; accompanying plots of standard errors are given in the appendix C (Figures C.20 and C.21).

Highest mean variance explained was observed for demographic information such as deprivation, BMI and smoking status, factors which have clear associations with schizophrenia. These were strongest for demographic-only models. Demographic and combined models also

showed generally higher variance explained for schizophrenia-related factors than genetic-only models for both cognitive and demographic factors. Despite diverging age distributions for cases and controls, age was not associated with predictions in controls, indicating no age-related confounding in unaffected participants.

Cognitive tests show mixed results, with low R^2 for fluid intelligence, symbol digit substitution and trail making test B, while trail making test A, pairs matching, digit span and reaction time show little to no variance explained in schizophrenia risk scores. Results indicate aspects of reasoning, complex processing speed and visual attention were most detected in machine learning models. Combining all cognitive tests into a multi-variable beta regression showed the strongest predictive ability to predict schizophrenia risk scores, though sample size was much lower ($n = 7,919$) as relatively few individuals completed all tests. Height generalises with demographic and combined risk scores, but not genetic models, as expected (Figure 5.9).

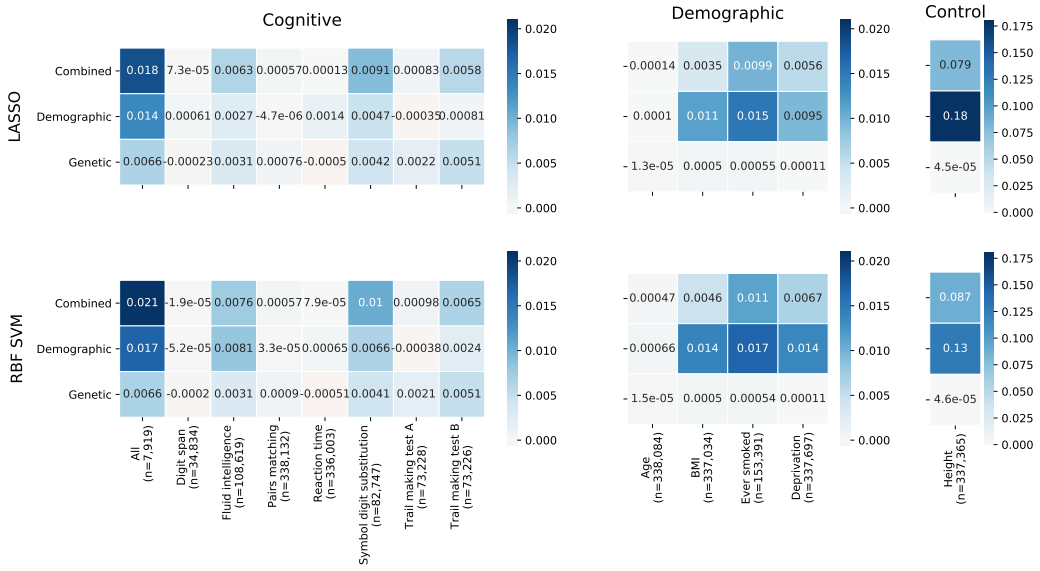


FIGURE 5.9: Prediction of risk scores using cognitive tests and widely-available demographic information in UK Biobank controls. Height is included as a negative control for genetic liability to schizophrenia and a positive control for models including demographic factors, as average height differs by sex, which is included as a predictor in demographic and combined datasets. LASSO and RBF SVM models are shown for genetic (0.05 ePRS), demographic or combined models. Values are the test-fold R^2 between calibrated machine learning risk scores and predictions from a beta regression of cognitive or demographic factors regressed against model predictions, averaged across a 5-fold cross-validation. Values may be negative or positive, with small negative values expected for the test-set when true $R^2 = 0$. Colour mapping differs for height, as it shows much higher average R^2, but is consistent for other subplots. Cognitive and demographic outcomes are detailed in appendix C.1.2.

Prediction of machine learning risk scores using psychiatric or neurological outcomes, visualised on the same scale as cognitive and demographic factors, showed a mean R^2 of around 0 for most outcomes (Figure 5.10). For psychiatric disorders, the highest variance explained

in risk scores was observed for depression, but all disorders showed extremely low or no generalisation with model predictions. Having seen a psychiatrist for nerves or anxiety showed the highest R^2 for all psychiatric factors. Neurological outcomes showed highest mean R^2 for stroke in demographic and combined models. Genetic models showed no generalisation with neurological factors.

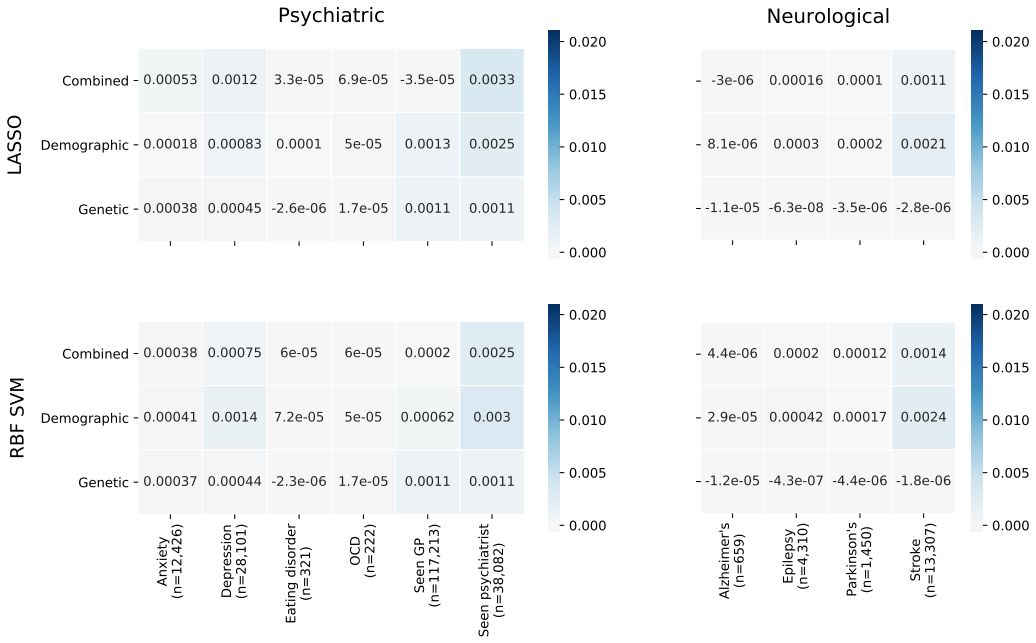


FIGURE 5.10: Prediction of risk scores using neurological and psychiatric outcomes in UK Biobank controls. LASSO and RBF SVM models are shown for genetic (0.05 ePRS), demographic or combined models. Values are the test-fold R^2 between calibrated machine learning risk scores and predictions from a beta regression of cognitive or demographic factors regressed against model predictions, averaged across a 5-fold cross-validation. Values may be negative or positive, with small negative values expected for the test-set when true $R^2 = 0$. Psychiatric and neurological outcomes are detailed in appendix C.1.2.

5.3.6 Deconfounding

Deconfounding was performed within cross-validation for 40 principal components and genotyping array on all genetic predictors. Presence of residual confounding in unsubsampled controls was assessed by cross-validated prediction of machine learning risk scores using 15 principal components, genotyping array, or their combination. Mean test fold R^2 is around 0 for genotyping array, but is consistantly elevated for principal components (Figure 5.11a and b), suggesting some population structure in controls was not accounted for in model development. Schizophrenia also shows weak prediction in the nested sample (0.535 AUC) from principal components, which may be expected for a confounder, given that the sample is only restricted to individuals self-reporting as white British or Irish. Participants were subsequently further restricted to those for whom genetic ancestry appears homogeneous by principal components (using field 22006 supplied by UK Biobank). This decreased the number of nested cases from 738 to 643 (~8% reduction), and nested controls from 3,690

to 3,410 (~13% reduction) and showed substantially less structure in plots of principal components in the full UK Biobank sample (Figure C.22). Prediction of schizophrenia status from principal components maintained an AUC of 0.535 in the more homogeneous nested sample. Variance in machine learning risk scores explained by principal components did not decrease upon restriction to homogeneous genetic ancestry for genetic models, but roughly halved for demographic models (Figure 5.11d and e).

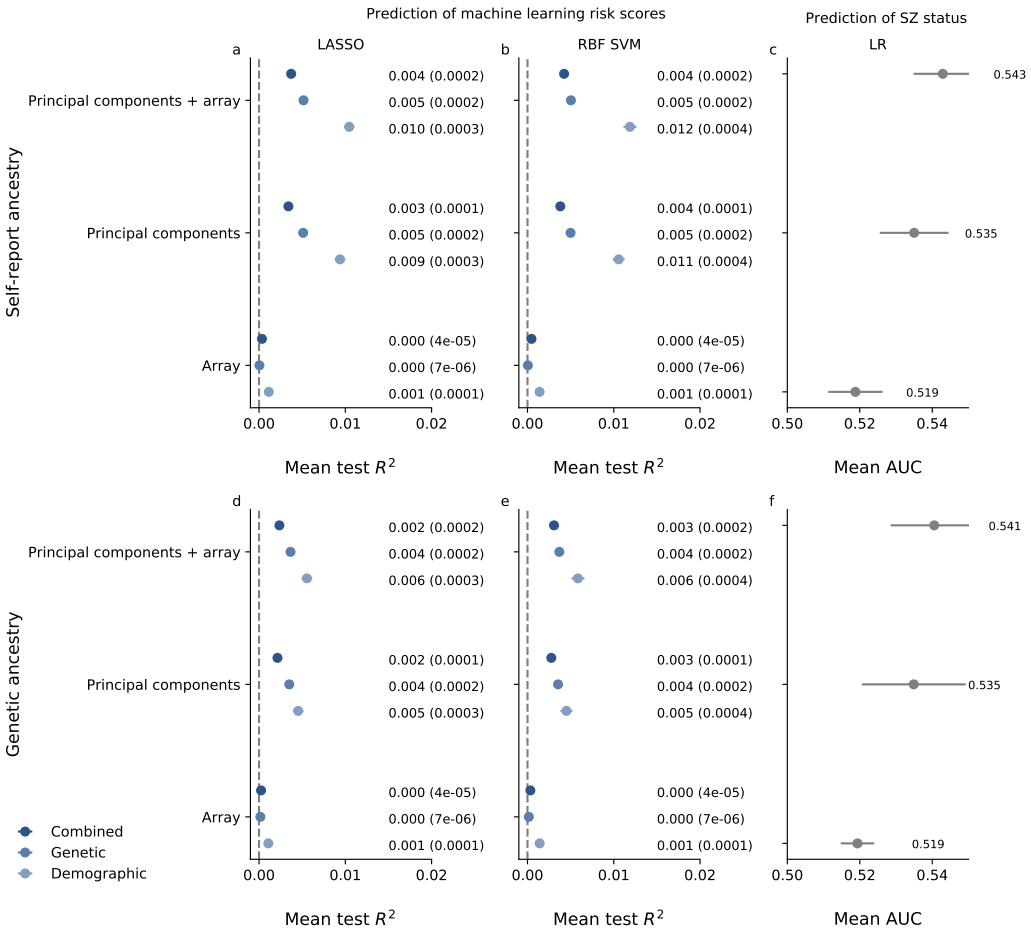


FIGURE 5.11: Assessment of deconfounding in unsubsampled controls. Model predictions were evaluated using genetic (0.05 ePRS), demographic or combined predictors for how well they could be predicted using principal components, genotyping array, or a combination of the two. LASSO (a, d) and RBF SVM (b, e) models were evaluated. The ability of genetic confounders to predict schizophrenia in the nested sample using logistic regression was also assessed (c, f). Standard error bars are shown for all points. Mean values across cross-validation and standard errors are annoted. SZ: schizophrenia., LR: logistic regression.

Comparison of the three most common deconfounding methods to models with no deconfounding procedure showed marked differences in ability to control for population structure (Figure 5.12). Results from RBF SVM and LASSO models show that all methods for deconfounding reduce variance in risk scores explained by principal components. Estimating coefficients for deconfounding once in each training fold in cross-validation, used in all main

analyses and detailed Figure 5.1c, shows higher generalisation with genetic ancestry. Together with results from restricting to genetic ancestry, this suggests the elevated R^2 in Figure 5.11 is consistent with a broader distribution of population structure in the unsubsampled controls, rather than the consequence of outliers for genetic ancestry. Applying a single fit and transformation to the whole dataset prior to cross-validation showed very similar results to estimating coefficients separately in the train and test data and gave the best control for population structure.

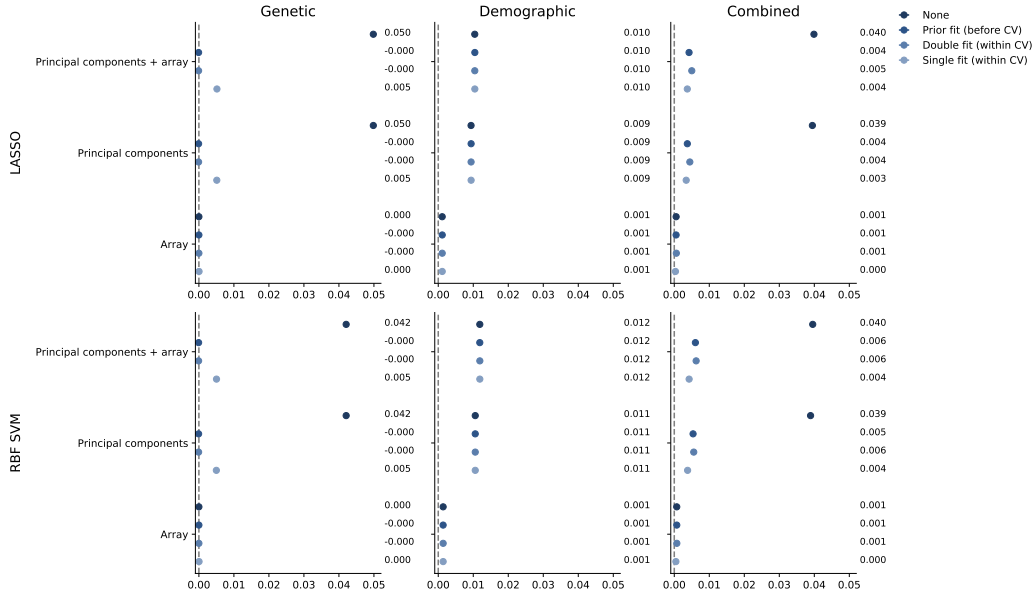


FIGURE 5.12: Comparison of deconfounding methods. LASSO and RBF SVM models were assessed for genetic (0.05 ePRS), demographic or combined predictors in unsubsampled controls for different deconfounding procedures. Methods were evaluated for how well model predictions could be predicted using 15 principal components, genotyping array, or a combination of the two. Methods either applied no deconfounding procedure (none), deconfounding on the whole dataset prior to cross-validation (prior fit), a separate fit for the effects of principal components on predictors in both train and test folds within cross-validation (double fit), or a single fit in the training fold within cross-validation (single fit). Standard error bars are shown for all points.

5.3.7 Sampling fraction in nested case-control designs

A nested case-control design was chosen for computational efficiency. Learning under class imbalance is an open field of research within machine learning, as classifiers may predict the majority class or attain reduced discrimination if imbalance is not handled correctly. The optimal ratio of controls to cases is not always clear; however, a choice of less than 1:5 cases to controls is often used in epidemiology (Biesheuvel et al., 2008). Predictive ability and its variance between folds is shown for 1, 2, 5 and 10x controls for logistic regression using all predictors (Figure 5.13). Discrimination generally remains constant as controls increase; by contrast, variance of AUC scores across folds largely decreases with increasing controls.

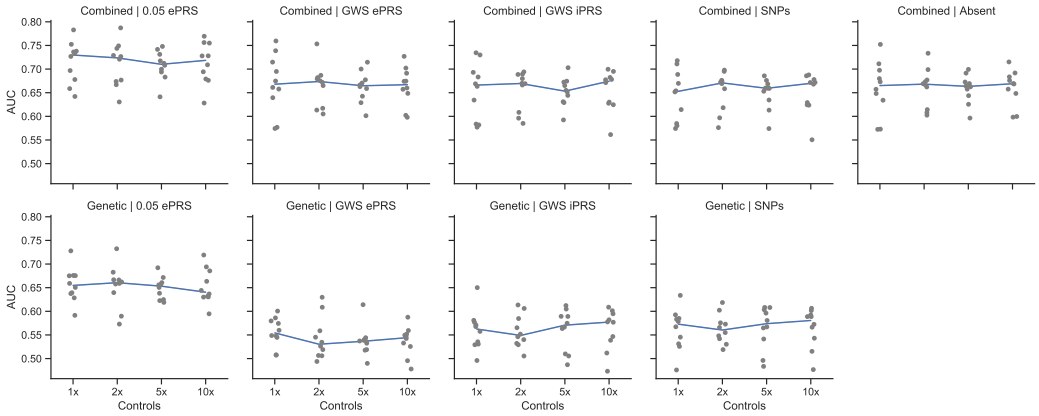


FIGURE 5.13: Discrimination across sampling fractions in nested case-control design. Headings indicate which genetic or demographic predictors were used, with combined-absent denoting demographic predictors only. Controls were sampled at 1, 2, 5 or 10 times the amount as cases, corresponding to sampling fractions of approximately 0.002, 0.004, 0.01 and 0.02 in controls. Discrimination was evaluated for logistic regression across all datasets. Mean AUC across folds remains reasonably stable with increasing sampling fraction for all datasets.

5.4 Discussion

5.4.1 Discrimination

Supervised machine learning models were successfully trained and evaluated in the first comprehensive assessment of discrimination and calibration for genetic prediction of schizophrenia in machine learning. Models trained on genome-wide significant SNPs alone demonstrated improved discrimination for SNPs and iPRS (which uses effect sizes from UK Biobank) compared to ePRS (which uses effect sizes from Pardiñas et al., 2018), suggesting the estimation of effect sizes within UK Biobank is more useful than external sources; the divergent sample ascertainment strategies in the CLOZUK-PGC2 sample and the UK Biobank may underly this difference. Genome-wide significant SNP and iPRS models showed similar discrimination, but more variance across folds was observed with SNPs, likely due to overfitting.

Further assessment of the most predictive linear and flexible models, LASSO and RBF SVM respectively, suggest that demographic and genetic predictors are equally important for discrimination. However, the aetiology of schizophrenia is multifactorial, and inclusion of additional genetic or demographic predictors may alter the strength of these contributions to the model. Importance scores give only the relative contribution of predictors in a particular model and should not be interpreted as an estimate of the overall contribution of genetic or other factors to schizophrenia in general. Despite this, the importance scores and significance tests indicate that building multivariable models of genetic and demographic factors together can provide improved discrimination for schizophrenia.

No improvement was seen from ensembles of machine learning approaches. Stacked models do best with diverse predictors, samples and modelling approaches in each of the base learners of the ensemble. A more extensive investigation of stacking could combine such diversity by

incorporating a wider range of modelling approaches, trained using different subsets, codings and transformations of predictors. Such an approach would lie at the extreme end of the accuracy-interpretability trade-off.

No benefit was observed from more flexible modelling approaches. It is possible that no improvement from flexible modelling approaches was seen due to the heterogeneity of schizophrenia, shown by the subgroups captured in the derived definition of schizophrenia in this study (Figure C.2), such that averaging of any non-additive effects is seen. The number of cases available for the detection and generalisation of interactions between predictors is also limited. Furthermore, the coding used for predictors may favour additive models, as variables for handedness, educational attainment and parental depression could alternatively be encoded as separate predictors. These explanations withstanding, it is possible that the true function that maps the predictors used here to the derived schizophrenia status is indeed linear in UK Biobank. Under such a situation, it is expected that more flexible approaches show similar or worse discrimination than linear approaches, and that tree-based models perform slightly worse than linear models, despite the smoothing effects of ensembling. Results in Figure 5.3 appear to support this scenario.

It has been suggested that machine learning provides no improvement over logistic regression (Christodoulou et al., 2019). Chapter 3 also intimates that what has been observed as differences between modelling approaches in psychiatric genetics may be due to within-study bias in model development - such as more tinkering with difficult-to-train models like neural networks or novel methodological approaches, more possibility of information leakage in more complex pipelines, and more potential for overfitting to information leakage with more flexible models - and that there may be no difference between traditional statistical approaches and ML for prediction of schizophrenia. This chapter offers tentative support for this suggestion.

However, there are important caveats to this. As noted in the discussion for chapter 4, the comparison of many different models implicitly favours those which are the easiest to train. Methods with more hyperparameters can require more hands-on tuning to ensure global minima are reached, or within the search space, during optimisation. Neural networks, for example, have an extremely large potential number of hyperparameters to tune, and new algorithmic developments and options are emerging continuously. This is the most plausible cause of the sometimes diffuse distribution of AUCs for neural networks. Furthermore, effective training of both gradient boosting and neural networks requires some monitoring of convergence, which is made more difficult when many models are run. Comparison of a smaller number of approaches using a split-sample validation may bring AUCs for neural networks into line.

5.4.2 Calibration

Models showed good calibration within the nested sample. Adjusting for the sampling fraction to obtain predictions for all observations caused probabilities to deviate from the diagonal more, particularly for RBF SVMs, indicating that additional steps may be needed to gain well-calibrated predictions from a nested case-control design.

Even well-calibrated models developed in UK Biobank would likely be unsuitable for predicting schizophrenia in the general population, however. The volunteer-based recruitment for UK Biobank reduces the chance of more affected individuals participating. Indeed, the prevalence of schizophrenia in the UK Biobank, at just over 0.2%, is lower than estimates of lifetime prevalence given in chapter 1, despite reasonably broad inclusion criteria. While it has been argued that the cohort may be generalisable despite not being representative (Manolio and Collins, 2010), with some support found for this (Batty et al., 2020), this is less likely to be true for severe psychiatric disorders. In general, participants in the UK Biobank are healthier, from less socioeconomically-deprived areas, and more likely to be female than the general population (Fry et al., 2017). UK Biobank also has a slightly higher percentage of individuals of self-reported white ethnicity (94.6%) than the 2011 UK census (91.3%).

It is likely that more affected cases are not present and UK Biobank represents a more high-functioning sample of individuals with schizophrenia than the UK population. This affects both discrimination and calibration. Assuming the UK population as the target population for a model, ascertainment bias has likely caused a systematic deviation of sample effect sizes from population parameters. As a consequence, discrimination is diminished compared to a model trained in a sample with more severe cases. In the context of logistic regression, effect sizes for associations between predictors and outcome are reduced, causing predicted probabilities to be constrained compared to those from a model with more affected cases. Contraction of the linear predictor is associated with an anti-clockwise tilting of calibration curves from the diagonal (Austin and Steyerberg, 2014), as probabilities move closer to 0.5, leaving controls typically assigned a probability which is too high, and cases too low. Graphical calibration of UK Biobank-trained models in the general population would likely show the loess curve following such a pattern.

5.4.3 Study design

Concerns may also be raised due to the small effective sample size present for schizophrenia in UK Biobank, and the potential impact of missingness on models. Though there is a small number of cases compared to common variant association studies, this is mainly a concern for models built with individual SNPs as predictors, as externally-weighted PRS are mostly dependent on the discovery sample for power (Dudbridge, 2013), and the sample size is reasonable for standard clinical prediction models. Events per candidate predictor parameter (EPP) counts the number of cases per candidate predictor, including all those considered for

association within the sample, and any additional parameters included in the model, such as categorical variables requiring more than 1 coefficient. The EPP for models of combined demographic and ePRS predictors in a logistic regression is 92 (738 cases/8 parameters, as parental depression has 3 levels), sufficiently high for clinical modelling contexts. Combined models of demographic factors and SNPs have a relatively large number of predictors, with an EPP of 3 (738 cases/240 parameters, as 116 SNPs and parental depression have 3 levels), an increase in cases may be beneficial, as models with more coefficients, such as deep neural networks, will have a lower EPP, so these LR-based EPPs represent an upper bound compared to ML models. However, AUCs across folds from these show a similar distribution to those for other predictors, with the exception of random forests and XGBoost in SNP-only models, suggesting that discrimination is reasonably stable under internal validation even with low effective sample size. Similarly, despite differential missingness displayed by several predictors, participants with complete records are demographically highly similar to those with missing data for controls (Figure C.5). A particular concern is that more severely affected cases may be less likely to fully participate in all assessments undertaken by UK Biobank; comparisons indicate that cases with missingness are not outliers for genetic ancestry and do not diverge from those cases without missingness for key demographics. However, a preferred solution to missingness is multiple imputation (Janssen et al., 2010), which was not feasible here due to the computational burden of performing this within cross-validation for all models.

This chapter set out to assess the challenge of learning under severe class imbalance. To address this, techniques from epidemiology and machine learning were combined. A nested case-control design was undertaken as an established technique to greatly reduce the number of controls in a cohort, combined with the model-based technique of re-weighting observations by class imbalance. Results confirm it is an efficient strategy for building machine learning models in a cohort with large class imbalance. Diminishing returns in predictive performance are seen with increasing controls; the often-recommended 1:5 ratio of cases to controls appears reasonable in UK Biobank.

5.4.4 Generalisable model associations

LASSO and RBF SVM models were selected for further investigation as the most discriminative approaches capable of linear or non-linear models, respectively. Analysis in controls indicates that several features in the general population are associated with a higher predicted risk of schizophrenia. Rather than association, often used in explanatory modelling, the ability to predict schizophrenia risk from factors is utilised through a cross-validated beta regression approach. This is potentially more useful in a predictive modelling context as it stresses generalisation, which is more aligned with real world utility than association (Bzdok, Varoquaux, and Steyerberg, 2020).

This analysis demonstrates weak generalisation of factors associated with schizophrenia, such as deprivation, BMI, smoking and cognitive tests, with machine learning predictions.

By contrast, height, for which association with genetic risk for schizophrenia is not expected, shows R^2 close to 0. It also explained the greatest variance in schizophrenia risk for demographic and combined models, which is expected as these use sex as a predictor. Though age shows association with schizophrenia status, it is not also associated with the most important predictors in the model, and so does not meet the criteria for a confounder. This is born out by the low R^2 demonstrated between age and machine learning predictions. Together these give confidence that the technique is effective in tagging associations which generalise with model predictions.

Results also illustrate the benefit of incorporating demographic factors, as known schizophrenia associations generalise with predictions from these, while little or no variance in genetic scores is explained by factors such as smoking and BMI. Given the differing profiles of R^2 for genetic and demographic predictors, and the increase in discrimination seen from combining them, the included genetic and demographic predictors incorporate partly distinct features of schizophrenia risk in this dataset.

5.4.5 Deconfounding

Generalisable associations may also highlight unintended consequences of model development which warrant improved specification of the model. This chapter presents the first detailed dissection of model confounding for a machine learning-based genetic prediction model in schizophrenia. Results show clearly that deconfounding procedures do reduce generalisation of population structure with model predictions. Applying deconfounding before or within cross-validation, where the latter involves learning coefficients separately in the training and testing folds, provide similar levels of adjustment for genetic ancestry. Though prior deconfounding is far easier to implement, it should be avoided as it breaks the statistical validity of cross-validation. As a result of prior deconfounding, train and test folds are no longer independent, a core assumption of cross-validation (Hastie, Tibshirani, and Friedman, 2009). Furthermore, separately estimating the linear effect of population structure on predictors in both training and testing data also breaks predictive modelling assumptions. This assumes that new predictions will only ever be needed on batches of individuals for which principal components can be computed, and so precludes prediction on a single individual. Despite higher variance explained by genetic ancestry, use of a single fitting procedure in the training set for deconfounding, as applied here, is still preferred. While deconfounding was assessed in a large sample of controls, care should be taken if assessing deconfounding from cross-validation results. Amalgamated predictions from across all folds are from distinct but correlated models. Though their combination may be a useful simplification for visual calibration, measures such as the AUC are only viable if calculated separately for each test fold. In the same vein, cross-validated beta regression of confounding variables must be run separately for each test fold, rather than once on the pooled predictions, and so risks overfitting on small datasets with many principal components.

Despite attempts to remove the linear effects of population structure, principal components still show some predictive ability in controls from the full UK Biobank cohort. Though the distribution of principal components appears similar between unsubsampled controls and those in the nested case-control sample, a greater spread is a consequence of a larger sample; variance explained by population structure is expected to reduce as sample size for model development increases. It is also possible that though linear effects were removed from each predictor individually, learning multivariable functions between SNPs and schizophrenia status still approximates some population structure which is not detected here. Evidence for this process occurring in deconfounding of neuroimaging data when using machine learning models has recently been provided, along with suggestions for a post-hoc approach as a solution (Dinga et al., 2020).

5.5 Conclusion

Chapter 3 demonstrated widespread risk-of-bias in machine learning models in psychiatric genetics, in addition to a lack of assessment of calibration and no correction for genetic ancestry or investigation of the potential effects of genetic confounders on the model. Here, supervised machine learning methods were assessed in a large, deeply-phenotyped cohort with severe class imbalance. Unlike previous studies assessing machine learning models of schizophrenia, approaches were systematically compared to logistic regression. Overall, no clear benefit was observed for applying flexible learning approaches. Improved prediction of schizophrenia was demonstrated by jointly modelling genetic and demographic factors. Collectively, assessment of both discrimination and calibration, in addition to use of per-predictor and group-wise importance scores, deconfounding within cross-validation, and cross-validated beta regression outline a framework for evaluating machine learning models in psychiatric genetics.

Chapter 6

Discussion

6.1 Overview

Schizophrenia is a complex psychiatric disorder characterised by a triad of negative, positive and cognitive symptoms. Decades of family, twin and adoption studies have supported the notion that schizophrenia is genetic, but only relatively recently have common and rare variants been reliably associated with the disorder. Despite high heritability, discrimination is generally modest for genetic factors. As outlined in chapter 1, the aetiology of schizophrenia is multifaceted, with contributions from common polymorphisms, rare variants and environmental factors suggested to interact in effecting increased susceptibility to schizophrenia.

Prediction of outcomes in psychiatry, and schizophrenia specifically, is important and increasingly used by researchers. Prediction models typically take the form of PRS which are passed to a logistic regression to evaluate fit or rescale predictions to between 0 and 1. Such models have proved extremely useful and are now a common method for assessing the polygenic common variant component of genetic liability to a disease or disorder. They are attractive for their discrimination and simplicity, as they require only an estimate of the effect size for each SNP and employ a model which is typically additive within and between loci. This makes their application practical where genetic data on individuals cannot be freely shared due to permissions, as summary statistics from association studies are widely available. However, PRS only explain a small portion of risk, with current estimates at around 7-8% of variance in liability (Pardiñas et al., 2018; Ripke et al., 2020). Furthermore, the assumption of additivity, restriction to common variants and averaging of risk alleles into a single score reduce the likelihood of capturing more complex aspects of risk, such as interactions between genetic and environmental factors. An emphasis on prediction in psychiatry also opens it up to methods in machine learning. These approaches contrast with standard techniques such as PRS in their complexity, allowing dominance deviation, interactions between predictors and multivariable modelling when $p > n$. Furthermore, while machine learning approaches place emphasis on learning from data, standard PRS do no learning in the target sample. In an era of increasingly large datasets, both in sample size, the depth of phenotyping and range of omics data, machine learning holds promise of combining heterogeneous data to give personalised predictions and improve patient outcomes.

This is perhaps most exciting for schizophrenia, as it increases the likelihood of capturing risk to a disorder with markedly diverse aetiology. Understandably, this has garnered increased interest in the use of machine learning, but predictive ability has not been systematically assessed across methods or consistently compared to PRS and LR.

A systematic review was undertaken to assess the performance of published models in chapter 3. Several narrative reviews, editorials and comments have been previously written addressing the wealth of flexible tools available from machine learning and their promise for improving prediction in psychiatry. Notably lacking has been a critical assessment of the potential for inflated estimates of predictive performance which can easily arise from small decisions made during model development. Semi-quantitative assessment using the prediction model risk of bias assessment tool (PROBAST) tool identified high risk of bias in all stages of model development and validation for all 77 models. The results highlighted the need for standardised reporting and practices with regards to use of pipelines to avoid data leakage and nested cross-validation when applying machine learning to small and moderate sample sizes. In particular, many studies performed predictor transformations or selection within the whole sample before cross-validation, a process well-known to inflate discrimination and classification measures. Also highlighted was the tendency of authors to use the same rounds of cross-validation to choose hyperparameters and assess prediction performance. This is akin to repeatedly tweaking analyses in explanatory modelling and reporting the best p-value; model selection and validation should be separate, preferably using nested cross-validation where computation is feasible. Issues were also identified in the use of samples which were inappropriate for prediction modelling or a lack of investigation into the impact of population structure on models. These observations were incorporated into subsequent chapters.

Simulations were devised to assess prediction from main or interaction effects in chapter 4. They also provide a useful baseline for what to expect when applying approaches to real data. Gene-gene interactions have received much attention due to their potential for explaining missing heritability in schizophrenia and other psychiatric disorders. Though their detection using machine learning has been studied, prediction has attracted surprisingly little interest. Studies have typically evaluated use of machine learning, often random forests, to detect interactions before applying statistical tests on promising loci; this implicitly takes an explanatory modelling approach by assuming the desired end-point of modelling is the detection of risk factors. Alternatively, a focus on prediction provides a more intuitive fit for machine learning, better assesses how well any interaction effects generalise, and assumes individualised prediction is the focus of research. Analysis found that sparse models, including LASSO, random forest and XGBoost, were favoured under additive effects when a small proportion of variants accounted for the majority of variance in liability, with dense approaches (ridge, support vector machines and neural networks) preferred as the proportion of causal variants increased. The latter is likely to be more appropriate for common variation in schizophrenia; diseases with a small number of loci of large effect may be best

modelled by sparse machine learning models. Findings from interactions highlighted XGBoost and random forest as the most consistently competitive across different interactions models, with good performance also shown by RBF SVMs. Neural networks, while capable of fitting highly non-linear models, showed high variance across simulations and datasets. Discrimination for interactions also showed sensitivity to decrease in linkage disequilibrium with the causal variant, changing minor allele frequency or addition of unassociated variants. This suggests that careful curation of variants for prior knowledge of association, and possibly less-stringent pruning for LD, may increase the likelihood of good prediction from interactions. Overall, results demonstrated best performance for linear sparse approaches under additive simulations, and tree-based ensembles in the presence of interaction effects.

Chapter 5 took up several leads from the systematic review and applied them to prediction in UK Biobank. In particular, it made use of a cohort to apply nested cross-validation on a nested case-control design in which calibration was assessed. Results demonstrated using nested case-controls was an efficient design which did not sacrifice prediction performance, and still allowed for predictions to be generated on the full cohort. Models were recalibrated within cross-validation in addition to being subsequently adjusted for the sampling fraction. The former was necessary for SVMs, which provide predictions as distance from the hyperplane, and tree-based models which have a tendency to retract predicted probabilities toward 0.5. This recalibration, via Platt scaling, also meant base classifiers gave predictions on the same scale to the stacking meta classifier, and that all calibrated predictions in unsubsampled controls were between 0 and 1, and so suitable for later analysis in a beta regression. The second calibration to predicted probabilities accounted for the study design by adjusting for the different proportion of cases in the nested and full samples. Together, these produced predictions from LASSO and RBF SVM models with good calibration in the full cohort. This procedure is extremely useful where generalisation from a sample to the larger population is desired, but is only practical when these are well-matched. Chapter 5 and appendix C demonstrated that the nested sample is similar to the full sample for predictors and key demographic variables. While such recalibration can be reasonable when this occurs, extrapolation from case-control studies to other populations by adjustment for the prevalance may prove hazardous as predictor distributions will differ. Models from association studies are likely to require full recalibration in a target dataset using Platt scaling.

6.1.1 Associations and deconfounding

In addition to calibration, chapter 5 used pipelines for all predictor transformations, evaluated models for generalisable associations, attempted to account for population structure in model development and assessed its impact in model validation. To assess model predictions, it is often desirable to understand how predictions are associated with other variables. This analysis is often used in genetic epidemiology where the prediction, in the form of a PRS, is assessed for association with other risk factors such as cognitive measures. Following calls in the literature for assessment of models under a prediction paradigm, a cross-validated beta

regression was introduced to understand how well the prediction (from a machine learning model) can be predicted by risk factors including demographic, cognitive and psychiatric outcomes. This gives an indication of generalisable associations, which are of greater interest for prediction. It was made possible through the use of the nested case-control design, as this allowed for a large sample of controls to be assessed for prediction from a single model, as opposed to the set of correlated models produced by cross-validation. Results demonstrated that predictions from cognitive measures could explain a small amount of the variance in machine learning predictions; factors such as BMI, deprivation and smoking only showed generalisable associations for models including demographic predictors. More generally, the procedure demonstrated a proof-of-principle for how such an analysis can be used to assess prediction models in a broad dataset such as UK Biobank.

To account for the linear effects of population structure, a deconfounding transformer was devised in scikit-learn that could be applied in all pipelines, with careful attention given to correct assignment of principal components to samples, which can easily become misaligned as sample IDs are not maintained in pipelines and row shuffling is used in cross-validation. While previous studies in psychiatric genetics have applied deconfounding to variants before cross-validation, none had included it within cross-validation or compared approaches. Results identified separate fitting of coefficients in the training and test folds, or prior fitting before cross-validation, as most able to reduce the variance in predictions explained by principal components. Despite this, a single fit of coefficients in each training fold of cross-validation is still preferred, as detailed in chapter 5. However, this implementation of deconfounding is costly. While the procedure is fast for a single predictor, use of 116 SNPs in 10-fold nested cross-validation and 100 iterations of hyperparameter tuning requires around 1.2 million separate regression models, which increases to over 315 million models at 31,603 SNPs.

It is also unclear the degree to which adjustment for principal components is necessary. Selection of individuals which self-report as the same ethnicity should account for social effects which increase risk to a disorder or likelihood of diagnosis. Further restriction to a homogeneous group by genetic ancestry can be used, though this was insufficient to reduce prediction of schizophrenia by principal components in UK Biobank. Deconfounding also faces practical difficulties in that prediction in a single new individual in a real-world setting would require either recomputation of principal components after combining the existing dataset with the new individual, in the case of separate coefficient estimates for train and test folds, or projection of new observations onto components derived in the training data in the case of training fold-only estimates. Ultimately, accounting for allele frequency differences across populations and subpopulations in prediction will require recruitment of more diverse datasets.

6.2 How well can machine learning predict schizophrenia from genetic data?

A motivating question for this thesis is how well classifiers perform in genetic prediction of schizophrenia. As noted in chapter 2, there are several ways of defining what good predictive performance is. It may simply be the ranking of cases and controls by risk (discrimination), or how well predicted and true probabilities concur (calibration). Performance may also be a relative improvement over another model, reclassification of an important subset of individuals, or detection of biologically relevant phenomena.

Discrimination estimates in chapter 3 identified the most likely range as 0.54-0.66 AUC, but noted the caution needed as just two studies contribute to this, with both assessed as at risk of inflated performance estimates. Careful interpretation of simulation results is also required, as assumptions around the number of causal loci, the presence of interaction effects and the homogeneity of populations may be incorrect. However, results from additive simulations indicate ranges of 0.606-0.669 AUC, for 1000 causal variants under additive simulations on the liability scale, and 0.546-0.563 AUC when using 100 causal variants of independent additive effects. Together these estimates span the range of discrimination from the systematic review. Analysis in UK Biobank using 116 genome-wide significant SNPs reached an AUC of around 0.56, matching that expected under additive simulations of a similar number of variants. This indicates that around 0.55 AUC is expected for GWS SNP-based prediction in schizophrenia, and that prediction from machine learning models in UK Biobank can be explained by additive effects.

6.2.1 Does performance differ between machine learning, logistic regression and PRS?

Comparison of machine learning to simpler approaches is important, as more interpretable models are preferable if they provide similar or better discrimination. In comparing machine learning to logistic regression and PRS in published studies, chapter 3 did not identify strong differences between ML and LR, but found slightly better performance for PRS. Though this is attached to the caveat that only 3 studies reported comparisons, it indicates that addition of independent effects or external estimates of effect sizes boost prediction. Simulations in chapter 4 established that both of these phenomena contribute to discrimination in the presence of additive effects. PRS generated using population parameters from simulations represented the best possible prediction, with discrimination far above other models. PRS with internal estimates of effect sizes, calculated in the training data, also showed superior discrimination over ML in most simulations including those incorporating strong linkage disequilibrium. PRS are only improved-upon where a small proportion of variants are associated with the outcome, such that sparse machine learning models thrive, or when trained on only interaction effects, where flexible ML methods excelled, indicating highly polygenic disorders are best modelled by PRS. Logistic regression typically showed equal or worse performance

when compared to ML approaches, which deteriorated as p approached n. In the UK Biobank, where dimensionality was reasonably low, LR and ML showed similar performance; their results may diverge as a larger number of SNPs are introduced to models.

While comparison of ML to PRS (or LR) is informative, it risks creating a false dichotomy. PRS is often treated as a single prediction, but in machine learning it can be conceptualised as feature engineering, the construction of new variables to improve prediction. There is therefore no reason why one or many PRS, weighted by different effects, could not be incorporated into machine learning models. Such an approach was highlighted in chapter 3 as displaying high risk of bias due to sample overlap; increasing the number of separate summary statistics used generally increases the risk of inflating estimates of predictive performance unless overlapping samples can be removed. Similarly, comparison to logistic regression might indicate that ML and LR are independent, but LASSO or ridge with weak penalties perform equivalently to LR. It is perhaps more appropriate to prefer a penalised regression approach in prediction modelling, where a weaker penalty will be chosen in cross-validation if this maximises discrimination.

6.3 Limitations

Several limitations have been highlighted in the thesis. Principle among these are sample size and hyperparameter search. The former is present in both the systematic review and UK Biobank. The effective sample size is low for reviewed papers as many make use of the same datasets compiled for association studies. This overlap further precluded meta-analysis, which would have aided in interpretation of discrimination. However, previous systematic reviews in genetic prediction have also had reasonably small sample sizes, and none had systematically assessed risk of bias. UK Biobank was primarily devised for common diseases and those for which numbers will increase as the cohort ages (Collins, 2012). Small sample size for schizophrenia means variants with lower minor allele frequency or small effects may not contribute to predictions, and that predictions may not generalise to new data upon external validation. This likely also negatively impacted the predictive performance of ML approaches, as flexible methods need more observations to accurately estimate effects from interactions. These limitations are balanced by the depth of phenotyping in UK Biobank, which is rare in large genetic studies, and the large number of controls which underwent common assessments and genotyping. Together these allowed for further investigation of prediction models that would otherwise not have been possible.

As noted in chapters 4 and 5, comparison of many machine learning models negatively impacts those which require more careful hyperparameter tuning. Neural networks are particularly difficult to train across many datasets without manual intervention. Training of a single flexible model may therefore show better performance. Recent improvements to tuning, such as hyperband (Li et al., 2017b), may make for more fair comparisons. Automated tuning procedures such as those based on genetic algorithms (for example, Le,

Fu, and Moore, 2020), collectively called "Auto-ML", may also be useful in this regard. An additional limitation is the exclusive focus on common variants. Incorporation of rare copy number variants, which have relatively large effect sizes for schizophrenia, was not deemed useful in chapter 5. This is because the small number of cases in the UK Biobank, which is biased toward less-affected individuals, would mean CNVs would not be observed in both train and test folds in cross-validation and so would not be informative for prediction. Additionally, exome data for UK Biobank was not available at the point of analysis.

6.4 Future work

An outstanding question is whether increasing the number of SNPs in machine learning models in UK Biobank to the 31,603 used in the 0.05-threshold PRS would improve discrimination. Computation for this was deemed prohibitive for the large number of datasets and models, broad hyperparameter search and deconfounding within cross-validation, and so was not conducted. Given the small number of cases in UK Biobank it is likely that results from such a model would not reach the discrimination achieved by schizophrenia PRS (around 0.66 AUC) but may improve over GWS SNPs if sufficient control for overfitting is in place. In addition, it can be asked whether increasing the sample size would allow for better estimation of any non-additivity or interactions in flexible approaches. While this was not possible here, future work could assess both increased sample size and predictors by making use of recent speed improvement in penalised regression approaches. A fast LASSO procedure for large genomic datasets was recently reported, which could be used within cross-validation to reduce dimensionality before flexible modelling approaches (Qian et al., 2020).

Finally, models employed here are tabular in that they take each variable as a separate predictor without encoding structure between them. Deep learning models which include network information may improve modelling of the relationships between predictors and consequently improve discrimination. Such an approach would require large sample sizes, such as found in case-control association studies. As these were identified as unsuitable for prediction models, a preferred approach may be to select variants present in both a large sample, such as the CLOZUK study used by Pardiñas et al., 2018, and a cohort such as UK Biobank, and train deep learning models in the case-control sample. The network of weights can then be frozen, with the final layer or layers retrained in the UK Biobank by cross-validation, in a process called "transfer learning". This would provide increased sample size and prior information, in addition to a deeply phenotyped cohort in which to investigate the consequences of model predictions.

6.5 Conclusion

Machine learning models hold promise for improved prediction of outcomes in psychiatry, but had previously not been systematically assessed for predictive performance. In this thesis

they were compared for discrimination, calibration and biological interpretation. Results from a systematic review and simulations indicate methods can achieve high discrimination, but are unlikely to out-compete traditional statistical approaches under additive effects. Though this finding was replicated in real data, results from simulations of a small number of causal variants and interactions demonstrate potential for improved prediction in schizophrenia and other psychiatric disorders where interactions between genetic and environmental factors are expected. Increasingly large data in psychiatry are being obtained from multiple sources including common and rare variants in genomics, neuroimaging and health records. Precision medicine will require models which can handle large heterogeneous data and learn complex patterns to predict risk for individuals. Machine learning approaches will play an important role in achieving this and improving outcomes in psychiatry.

Appendix A

Systematic Review of Machine Learning Methods for Genetic Prediction of Psychiatric Disorders

A.1 Methods

A.1.1 Extraction

Events per candidate predictor, or events per variable (EPV), were extracted for all models. Candidate predictors include all predictors considered for inclusion in a model by their association with the outcome. Predictors removed due to association only with other predictors were not counted. As coding of variables is not supplied by most authors, categorical predictors that may be converted to multiple indicator variables by methods are considered only as a single candidate predictor. Similarly, where methods consider additional parameters in the model, such as hidden layers in deep neural networks, only the number of actual predictors is used, not including all possible additional parameters estimated in the model. EPV should therefore be considered an upper bound. Where authors were ambiguous in their reporting of sample size or number of predictors, bounds of the highest and lowest possible EPV are given.

Model discrimination was extracted independently by two individuals. AUC extracted were the same for both authors, except for 3 of 77 models from a single study; consensus was reached after reviewing the text. Studies often included many models; logistic regression models were only extracted where they received the same predictors as ML methods, in order to keep models comparable.

A.1.2 PROBAST method

Risk of bias (ROB) was assessed using the prediction model risk of bias assessment tool (PROBAST). No studies dictated if a model was intended for prognostic or diagnostic use. For the purpose of assessing ROB, models are assumed to be diagnostic; changing intended model use to prognostic does not alter the final ROB assessments for models.

Where databases or publications were referenced for a study, these were assessed for information relevant to ROB. As large genetic datasets may change composition over iterations as smaller studies are added, additional publications that may describe an iteration of a publicly available dataset, but which were not referenced in the included study, were not examined.

Questions in PROBAST are formatted such that answering "Yes" indicates low risk of bias, and answered "no" indicates high risk of bias. Normally, if any questions within a domain are rated "no" or "probably no" (N/PN), then the rating is considered to be "high" ROB for that domain. In the absence of any N/PN responses, if any questions are reported as "no information" (NI), then the domain is taken to have "unclear" ROB. If instead all questions were answered as "yes" or "probably yes" (Y/PY), then the domain is rated as "low" ROB. Select situations where questions are rated NI or N/PN were allowed to be rated "low" ROB overall. For predictors, if question 2.2 ("Were predictor assessments made without knowledge of outcome data?") was rated as NI or Y/PY, overall rating for ROB of predictors was allowed to be "low". Knowledge of the outcome can enable careful design of cases and controls across arrays and batches, and exclusion by a more stringent threshold of Hardy-Weinberg equilibrium in controls. These may allow for reduced ROB for predictors, rather than increased. For outcome, question 3.5 ("Was the outcome determined without knowledge of predictor information?"), if NI or Y/PY, was allowed to be rated "low" ROB for outcome overall if it was considered that genotypes or other predictors would have been extremely unlikely to influence the outcome of standard assessments, or that outcomes were likely to have been assessed prior to genotyping. For question 4.1 ("Were there a reasonable number of participants with the outcome?"), events per candidate predictors were assessed against recommendations using machine learning methods with default hyperparameters, and therefore represent the worst-case scenario. If EPV was determined to be near to the cut-off, and all other modelling procedures indicated low ROB, including appropriate regularisation and handling of predictors, analysis was allowed to be rated "low" overall. In practice, this situation did not occur.

PROBAST requires a ROB assessment of each evaluation of each distinct model (Moons et al., 2019). Development and validation are therefore both assessed for each model and contribute separately to overall counts. Restricting counts to development-only does not appreciably change results. ROB was assessed for all studies by the author, with the exception of a single publication where co-authorship necessitated an additional assessor (Vivian-Griffiths et al., 2019). Here two authors independently assessed ROB, the second being uninvolved in the original study. Differences were overcome through consensus. A third colleague not included in the original study was designated as arbiter should disagreements be unable to be resolved. This situation did not occur.

A.1.3 PROBAST questions

A list of the complete extraction form use for assessing PROBAST is given in the results appendix.

1.1 Were appropriate data sources used, e.g. cohort, RCT, or nested case-control study data? Studies may be made of multiple smaller studies, some of which are cohorts or where cases are from cohorts but controls are from elsewhere. If cases and controls are sampled from different sources to give a roughly balanced (equal events and non-events) combined sample, denote the combined sample as case-control. If absolute risk cannot be estimated from the combined sample, rate as N/PN.

1.2 Were all inclusions and exclusions of participants appropriate? If the target population for the prediction model is undefined, rate as NI, as this cannot be assessed.

2.1 Were predictors defined and assessed in a similar way for all participants? If genotypes measured on different arrays and there has been no effort to demonstrate similarity across arrays or lack of batch effects, rate N/PN. If genotypes from different arrays have been imputed to the same panel of reference genomes to infer untyped or missing variants, rate Y/PY.

3.1 Was the outcome determined appropriately? Consensus best-estimate diagnosis using medical records and structured interview is considered appropriate. Use of only a structured interview is also considered appropriate, but use of only interviews with family members and records is rated N/PN. Routine care registry data are appropriate only if studies confirming comparability with standard diagnostic methods are available. If method is appropriate only for cases, rate 3.1 as Y and 3.4 as N.

3.2 Was a prespecified or standard outcome definition used? Diagnostic and Statistical Manual of Mental Disorders (DSM) or International Classification of Diseases (ICD)-based outcomes are accepted.

3.4 Was the outcome defined and determined in a similar way for all participants? If the same assessments tool was used for all participants, rate Y/PY. If cases were assessed differently to controls, rate N/PN.

3.6 Was the time interval between predictor assessment and outcome determination appropriate? If predictors are genetics-only, rate Y/PY. If predictors include gene-expression data sampled after diagnosis or onset, rate N/PN.

4.1 Were there a reasonable number of participants with the outcome? No recommendations are available for assessing events per variable (EPV) in machine learning models. To our knowledge, only one paper has attempted to assess EPV needed for machine learning models across multiple datasets (Ploeg, Austin, and Steyerberg, 2014), which we use here as a guide in lieu of a more rigorous alternative. For the purpose of assessing ROB in this review, support vector machines are required to have greater than 200 EPV. Neural networks require

at least 200 EPV, but a cut-off of at least 500 EPV should be imposed as architecture can vary greatly. Random forests are also required to have greater than 500 EPV. For other machine learning methods not specified above, 200 EPV is taken as the minimum requirement. Everything below these cut-offs is rated as N/PN. It should be noted that the models these estimates are based on were run using default (hyper)parameters [4] on non-genetic data. Final assessment of ROB for "analysis" should therefore take into account regularisation and model architecture, as models with an EPV of less than 200 may still be rated as "low" ROB for the domain. However, given that all models had multiple aspects of analysis which introduced ROB, changing these thresholds would not affect the final rating for the 'analysis' domain in any models.

4.4 Were participants with missing data handled appropriately? For imputation using a genetics-specific application or server, such as IMPUTE2, rate Y/PY. For imputation in the sample using other methods, rate N/PN. For complete-case analysis, rate N/PN.

4.5 Was selection of predictors based on univariable analysis avoided? If any plink-based univariable tests for association in the current dataset were used, rate N/PN. If information from an external published genome-wide association study (GWAS) was used to select predictors, rate Y/PY.

4.8 Were model overfitting, underfitting, and optimism in model performance accounted for? If nested cross-validation was used, rate Y/PY, assuming other standard procedures were followed. If any method of repeated cross-validation on the whole dataset where both tuning and evaluation of models were done in the same k-fold cross-validation loop was used, or where test data were observed during tuning of hyperparameters, rate N/PN.

4.9 Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis? If no model coefficients or assigned weights clearly reported, rate NI, as this cannot be assessed.

A.2 Results

Where percentages are reported in any table, they are taken from the total number of models, 77, and rounded to the nearest integer unless stated otherwise. Some aspects of methodology differed between models within studies. Where this occurs, studies are counted under each category that has been met unless stated otherwise, and total counts may not sum to 13.

A.2.1 Literature search

Order	Search Command
1	(schizophreni* or schizoaffective or schizotyp* or anxiety or depressi* or autis* or adhd or anorexi* or bullimi* or psychos?s or psychotic or manic or mania or hypomani* or tourette* or obsessive compulsive disorder or ocd).ti,ab. or (exp SCHIZOPHRENIA/ or Bipolar Disorder/ or exp ANXIETY DISORDERS/ or exp Autism Spectrum Disorder/ or exp Depressive Disorder/ or Attention Deficit Disorder with Hyperactivity/ or Anorexia Nervosa/ or Bulimia Nervosa/ or exp Obsessive-Compulsive Disorder/ or Tourette Syndrome/)
2	(machine learning or statistical learning or pattern analysis or pattern recognition or ensemble or Bayesian network* or relevance vector machine* or support vector machine* or decision tree* or classification tree* or regression tree* or elastic net or bagging or gradient boosting or neural network or perceptron or nearest neighbo?r or gaussian process* or ridge or lasso or regulari#ed regression or penali#ed regression or naïve Bayes or (deep adj3 learning) or (boosted adj2 trees) or (deep adj2 network) or (random adj2 forest) or (supervised adj2 learning)).ti,ab. or exp Machine Learning/
3	(rare variant* or rare variation or copy number variant* or copy number variation* or dna variant* or polygenic or genetic* or polymorphism* or genotype* or genome* or genomic* or exome*).ti,ab. or exp Polymorphism, Genetic/
4	1 and 2 and 3
5	limit 4 to english language
6	limit 5 to journal article
7	remove duplicates from 6

TABLE A.1: example literature search from Medline (Ovid).

A.2.2 Extraction

Domain	Item
Background	Reference
	Disorder
	Study design
	Publication number
	Model type (diagnostic/prognostic)
Participants	Recruitment method
	Study setting
	Retrospective or Prospective?
	Number of Centres
	Inclusion/Exclusion criteria
	Sample description
	Study Dates
	Dataset names or identifiers
Sample size	Total number of observations before QC
	Total number of observations after QC
	Case:control ratio in final dataset
	Number of cases in training set/fold
	Events per variable in the training set/fold
Outcome	Definition of outcome
	Measurement
	Same for all patients?
	Type of outcome (single/combined)
	Were assessors blinded to knowledge of predictors?
	Predictors in outcome?
Predictors	Genotyping/sequencing method
	Imputation method and reference
	Types of genetic data
	Method of choice of variants to genotype/sequence
	Genetic Predictor QC
	Number of candidate predictors
	Number predictors in final model
	Coding of genetic data
	Risk allele definition for coding at a single locus
	Knowledge/annotation information included?
	Knowledge/annotation inclusion method
	Was measurement of predictors blinded to outcome/other predictors?
	Any other handling of predictors
	Was leakage handled appropriately?

Domain	Item
Participant QC	Genetic sample QC
	Method for accounting for genetic ancestry
	Method of accounting for plate/batch/site effects
	Method for accounting for relatedness
Missing Data	Number participants with any missing value[1,2]
	Number of participants with missing data for each predictor[1]
	Handling of missing data
	Modelling method/representation
Model Development	Model implementation (programming language)
	Model modifications
	Predictor selection types used
	Method for selection of predictors prior to modelling (filter)
	Method for selection of predictors during modelling (wrapper)
	Method for selection of predictors as part of model (embedded)
	Hyperparameter search method
	Tuned Hyperparameters
	Class imbalance method[1]
Model Performance	Discrimination measures reported
	Calibration measures reported
	Classification measures reported
	Other measures reported.
	A-priori decision threshold cut-off used for classification?
Model Evaluation	Method for testing model performance internally
	Method for testing model performance externally
	Model adjusted or updated after poor validation?[3]
Results	Model AUC
	Model Accuracy, sensitivity and specificity
	Model calibration[1]
	Comparison of distribution of predictors[1]
	Data/code available (link)
Extra	Resources
	Notes

TABLE A.2: extraction form, modified from the checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies (CHARMS) checklist (Moons et al., 2014). Items which overlap heavily with prediction model risk of bias assessment tool (PROBAST) signalling questions, such as participant information, are

TABLE A.2: reported in risk of bias summaries. AUC: area under the receiver operating characteristic curve, QC: quality control. [1]Not reported in any publications. [2]Number of participants excluded above a threshold of missingness was reported in many studies. [3]No for all publications.

Step	Domain	Item/Questions
Classify model type		Type of prediction model
		Publication Reference
		Models of Interest
		Outcome of Interest
Assess ROB & applicability	Participants	1.1 Were appropriate data sources used, e.g. cohort, RCT, or nested case-control study data?
		1.2 Were all inclusions and exclusions of participants appropriate?
		Risk of bias introduced by selection of participants
		Rational of bias rating
		Describe included participants, setting and dates
		Concern that the included participants and setting do not match the review question
		Rationale of applicability rating
	Predictors	List and describe predictors included in the final model, e.g. definition and timing of assessment
		2.1 Were predictors defined and assessed in a similar way for all participants?
		2.2 Were predictor assessments made without knowledge of outcome data?
		2.3 Are all predictors available at the time the model is intended to be used?
		Risk of bias introduced by predictors or their assessment
		Rationale of bias rating
		Concern that the definition, assessment or timing of predictors in the model do not match the review question
		Rationale of applicability rating
	Outcome	Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination

Step	Domain	Item/Questions
		3.1 Was the outcome determined appropriately? 3.2 Was a prespecified or standard outcome definition used? 3.3 Were predictors excluded from the outcome definition? 3.4 Was the outcome defined and determined in a similar way for all participants? 3.5 Was the outcome determined without knowledge of predictor information 3.6 Was the time interval between predictor assessment and outcome determination appropriate? **Risk of bias introduced by the outcome or its determination** *Rationale of bias rating* At what time point was the outcome determined? If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome **Concern that the outcome, its definition, timing or determination do not match the review question** *Rationale of applicability rating*
	Analysis	Describe numbers of participants, number of candidate predictors (for DEV only), outcome events and events per candidate predictor (for DEV only) Describe how the model was developed (predictor selection, optimism, risk groups, model performance) Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants) Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit Describe any participants who were excluded from the analysis Describe missing data on predictors and outcomes as well as methods used for missing data 4.1 Were there a reasonable number of participants with the outcome? 4.2 Were continuous and categorical predictors handled appropriately?

Step	Domain	Item/Questions
		4.3 Were all enrolled participants included in the analysis? 4.4 Were participants with missing data handled appropriately? 4.5 Was selection of predictors based on univariable analysis avoided? 4.6 Were complexities in the data (e.g. censoring, competing risks, samples of control participants) accounted for appropriately? 4.7 Were relevant model performance measures evaluated appropriately? 4.8 Were model overfitting, underfitting, and optimism in model performance accounted for? 4.9 Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis? **Risk of bias introduced by the analysis** *Rationale of bias rating*
Overall judgement		**Overall judgement of risk of bias** *Summary of sources of potential bias* **Overall judgement of applicability** *Summary of applicability concerns*

TABLE A.3: signalling questions in PROBAST.

A.2.3 Samples

Titles and descriptions of studies making up a dataset are recorded as given in the extracted publication. Where references are supplied, these were given in the text, or clear from an online repository, such as the database of Genotypes and Phenotypes (dbGaP) (Mailman et al., 2007). Where datasets appear to overlap, this has been noted.

Study	Disorder	Dataset
Yang et al. (2010)	Schizophrenia	No name/reference given
Ghafouri-Fard et al. (2010)	Autism	No name/reference given
Aguiar-Pulido et al. (2010;2013)	Schizophrenia	External sample[1]
Wang et al. (2018)	Schizophrenia Bipolar disorder Autism	PsychENCODE[2]

Study	Disorder	Dataset
Pirooznia et al. (2012)	Bipolar disorder	BGSC[3†‡] (DEV), WTCCC[4*] (VAL)
Laksshman et al. (2017)	Bipolar disorder	Not clearly reported[5]
Acikel et al. (2016)	Bipolar disorder	Whole-Genome Association Study of Bipolar Disorder[6†‡]
Li et al. (2014)	Bipolar disorder	Whole-Genome Association Study of Bipolar Disorder[6†‡]
	Schizophrenia	Genome-Wide Association Study of Schizophrenia[7†]
Guo et al. (2016)	Anorexia	GCAN[8], WTCCC[4*], CHOP[9], PFCG[10]
Trakadis et al. (2019)	Schizophrenia	Sweden-Schizophrenia Population-Based Case-Control Exome Sequencing[11**]
Engchuan et al. (2015)	Autism	AGP[12]
Chen et al. (2018)	Schizophrenia	MGS[13†], SSCCS[14**] (DEV), CATIE[15†] (VAL)
Vivian-Griffiths et al. (2019)	Schizophrenia	CLOZUK[16*]

TABLE A.4: sample overlap between studies. [1]Galician sample described elsewhere (Domínguez et al., 2007). [2]PsychENCODE, made up of 8/9 studies, where only 6 are listed in the supplementary as having genotype data - study 1 (BrainGVEX, consisting of the Banner Sun Mental Research Institute, BSHRI (Beach et al., 2008), and Stanley Medical Research Institute, SMRI); study 2 (BrainSpan), no genotype data; study 3 (CommonMind (Fromer et al., 2016)); study 4 (Yale-ASD); no genotype data; study 5 (UCLA-ASD (Parikshak et al., 2016)); study 6 (BipSeq); study 7 (CMC-HBCC); study 8 (LIBD-szControl and BipSeq); study 9 (not reported). Information and data also available through an online repository (*PsychENCODE Integrative Analysis*). [3]Bipolar Genome Studies Consortium (BGSC) (Mahon et al., 2009), made up of the Genetic Association Information Network European American (GAIN) (Manolio et al., 2007), and the Translational Genomics Research Institute (TGRI) samples. Controls obtained through Knowledge Networks (KN) (Sanders et al., 2008), and recruitment described elsewhere (Dick et al., 2003; Kassem et al., 2006). [4]Wellcome Trust Case Control Consortium (WTCCC). Bipolar Disorder cases are described in methods, with further information provided elsewhere (Green et al., 2005; Green et al., 2006). Controls include the 1958 British Birth Cohort (58BC) (Power and Elliott, 2006) and the UK Blood Service (UKBS) (Consortium et al., 2007). [5]part of the Critical Assessment of Genome Interpretation (CAGI)-4 challenge. Laksshman et al., 2017 reference Daneshjou et al., 2017, from which a third reference (Monson et al., 2017) gives information on an exome dataset with only bipolar cases recruited for a suicide study, but not controls. [6]Whole-Genome Association

TABLE A.4: Study of Bipolar Disorder, dbGaP study accession "phs000017.v3.p1". References on dbGaP provide further details on sample recruitment (Dick et al., 2003; McInnis et al., 2003). Acikel et al., 2016. acquired Bipolar Disorder Only (BDO) participants; Li et al. report using the Bipolar and Related Disorders (BARD) subset (Li et al., 2014). Controls, obtained through KN, are described under "Clinical Procedures" of the relevant dbGaP entry, and by other studies (Sanders et al., 2008). [7]Genome-Wide Association Study of Schizophrenia, dbGaP study accession "phs000021.v3.p2". Cases described on dbGaP, controls obtained through KN. [8]the Genetic Consortium for Anorexia Nervosa (GCAN). [9]Price Foundation Collaborative Group and the Children's Hospital of Philadelphia (CHOP). Methodological details for Guo et al. are also referenced to a previous study (Boraska et al., 2014). [1]0the Price Foundation Collaborative Group (PFCG). [1]1Sweden-Schizophrenia Population-Based Case-Control Exome Sequencing, dbGaP study accession "phs000473.v1.p1". Described in more detail elsewhere(Purcell et al., 2014). [1]2Autism Genome Project (AGP); three references supplied for methodology and participants (Pinto et al., 2010; Pinto et al., 2011; Pinto et al., 2014). [1]3Molecular Genetics of Schizophrenia (MGS) (Shi et al., 2009), with controls from KN. [1]4Swedish Schizophrenia Case Control Study (SSCCS) (Bergen et al., 2012). [1]5Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) (Stroup et al., 2003; Sullivan et al., 2008), with controls from KN. Imputation for Chen et al. is also given elsewhere (Ware et al., 2016). [1]6CLOZUK (Hamshere et al., 2013); controls from 58BC and UKBS. *Includes controls from the 1958 British Birth Cohort and the UK Blood Service. †Includes controls from Knowledge Networks. ‡Publications do not all give the same dataset name or description, but do include a common reference for recruitment or inclusion criteria. **Studies refer to a Swedish population-based sample with the same outcome definition, but no clear statement or reference describing sample overlap.

A.2.4 Missingness

Methods	Studies	Models
Reported	7	43 (56%)
Exclusion (complete-case analysis)	1	1
Code missingness as category in predictor	1	12
Imputation after excluding high missingness	5	30
Imputation using genetics server/application	3	16
Imputation in-sample from binomial distribution[1]	2	14
Unclear/unreported	6	34 (44%)
Only exclusion for high missingness reported[2]	4	28
Not reported	2	6

TABLE A.5: missingness. Handling of missing data differed between the development and validation set for Pirooznia et al. (2012), where imputation is only reported for external validation (Pirooznia et al., 2012); these models are counted under the method reported in model development, "only exclusion for high missingness". [1]A study (Vivian-Griffiths et al., 2019) reported using unspecified imputation prior to quality control filters, before a second in-sample imputation and is recorded once as in-sample. [2]Includes high missingness filters for samples, predictors or both, with method for handling remaining missingness not reported.

A.2.5 Software

Language/Implementation/Method	Studies	Models
R	4	11 (14%)
glmnet (LASSO)	1	1
randomForest (RF)	2	2
party (CIF)	1	1
e1071 (SVM, NB)	2	2
gbm (GBM)	1	1
XGBoost (Histogram-based GBM)	1	1
*k*NN (*k*-NN)	1	1
MDR (MDR)	1	2
Python	4	16 (21%)
scikit-learn	3	12
SVM	1	8
Data handling	1	1
Unspecified	1	3
Keras (NN)[1]	2	4
Tensorflow (NN)	1	4
Java (WEKA)[2]	2	28 (36%)
Matlab	2	11 (14%)
Matlab (NN)	2	10
libSVM (SVM)	1	1
Not reported	3	11 (14%)

TABLE A.6: software and packages used in machine learning. [1]Backend to Keras not specified. [2]Methods used in WEKA: neural networks (linear, perceptron and radial basis function), evolutionary computation, multifactor dimensionality reduction, Bayesian networks, naïve Bayes, support vector machine, decision tables, decision tree-naïve Bayes, best-first tree, AdaBoost. LASSO: least absolute shrinkage and selection operator, RF: random forest, CIF: conditional inference forest, GBM: gradient boosting machine, XGBoost, eXtreme Gradient Boosting, *k*-NN: *k*-nearest neighbours, MDR: multifactor dimensionality reduction, SVM: support vector machine, NN: neural network, NB: naïve Bayes.

A.2.6 Bias

Methods	Studies	Models
Population substructure identified in current study but not accounted for	2	14 (18%)
Visualised by PCs for subsample after restricting to European	1	9
Table of ancestry for European American and African American[1]	1	5
Unclear[2]	9	50 (65%)
Population structure identified in dataset reference(s)	7	42

Methods	Studies	Models
Exclusion of non-European ancestry through PCs/MDS	5	35
Visualised but observations not excluded	3	11
Reported as European/Caucasian-only, no details given	2	8
Not reported in publication or reference	2	13 (17%)

TABLE A.7: methodology for accounting for population structure. Where development or validation sets are made-up of multiple datasets with separate ancestry filters, these are counted separately. [1]Method of establishing ancestry not specified. [2]Ancestry not clearly specified in current study. PCs: principal components, MDS: multi-dimensional scaling.

A.2.7 Model Performance and Validation

Reported Measures	Studies	Models
Discrimination	8	45 (58%)
AUC	8	45
ROC Plot	4	7
Classification	9	41 (53%)
Accuracy	8	39
Sensitivity/Recall/Hit-rate/TPR	6	16
Specificity/TNR	4	10
F_1-score (F-measure)	3	12
Precision/PPV	3	12
Confusion matrix	3	3
Other	5	29 (38%)
Variance explained on liability scale	1	9
p-value*	1	4
% correctly classified cases, averaged over repeats	1	4
Nagelkerke's pseudo-R^2	1	4
t-test comparisons between models	1	8

TABLE A.8: model performance. *The p-value "indicates that XGBoost algorithm is performing better than a random predictor simply predicting the majority class" (Trakadis et al., 2019). ROC: receiver operating characteristic, AUC: area under the ROC curve, TRP: true positive rate, TNR: true negative rate, PPV: positive predictive value. As many studies reported multiple measures, percentages do not combine to 100.

Methods	Studies	Models
a-priori	1	9 (22%)
Unclear	3	6 (15%)
Unreported	5	26 (63%)

TABLE A.9: method for choosing decision threshold when reporting classification metrics. Studies which were unclear either reported a general outline of how classification works for a given method, without stating this was used in the current implementation, or reported the use of 0.5 as the threshold but not how the number was chosen. Percentages are taken from the total number of models which reported classification measures, 41, and rounded to the nearest integer. Number of studies does not sum to 13 as not all studies reported classification metrics.

Study	Method	Data	Modifications	n/N	p/P	Imbalance	EPV	Risk allele	Sensitivity	Specificity	Validation
a	AB	SNP	Y	20/40	150/367	1	0.0054	NR	0.7175	0.76	CV
a	SVM	SNP	N	20/40	367/367	1	0.0054	NR	0.4	0.4	CV
b	NN	SNP	Y	487/942	15/15	0.93	32.5	NR	0.8275	0.6395	CV
c	NN	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	NN	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	EC	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	NN	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	MDR	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	BN	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	NB	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	SVM	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	DTb	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	DTNB	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	BFT	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
c	AB	SNP	N	260/614	40-48/40-48*	1.36	5.42-6.5*	NR	NR	NR	CV
d	NN	SNP/GE	N	355/710	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	Y	355/710	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	Y	355/710	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	N	94/188	NCR/NR	1	n/a	NR	NR	NR	CV

Study	Method	Data	Modifications	n/N	p/P	Imbalance	EPV	Risk allele	Sensitivity	Specificity	Validation
d	NN	SNP/GE	Y	94/188	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	Y	94/188	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	N	31/62	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	Y	31/62	NCR/NR	1	n/a	NR	NR	NR	CV
d	NN	SNP/GE	Y	31/62	NCR/NR	1	n/a	NR	NR	NR	CV
e	SVM	SNP	N	2191/3625	3514/NCR	0.65	n/a	NR	NR	NR	Ext
e	SVM	SNP	N	2191/3625	14632/NCR	0.65	n/a	NR	NR	NR	Ext
e	SVM	SNP	N	2191/3625	1252/NCR	0.65	n/a	NR	NR	NR	Ext
e	SVM	SNP	N	2191/3625	5366/NCR	0.65	n/a	NR	NR	NR	Ext
e	NN	SNP	N	2191/3625	3514/NCR	0.65	n/a	NR	NR	NR	Ext
e	NN	SNP	N	2191/3625	14632/NCR	0.65	n/a	NR	NR	NR	Ext
e	NN	SNP	N	2191/3625	1252/NCR	0.65	n/a	NR	NR	NR	Ext
e	NN	SNP	N	2191/3625	5366/NCR	0.65	n/a	NR	NR	NR	Ext
e	RF	SNP	N	2191/3625	3514/NCR	0.65	n/a	NR	NR	NR	Ext
e	RF	SNP	N	2191/3625	14632/NCR	0.65	n/a	NR	NR	NR	Ext
e	RF	SNP	N	2191/3625	1252/NCR	0.65	n/a	NR	NR	NR	Ext
e	RF	SNP	N	2191/3625	5366/NCR	0.65	n/a	NR	NR	NR	Ext
e	BN	SNP	N	2191/3625	3514/NCR	0.65	n/a	NR	NR	NR	Ext
e	BN	SNP	N	2191/3625	14632/NCR	0.65	n/a	NR	NR	NR	Ext
e	BN	SNP	N	2191/3625	1252/NCR	0.65	n/a	NR	NR	NR	Ext

Study	Method	Data	Modifications	n/N	p/P	Imbalance	EPV	Risk allele	Sensitivity	Specificity	Validation
e	BN	SNP	N	2191/3625	5366/NCR	0.65	n/a	NR	NR	NR	Ext
f	NN	Exome	Y	200/400	∼1000/>500000*	1	<0.0004*	Ref/alt	0.64	NR	Split
f	RF	Exome	N	200/400	∼1000/>500000*	1	<0.0004*	Ref/alt	0.55	NR	Split
f	DT	Exome	N	200/400	∼1000/>500000*	1	<0.0004*	Ref/alt	0.54	NR	Split
g	RF	SNP	N	604/2371	693/761830	2.93	0.00079	NR	0.998	NR	App.
g	NB	SNP	N	483/1414	693/761830	1.93	0.00063	NR	0.734	NR	Split
g	k-NN	SNP	N	483/1414	693/761830	1.93	0.00063	NR	0.954	NR	Split
g	MDR	SNP	N	604/2371	693/761830	2.93	0.00079	NR	0.664	NR	CV
g	MDR	SNP	N	604/2371	693/761830	2.93	0.00079	NR	0.883	NR	CV
h	Ridge	SNP	N	653/1158	298604/298604	0.77	0.0022	NR	NR	NR	CV
h	SVM	SNP	N	653/1158	98604/298604	0.77	0.0022	NR	NR	NR	CV
h	LASSO	SNP	N	653/1158	98604/298604	0.77	0.0022	NR	NR	NR	CV
h	Ridge	SNP	N	1170/2068	98604/298604	0.77	0.0039	NR	NR	NR	CV
h	SVM	SNP	N	1170/2068	98604/298604	0.77	0.0039	NR	NR	NR	CV
h	LASSO	SNP	N	1170/2068	98604/298604	0.77	0.0039	NR	NR	NR	CV
i	LASSO	SNP	N	1341/4402	1486/317481*	2.28	>=0.0042*	NR	0.11	0.97	Split†
i	SVM	SNP	N	1341/4402	1486/317481*	2.28	>=0.0042*	NR	NR	NR	Split†
i	GBM	SNP	N	1341/4402	1486/317481*	2.28	>=0.0042*	NR	NR	NR	Split†
j	LASSO	Exome	N	1782*/3564	1155/17138	1	0.1*	NR	0.720	0.773	Split
j	SVM	Exome	N	1782*/3564	1155/17138	1	0.1*	NR	0.708	0.706	Split

Study	Method	Data	Modifications	n/N	p/P	Imbalance	EPV	Risk allele	Sensitivity	Specificity	Validation
j	RF	Exome	N	1782*/3564	1155/17138	1	0.1*	NR	0.820	0.813	Split
j	GBM	Exome	N	1782*/3564	1155/17138	1	0.1*	NR	0.849	0.866	Split
k	CIF	CNV	N	1570/3486	21/21	1.22	74.6	NR	NR	NR	CV
k	RF	CNV	N	1570/3486	21/21	1.22	74.6	NR	NR	NR	CV
k	SVM	CNV	N	1570/3486	21/21	1.22	74.6	NR	NR	NR	CV
k	NN	CNV	N	1570/3486	21/21	1.22	74.6	NR	NR	NR	CV
l	NN	PRS	N	5018/10859	19/116	1.16	43.26	NR	NR	NR	Split/Ext.
l	NN	PRS	N	5018/10859	116/116	1.16	43.26	NR	NR	NR	Split/Ext.
l	NN	PRS	N	5018/10859	14/29-32*	1.16	156.81-173.03*	NR	NR	NR	Split/Ext.
l	NN	PRS	N	5018/10859	26/29-32*	1.16	156.81-173.03*	NR	NR	NR	Split/Ext.
m	SVM	SNP	N	3446/7731	125/125	1.24	27.57	Ref	NR	NR	CV
m	SVM	SNP	N	5554/11853	125/125	1.13	44.43	Ref	NR	NR	CV
m	SVM	SNP	N	3446/7731	4998/4998	1.24	0.69	Ref	NR	NR	CV
m	SVM	SNP	N	5554/11853	4998/4998	1.13	1.11	Ref	NR	NR	CV
m	SVM	SNP	N	3446/7731	125/125	1.24	27.57	Ref	NR	NR	CV
m	SVM	SNP	N	5554/11853	125/125	1.13	44.43	Ref	NR	NR	CV
m	SVM	SNP	N	3446/7731	4998/4998	1.24	0.69	Ref	NR	NR	CV
m	SVM	SNP	N	5554/11853	4998/4998	1.13	1.11	Ref	NR	NR	CV

TABLE A.10: overview of prediction models. n: number of cases used in model development in final model, N: number of total observations in model

TABLE A.10: development in final model, p: number of predictors in final model, P: number of candidate predictors, EPV: events per candidate variable/predictor, NR: not reported, NCR: not clearly reported, Ref: risk allele coded as reference allele, Alt: coded as alternative allele, SNP: single nucleotide polymorphism, CNV: copy number variant, PRS: polygenic risk score, GE: gene expression, AB: AdaBoost, SVM: support vector machine, NN: neural network, EC: evolutionary computation, MDR: multifactor dimensionality reduction, BN: Bayesian networks, NB: naïve Bayes, DTb: decision tables, DTNB: decision table naïve Bayes, BFT: best-first tree (BFTree), RF: random forest, DT: decision tree, *k*-NN: *k*-nearest neighbours, LASSO: least absolute shrinkage and selection operator, GBM: gradient boosting machine, CIF: conditional inference forests, CV: cross-validation, n/a: not applicable. †Study used a roughly equal 3-way split for predictor selection, training and testing, where 10-fold CV was used in the training fold (Guo et al., 2015). Splits were repeated, but reported AUCs in the main text are for only one of the repeats; the study is recorded here as split-sample. *Number reported is unclear; upper and lower bounds, or an approximation given by the authors in the text are used. Where insufficient information is provided to give a reasonable approximation for predictors, NCR or NR is recorded. Imbalance refers to class imbalance, given here as number of controls divided by number of cases in model development. Modification refers to whether a classifier was used "out-of-the-box", N, or was modified in some way, Y. Validation is *k*-fold CV, split-sample (Split), apparent (App.) or external (Ext.). A single study reported internal validation (split-sample) and external validation (but with partial sample overlap) (Chen et al., 2018). Studies: a (Yang et al., 2010a), b (Ghafouri-Fard et al., 2019), c (Aguiar-Pulido et al., 2010; Aguiar-Pulido et al., 2013), d (Wang et al., 2018), e (Pirooznia et al., 2012), f (Laksshman et al., 2017), g (Acikel et al., 2016), h (Li et al., 2014), i (Guo et al., 2015), j (Trakadis et al., 2019), k (Engchuan et al., 2015), l (Chen et al., 2018), m (Vivian-Griffiths et al., 2019).

A.2.8 Predictors

Coding	Studies	Models
Reported	6	35 (45%)
Continuous (weighted average of additive SNPs; PRS)	1	4
Counts of genes per gene set (CNV)	1	4
Counts of variants per gene (Exome)	1	4
Additive model (0, 1, 2), missing coded as 3 (SNP)	1	12
Z-transformation of additive model (0, 1, 2; SNP)	1	8
One-hot encoded (SNP)	1	3
Unclear/unreported	7	42 (55%)
Unclear[1]	2	3
Not reported	5	39

TABLE A.11: coding of predictors. [1]Coding implied through description as 'ordinal' or through an abstract description of the type of classifier, but not clear.

Methods	Studies	Models
Additional knowledge used	9	49 (64%)
Predictors	8	43
Array not genome-wide	3	15
Predictors only from brain-expressed genes	1	8
Selection by *p*-value cut-off from external GWAS	1	8
Annotation of gene and variant-type	1	4
Annotation of gene and gene set	1	4
Choice of phenotypes and weights from GWAS for SZ-PRS	1	4
Modelling	1	6
Non-zero matrix weights in cRBM determined from GE data	1	6
Unclear/unreported	6	28 (36%)
Not clear	1	3
Not reported	5	25

TABLE A.12: explicit use of additional knowledge in selecting or weighting of predictors and modelling. Implicit knowledge, such as choice of a linear machine learning method, or additive encoding of genotyping data, are not included. GE: gene expression, cBRM: conditional restricted Boltzmann machine.

Type	Studies	Models
Filter	8	48 (62%)
Association test in external dataset, clumping	1	8
Association test in current dataset, clumping	1	8

Type	Studies	Models
Association test in current dataset for brain-expressed genes only, clumping	1	8
Association test in split of current dataset, *p*-value cut-off	1	3
Pruning, association test in current dataset, *p*-value cut-off	1	5
Embedded (LASSO/RF/GBM combined)[1]	1	4
Embedded (LASSO) with *p*-value cut-off	1	2
Forward sequential feature selection (FSFS)[2]	1	1
Correlation with outcome or intermediate phenotype	1	9
Embedded	8	20 (26%)
Regression (LASSO)	3	4
Tree-based	7	13
RF (including CIF)	4	8
Boosting (GBM, AdaBoost)	3	3
DT	2	2
Other	2	3
DTb	1	1
DTNB	1	1
Feature-selective AdaBoost[3]	1	1
Unclear[4]	1	3 (4%)
None reported	6	18 (23%)

TABLE A.13: predictor selection technique. [1]Trakadis et al. (2019) report predictors being selected "in combination of" embedded methods, but do not state how such methods were combined (Trakadis et al., 2019). [2]FSFS is a wrapper on an embedded method, used as a filter. [3]Yang et al. (2010) modified AdaBoost to include univariable predictor selection within each iteration before training each weak learner (Yang et al., 2010a); as the modification is within each iteration it is listed as "embedded" here. This is counted once under feature-selective AdaBoost, and is not counted under 'Boosting'. dLaksshman et al. (2017) report using "L1-based feature selection" but no indication about what method the L_1-norm was applied to (Laksshman et al., 2017). LASSO: least absolute shrinkage and selection operator, RF: random forest, GBM: gradient-boosting machine, DTNB: decision table-naïve Bayes, DTb: decision table, DT: decision tree, CIF: conditional inference forest. Several models exploited both filter and embedded methods; these are counted in both sections.

A.2.9 Hyperparameters

Methods	Studies	Models
Reported	6	26 (34%)
SVM (RBF)		
C	2	9
Gamma	2	9
AdaBoost[1]		

Methods	Studies	Models
Iterations	1	1
Neural Networks		
Epochs	2	12
Optimiser	1	4
Activation function	1	4
Layers	1	4
LASSO		
Lambda	1	1
Unclear/unreported	9	51 (66%)
Not clearly reported	2	5
Not reported	8	46

TABLE A.14: hyperparameters tuned during model training. [1]Feature-selective AdaBoost (Yang et al., 2010a). Manual experiments with different hyperparameters are presented by Engchuan et al. (2015) in the supplementary: these are included as "not reported", as they appear to be post-hoc experiments rather than a search as part of learning (Engchuan et al., 2015). Several studies report either hyperparameter search method, or the hyperparameters that were tuned, but not both (see Table 3.5). A study (16 models) used the default hyperparameters (Table 3.5) and is counted here under 'not reported' (Pirooznia et al., 2012).

Appendix B

A Simulation Study of Binary Classification of Complex Traits

B.1 Methods

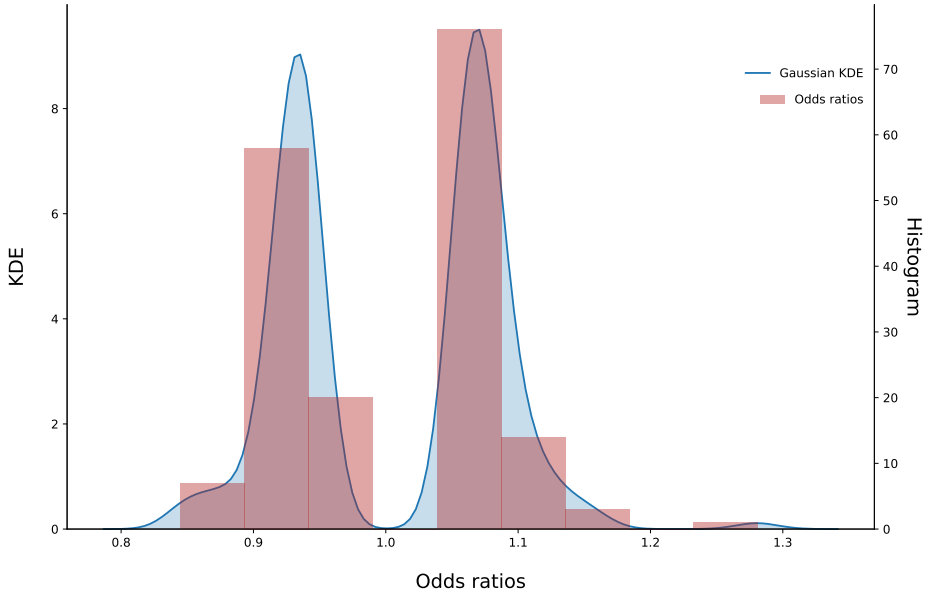


FIGURE B.1: Estimated probability density function (PDF) applied to genome-wide significant odds ratios from (Pardiñas et al., 2018) by kernel density estimation. The "gaussian_kde" class from SciPy's "stats" module was used, with a closer fit achieved by setting "bw_method" to 0.1. Odds ratios for simulations were drawn from the estimated PDF using the class's "resample" method.

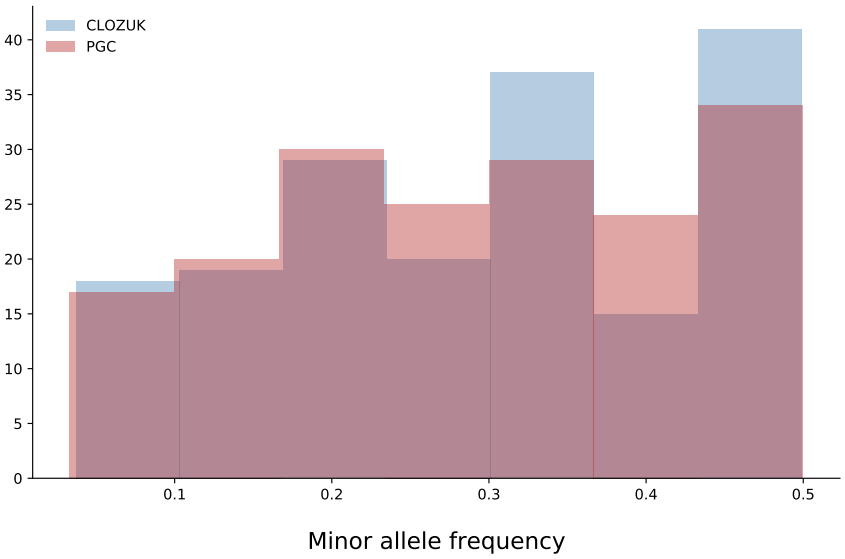


FIGURE B.2: Distribution of minor allele frequencies reported from psychiatric genetics consortium (PGC)2 and CLOZUK data for genome-wide significant SNPs (Pardiñas et al., 2018).

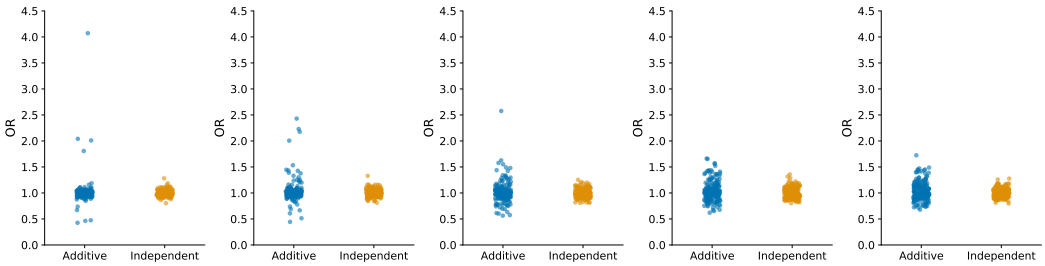


FIGURE B.3: Example distributions of odds ratios with $m \in \{0.05, 0.25, 0.5, 0.75, 1\}$ under additive and independent simulations. For both simulation types, $n = 5000$ and $p = 200$. Additive simulations set $h^2 = 0.2$ and $k = 0.0025$. Independent simulations used odds ratios drawn from an estimated PDF. Empirical odds ratios, rather than values set during simulations, are calculated on the observed scale. Odds ratios are shown for a single simulation for each scenario.

B.2 Results

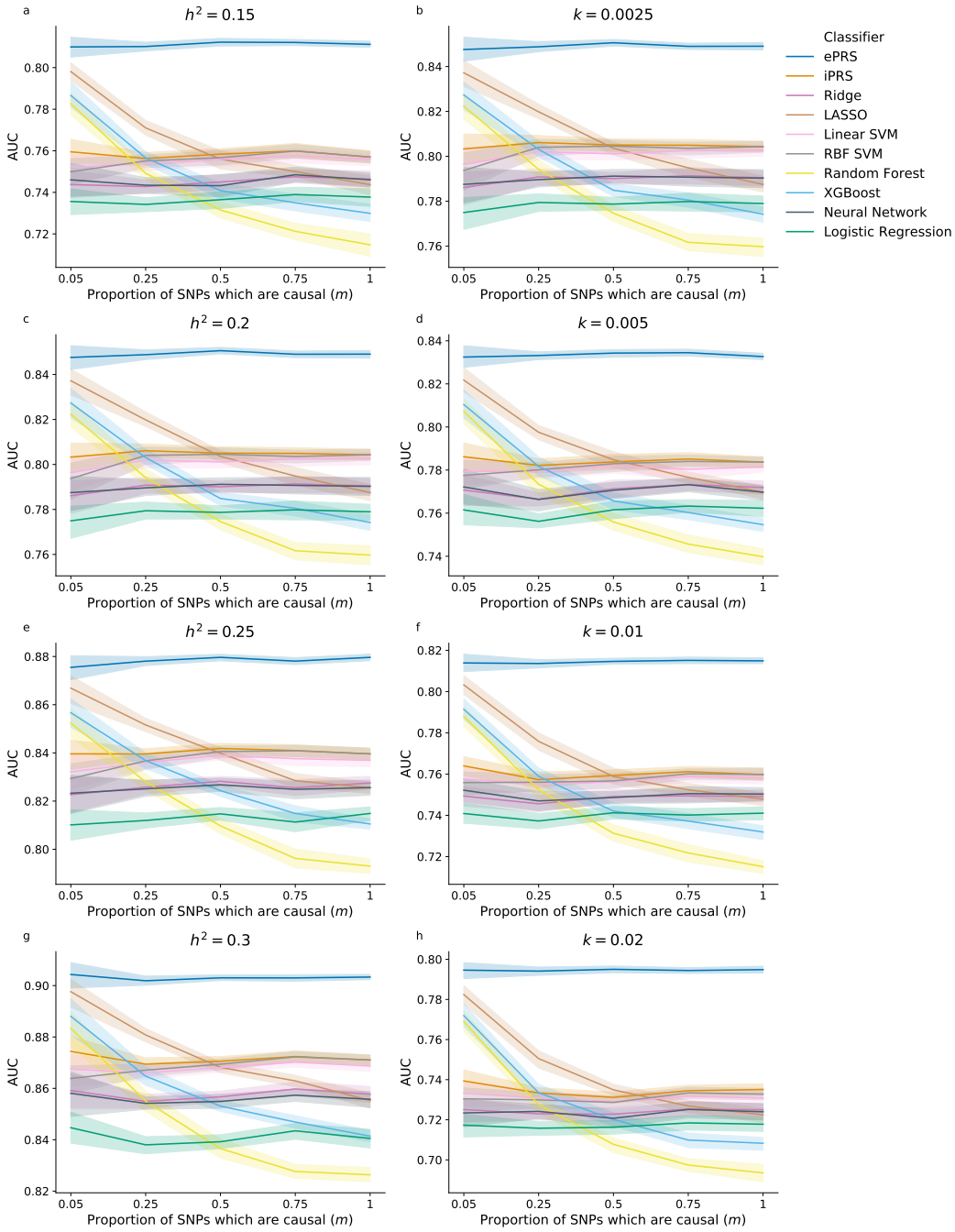


FIGURE B.4: Replication of simulations increasing m when $p < n$ for different values of h^2 and k. $p = 200$, $n = 1000$. Relationships seen in Figure 4.10 and clearly repeated across heritabilities and prevalences.

θ	MAF	P(AABB|case)	P(AABB|control)	N case	N control
0.1	0.1	0.00015	0.0001	0.29	0.2
0.1	0.2	0.0023	0.0016	4.7	3.2
0.1	0.3	0.012	0.0081	24	16
0.1	0.4	0.037	0.026	75	51

θ	MAF	P(AABB|case)	P(AABB|control)	N case	N control
0.1	0.5	0.092	0.062	1.8e+02	1.2e+02
0.2	0.1	0.00021	0.0001	0.41	0.2
0.2	0.2	0.0033	0.0016	6.6	3.2
0.2	0.3	0.017	0.0081	34	16
0.2	0.4	0.053	0.026	1.1e+02	51
0.2	0.5	0.13	0.062	2.6e+02	1.2e+02
0.3	0.1	0.00029	0.0001	0.57	0.2
0.3	0.2	0.0046	0.0016	9.1	3.2
0.3	0.3	0.023	0.0081	46	16
0.3	0.4	0.073	0.026	1.5e+02	51
0.3	0.5	0.18	0.062	3.6e+02	1.2e+02
0.4	0.1	0.00038	0.0001	0.77	0.2
0.4	0.2	0.0061	0.0016	12	3.2
0.4	0.3	0.031	0.0081	62	16
0.4	0.4	0.098	0.026	2e+02	51
0.4	0.5	0.24	0.062	4.8e+02	1.2e+02
0.5	0.1	0.00051	0.0001	1	0.2
0.5	0.2	0.0081	0.0016	16	3.2
0.5	0.3	0.041	0.0081	82	16
0.5	0.4	0.13	0.026	2.6e+02	51
0.5	0.5	0.32	0.062	6.3e+02	1.2e+02

TABLE B.1: Probability of observing the double homozygote, AABB, under a multiplicative model for $\theta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and MAF $\in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, where A and B are the risk alleles. N case and N control give the expected number of cases and controls when $n = 2000$.

θ	MAF	P(AABB|case)	P(AABB|control)	N case	N control
0.1	0.1	0.00011	0.0001	0.22	0.2
0.1	0.2	0.0018	0.0016	3.5	3.2
0.1	0.3	0.0089	0.0081	18	16
0.1	0.4	0.028	0.026	56	51
0.1	0.5	0.069	0.062	1.4e+02	1.2e+02
0.2	0.1	0.00012	0.0001	0.24	0.2
0.2	0.2	0.0019	0.0016	3.8	3.2
0.2	0.3	0.0097	0.0081	19	16
0.2	0.4	0.031	0.026	61	51
0.2	0.5	0.075	0.062	1.5e+02	1.2e+02
0.3	0.1	0.00013	0.0001	0.26	0.2
0.3	0.2	0.0021	0.0016	4.2	3.2

θ	MAF	P(AABB|case)	P(AABB|control)	N case	N control
0.3	0.3	0.011	0.0081	21	16
0.3	0.4	0.033	0.026	67	51
0.3	0.5	0.081	0.062	1.6e+02	1.2e+02
0.4	0.1	0.00014	0.0001	0.28	0.2
0.4	0.2	0.0022	0.0016	4.5	3.2
0.4	0.3	0.011	0.0081	23	16
0.4	0.4	0.036	0.026	72	51
0.4	0.5	0.087	0.062	1.8e+02	1.2e+02
0.5	0.1	0.00015	0.0001	0.3	0.2
0.5	0.2	0.0024	0.0016	4.8	3.2
0.5	0.3	0.012	0.0081	24	16
0.5	0.4	0.038	0.026	77	51
0.5	0.5	0.094	0.062	1.9e+02	1.2e+02

TABLE B.2: Probability of observing the double homozygote, AABB, under a threshold model for $\theta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and MAF $\in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, where A and B are the risk alleles.

θ	MAF	P(AABB|case)	P(AABB|control)	N case	N control
0.1	0.1	0.0001	0.0001	0.2	0.2
0.1	0.2	0.0016	0.0016	3.2	3.2
0.1	0.3	0.0081	0.0081	16	16
0.1	0.4	0.026	0.026	51	51
0.1	0.5	0.062	0.062	1.2e+02	1.2e+02
0.2	0.1	0.0001	0.0001	0.2	0.2
0.2	0.2	0.0016	0.0016	3.2	3.2
0.2	0.3	0.0081	0.0081	16	16
0.2	0.4	0.026	0.026	51	51
0.2	0.5	0.062	0.062	1.2e+02	1.2e+02
0.3	0.1	0.0001	0.0001	0.2	0.2
0.3	0.2	0.0016	0.0016	3.2	3.2
0.3	0.3	0.0081	0.0081	16	16
0.3	0.4	0.026	0.026	51	51
0.3	0.5	0.062	0.062	1.2e+02	1.2e+02
0.4	0.1	0.0001	0.0001	0.2	0.2
0.4	0.2	0.0016	0.0016	3.2	3.2
0.4	0.3	0.0081	0.0081	16	16
0.4	0.4	0.026	0.026	51	51
0.4	0.5	0.062	0.062	1.2e+02	1.2e+02

θ	MAF	P(AABB|case)	P(AABB|control)	N case	N control
0.5	0.1	0.0001	0.0001	0.2	0.2
0.5	0.2	0.0016	0.0016	3.2	3.2
0.5	0.3	0.0081	0.0081	16	16
0.5	0.4	0.026	0.026	51	51
0.5	0.5	0.062	0.062	1.2e+02	1.2e+02

TABLE B.3: Probability of observing the double homozygote, AABB, under M170 XOR, M78 XOR and M68 interference models for $\theta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and MAF $\in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, where A and B are the risk alleles. Expected counts when $n = 2000$ are unaffected by θ as AABB genotype combinations do not increase risk under XOR and interference models.

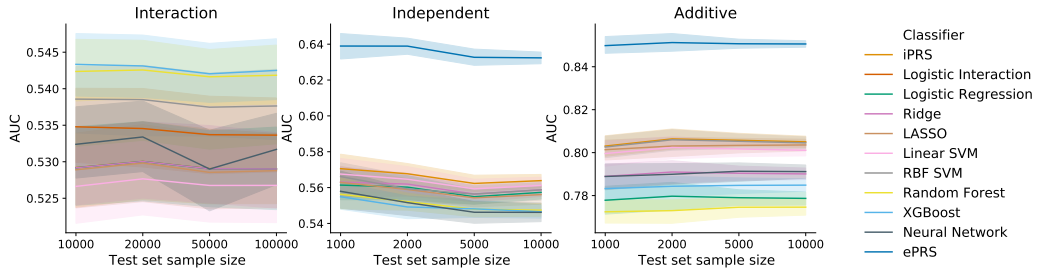


FIGURE B.5: AUC for classifiers trained on 2-SNP interaction models for decreasing size of the test set. To ensure any decrease in discrimination was observed, simulations set $\theta = 0.5$ and $MAF = 0.5$. However, predictive performance appears stable, suggesting the chosen test set sizes are reasonable and have not unduly influenced reported results. Interaction simulations are shown here with a two-SNP causal interaction and no noise SNPs. Independent and additive simulations use $p = 200$ and $m = 0.5$, with independent simulations using odds ratios drawn from a KDE fit to values in Pardiñas et al., 2018, and additive simulations setting $h^2 = 0.2$.

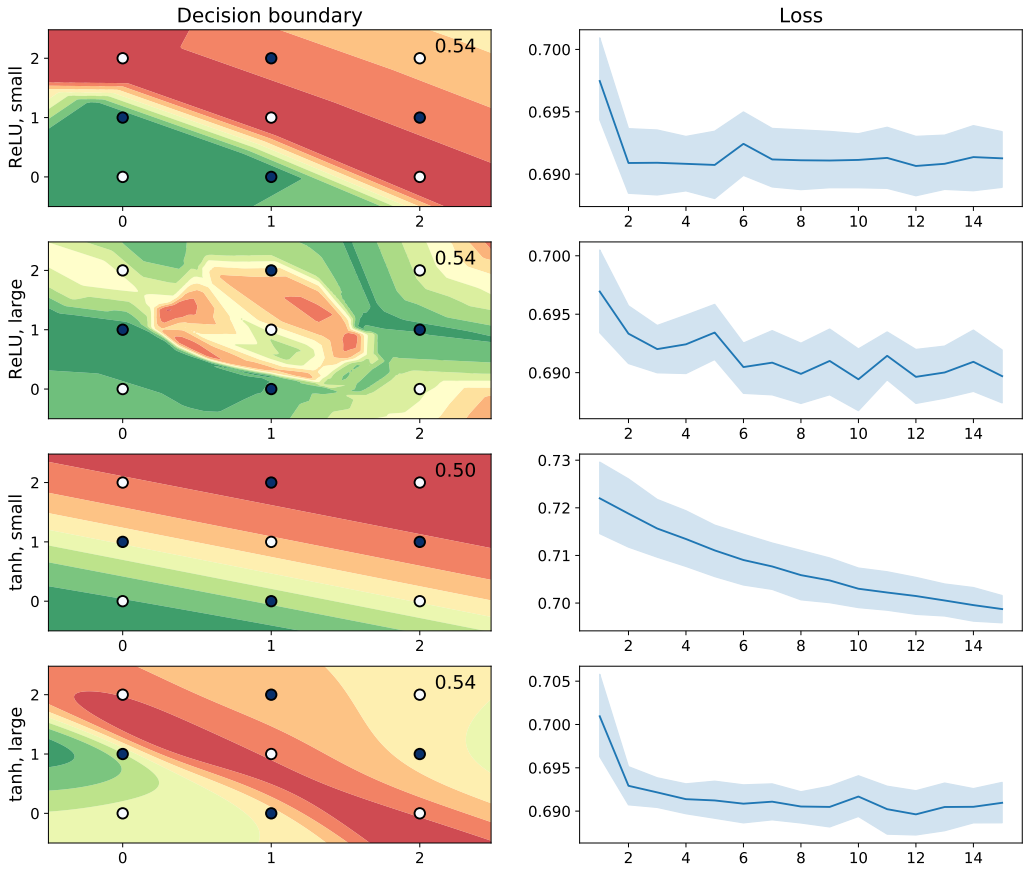


FIGURE B.6: Examination of different neural network architectures on the decision boundary for M170 XOR models for a single simulation. Contour plots of the decision boundary, and a plot of the loss function against epochs is given for each combination of ReLU or tanh and small or large architectures.

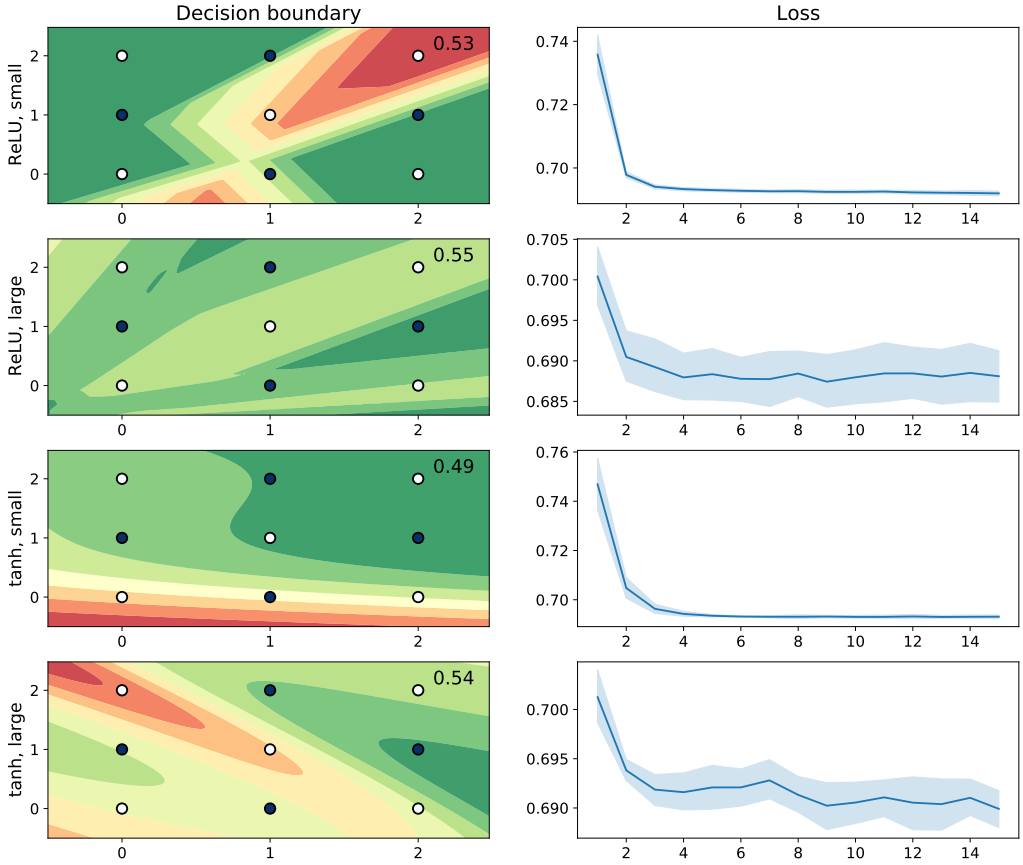


FIGURE B.7: An additional run of M170 XOR models showing differences between neural network architectures.

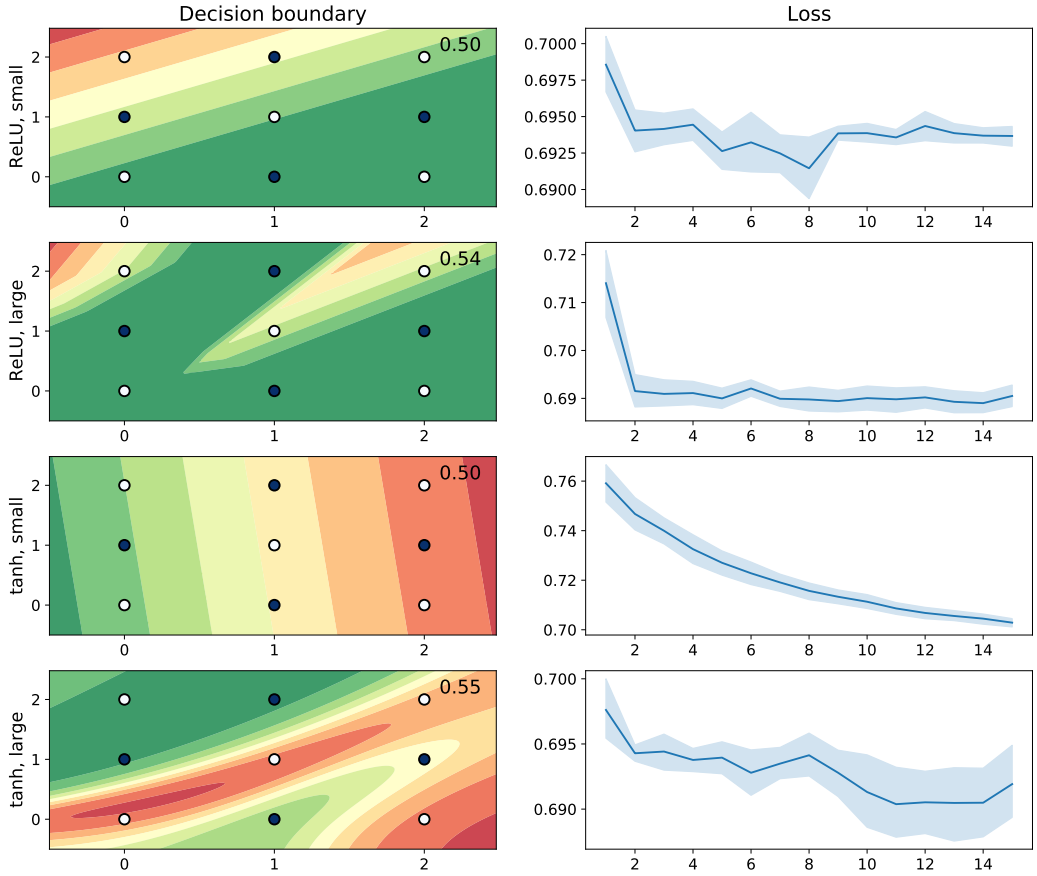


FIGURE B.8: A further run of M170 XOR models showing differences between neural network architectures.

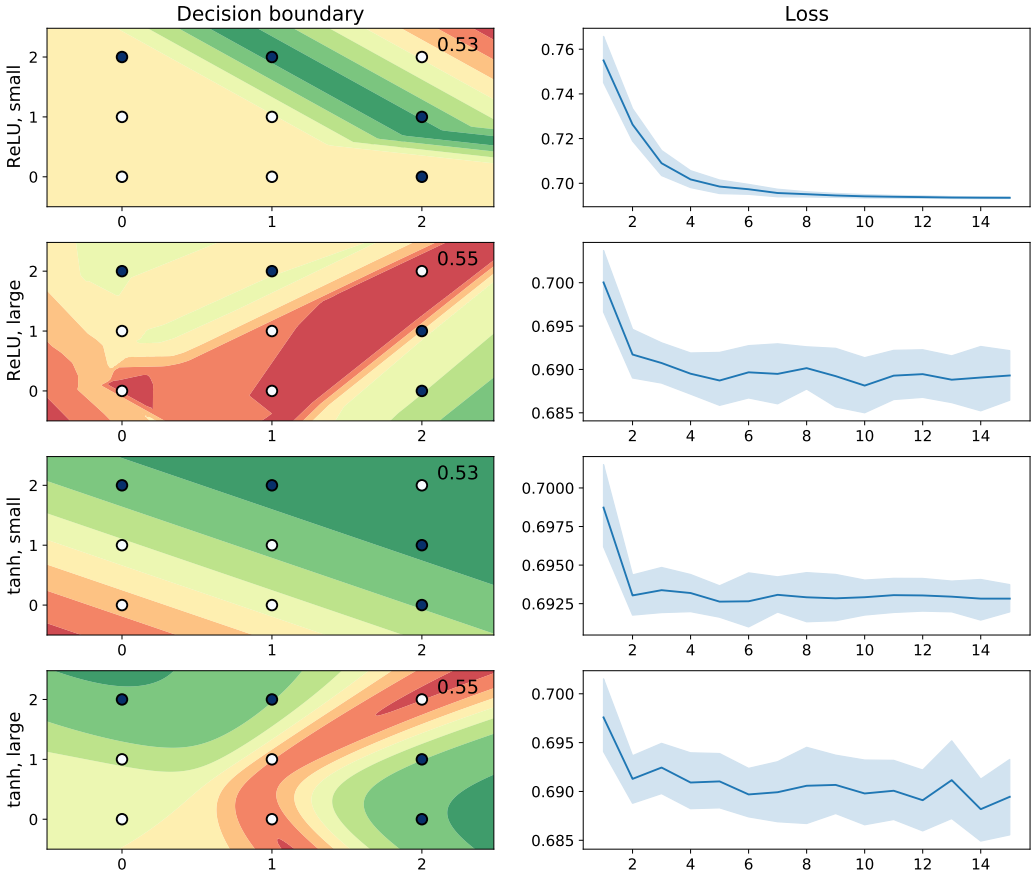


FIGURE B.9: An additional run of M78 XOR models showing differences between neural network architectures.

Appendix C

Multivariable machine learning models of schizophrenia in UK Biobank

C.1 Methods

C.1.1 Hyperparameters

Hyperparameters were searched using a Monte Carlo randomised search. All distributions were defined using SciPy and NumPy.

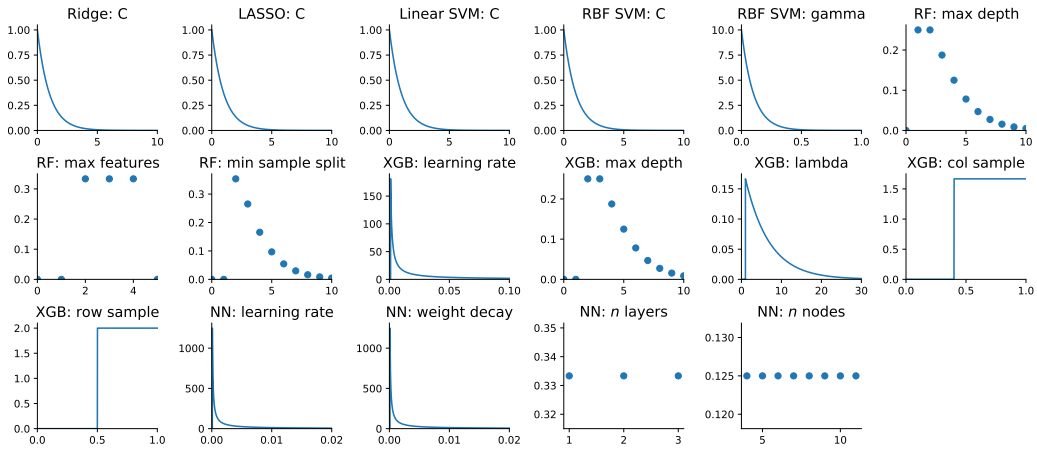


FIGURE C.1: Probability density functions (PDFs) and probability mass functions (PMFs) of hyperparameters for machine learning methods. Distributions were used as above for all models, with the exception of neural networks. While large networks of SNPs used 1-3 layers and $\frac{p}{2}$-$\frac{3}{2}p$ neurons per hidden layer, networks with $2 \leq p < 20$ used 2-6 hidden layers and p-$2p$ hidden neurons, and networks with $p = 1$ used 2-8 hidden neurons per 3-6 hidden layers. Lower limits on search for learning rate and weight decay were also decreased (from $10^{-4} to 10^{-7}$) for more narrow datasets to account for the stronger predictors available when using the 0.05-threshold ePRS or demographic predictors.

C.1.2 Generalisable model associations

C.1.2.1 Cognitive tests

Cognitive tests (category 100026) were evaluated for their ability to predict machine learning risk scores in controls. A cognitive battery was completed by participants in UK Biobank

on request. Here, tests were restricted to those completed by at least 20% of participants. Missingness varies between tests; differing numbers of participants were requested to complete each test by UK Biobank. Reaction time and pairs matching tests were completed by a large number of participants at the initial visit to assessment centres. Subsequent tests were performed at home in follow-up assessments. All tests were z-transformed within cross-validation.

Fluid intelligence score

The fluid intelligence score (field 20016) aims to measure fluid intelligence via a sum of the answers from 13 questions, coded as 1 for correct and 0 for incorrect, performed as part of touchscreen assessments. Questions assess ability to reason through tasks such as summing numbers together, identifying family relationships from descriptions and identifying the next word or number in a sequence. Values run from 0 to 13, with 0 assigned where questions were not completed.

Digit span

Numeric short term memory was assessed through touchscreen questions which showed participants a number, removed it and asked participants to recall it. The longest number which participants could recall was established by increasing the number appearing on the screen by 1 each round. Only the length of the number (digit span, field 4282) is used. Values range from 2 to 12. Abandoned tests were counted as missing.

Pairs matching

Pairs matching was assessed using another electronic card game, wherein participants were shown multiple pairs of cards and asked to recall the matching sets after cards had been turned over. The test evaluates episodic memory. Number of errors made by participants that completed the task were used after a log + 1 transformation. 338,132 complete entries were used in analysis.

Reaction time

Reaction time was assessed using an electronic form of the game Snap, where participants were asked to click as soon as a matching set of cards appeared. The test evaluates processing speed. 336,003 completed reaction time tests were retained for analysis in unsubsampled controls.

Symbol digit substitution

Symbol digit substitution was assessed using an online number-symbol matching game. The number of correct substitutions was restricted to the between 3 and 36. The test assesses complex processing speed. 82,747 completed assessments were used in analysis.

Trail making test A and B

Trail making tests were assessed using an online circle connecting task, where participants had to follow a numerical or numerical and alphabetical order for trail making tests A and

B respectively. This evaluated visual attention. Log-transformed completion time was used for 73,228 participants trail making test A and 73,226 in trail making test B in analysis.

C.1.2.2 Demographic variables

Demographic factors of interest which were completed by a large fraction of the cohort were selected for analysis. Variables were not included in the initial machine learning models as their value may be influenced by schizophrenia or its treatment.

Age
Age at recruitment was included as a common stratifier of risk (field 21022).

BMI
BMI (field 21001) is elevated in schizophrenia, which may be due to disease-associated lifestyle factors or side-effects of antipsychotic medication (Wirshing, 2004). Body mass index (BMI) was calculated by UK Biobank using height and weight measurements from initial visits to assessment centres. Values were log transformed.

Deprivation
Elevated deprivation has repeatedly been associated with schizophrenia (Allardyce and Boydell, 2006). Deprivation is measured using the Townsend deprivation index (TDI) (Townsend, Phillimore, and Beattie, 1988). TDI (field 189) was calculated by UK Biobank before participants joined the study using the most recent census prior to commencement. Scores are calculated for each individual using unemployment, non-home ownership, non-car ownership and household overcrowding. A higher scores implies higher deprivation. Values were log transformed.

Ever smoked
Individuals with schizophrenia consistently show greater likelihood of having smoked (De Leon and Diaz, 2005). Smoking status (field 20116) was given by participants at first assessment centre visits during touch screen questionnaires. Previously smoked and currently smoke categories were combined to give a single variables of 1 for ever smoked and 0 for never smoked.

Height
Standing height at baseline (field 50) was included as a negative control for genetic risk and a positive control for environmental risk (as demographic and combined models include sex as a predictor, and so assign a higher risk of schizophrenia in males).

C.1.2.3 Diseases and disorders

Neurological diseases and psychiatric disorders were derived using hospital, death and self-report records, where available. Codes used for each outcome are given in Table C.1. For all derived outcomes, including schizophrenia status, hospital records were searched in fields 41202.0.0 - 41202.0.65 for primary records, and 41204.0.0 - 41204.0.183 for secondary

records. Death records were searched in 40001.0.0 for primary cause of death, and 40002.0.0 - 40002.0.13 for secondary. All outcomes were coded as 1 if a code was present for a participant and 0 if not. Hospital and death records were search for partial matches beginning with the provided code, such that a search for F20 would match F201, F202 and all other codes under this heading. Self-report codes were checked in fields 20002.0.0 - 20002.2.33. All self-report codes were checked for an exact match only. Code for extraction of outcomes was defined in an R package, and prevalence of outcomes was checked against levels reported in the Biobank showcase for veracity. Code was also unit-tested against manually-annotated outcomes, using small subsets of UK Biobank data and synthetic tables of possible codes and fields, to ensure output was consistently correct.

C.2 Results

C.2.1 Outcome

Derived schizophrenia status consisted of both schizophrenia and schizoaffective disorder, with some participants having only self-report, only hospital records, or both (Figure C.2).

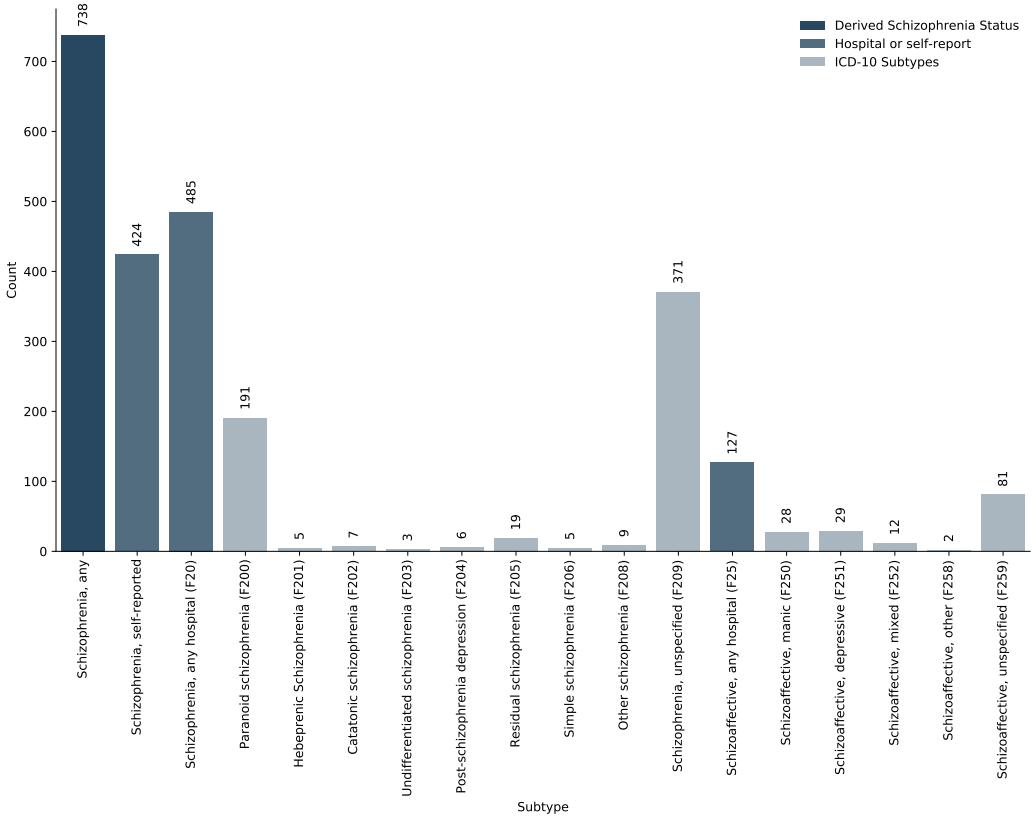


FIGURE C.2: Schizophrenia outcome subgroups. ICD-10 subtypes, shown in light blue, are present with the hospital definitions for schizophrenia and schizoaffective disorder. Categories are not exclusive.

Code	Outcome	Type
I69	Stroke	ICD-10
I68	Stroke	ICD-10
I67	Stroke	ICD-10
I66	Stroke	ICD-10
I64	Stroke	ICD-10
I63	Stroke	ICD-10
I62	Stroke	ICD-10
I61	Stroke	ICD-10
I60	Stroke	ICD-10
G46	Stroke	ICD-10
G45	Stroke	ICD-10
1583	Stroke	Self-report
1491	Stroke	Self-report
1086	Stroke	Self-report
1083	Stroke	Self-report
1082	Stroke	Self-report
1081	Stroke	Self-report
G22	Parkinson's	ICD-10
G21	Parkinson's	ICD-10
G20	Parkinson's	ICD-10
1262	Parkinson's	Self-report
G41	Epilepsy	ICD-10
G40	Epilepsy	ICD-10
1264	Epilepsy	Self-report
G30	Alzheimer's	ICD-10
1	Seen GP	Field 2090
1	Seen psychiatrist	Field 2100
F42	OCD	ICD-10
1615	OCD	Self-report
F50	Eating disorder	ICD-10
1470	Eating disorder	Self-report
F33	Depression	ICD-10
F32	Depression	ICD-10
1286	Depression	Self-report
F43	Anxiety	ICD-10
1287	Anxiety	Self-report
F90	ADHD	ICD-10

TABLE C.1: Search codes used for all psychiatric and neurological outcomes used in . Seen GP is participants answer to the question "Have you ever seen a general practitioner (GP) for nerves, anxiety, tension or depression?", for which answers are coded as 1 if true, and 0 otherwise. Similarly, seen psychiatrist is the touchscreen answer to the question "Have you ever seen a psychiatrist for nerves, anxiety, tension or depression?", where answers are coded in the same way. OCD: obsessive-compulsive disorder, ADHD: attention-deficit hyperactivity disorder, ICD-10: international classification of diseases 10.

C.2.2 Predictors

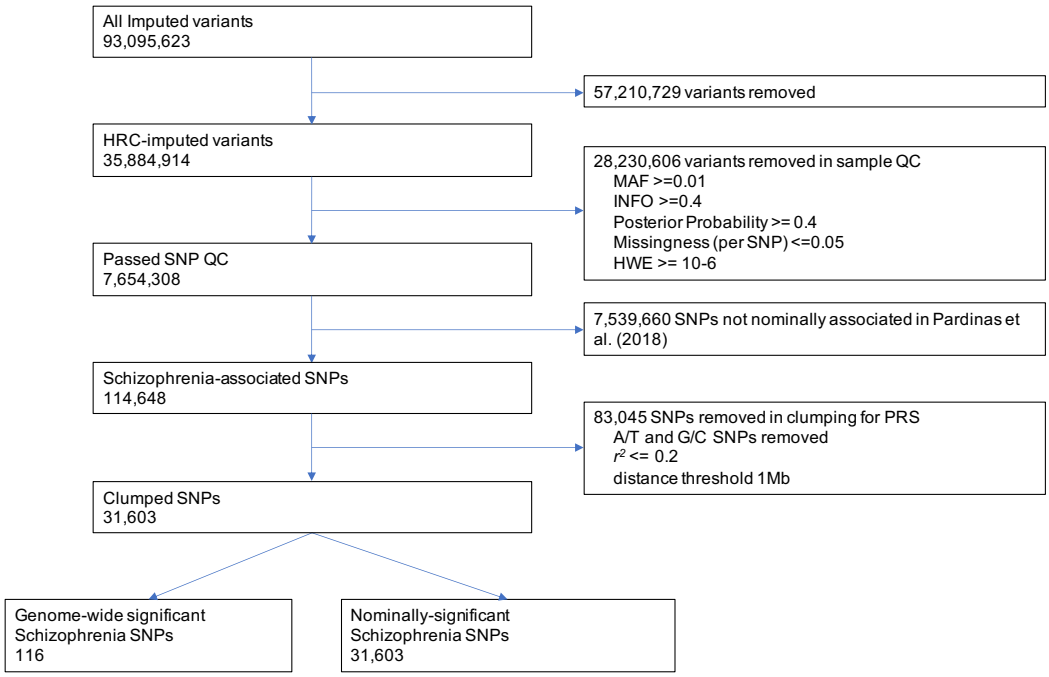


FIGURE C.3: Genetic predictor quality control pipeline. All imputed SNPs from UK Biobank were processed to derive two sets of SNPs for prediction models.

C.2.3 Missingness

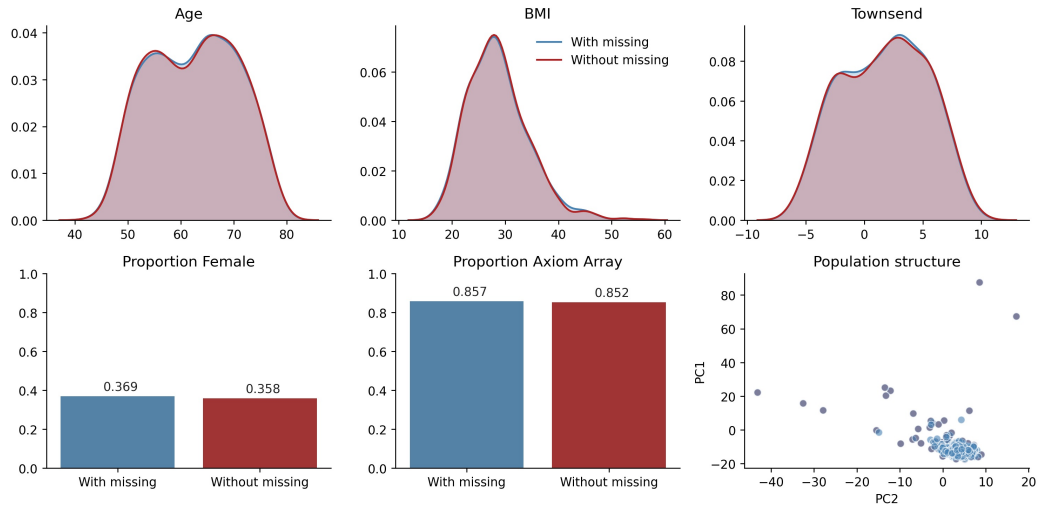


FIGURE C.4: Sample characteristics in cases with or without participants with missing values excluded. Distributions and proportions show extremely high similarity between groups.

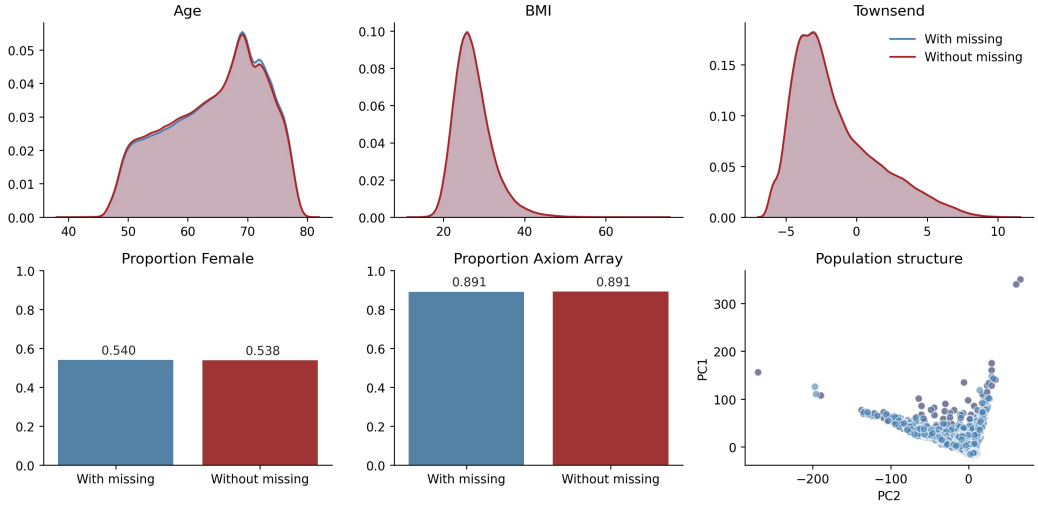


FIGURE C.5: Sample characteristics in controls with or without exclusion by missingness. Distributions and proportions show strong overlap groups. Distributions are shown for observations before subsampling.

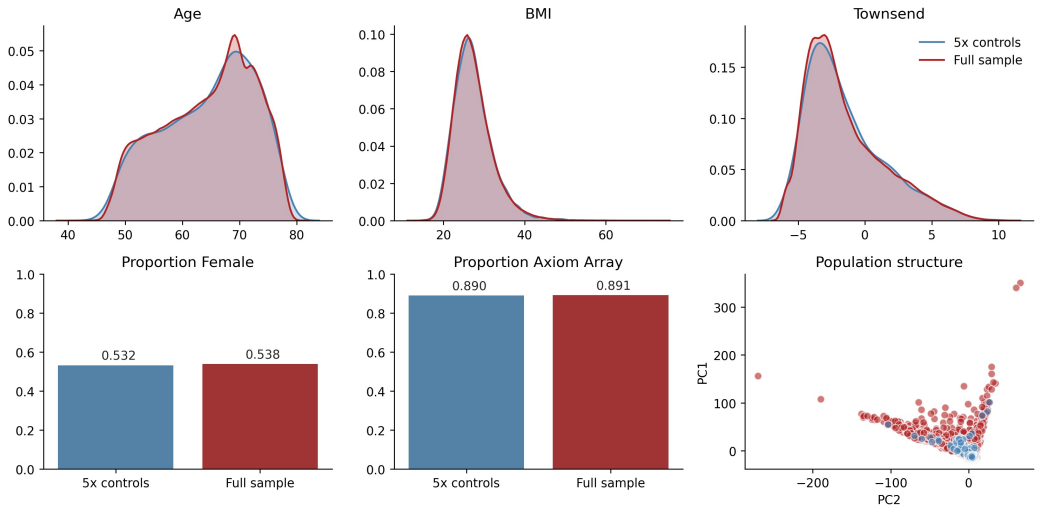


FIGURE C.6: Sample characteristics in controls with missingness removed, before and after subsampling. High similarity between the full and subsampled controls is shown.

Predictor	Test	Test-statistic	P-value (uncorrected)
Winter birth	None		
Handedness	Fisher's exact	0	1
Sex	None		
Number of siblings	Chi-squared	17.33	0.000031
Parental depression	Chi-squared	8.80	0.0030
Qualifications	Chi-squared	2.04	0.15

TABLE C.2: Tests of differential missingness for all non-genetic predictors. Tests were not run for sex or winter birth as all observations are non-missing. Fisher's exact test was used where cell counts were too low

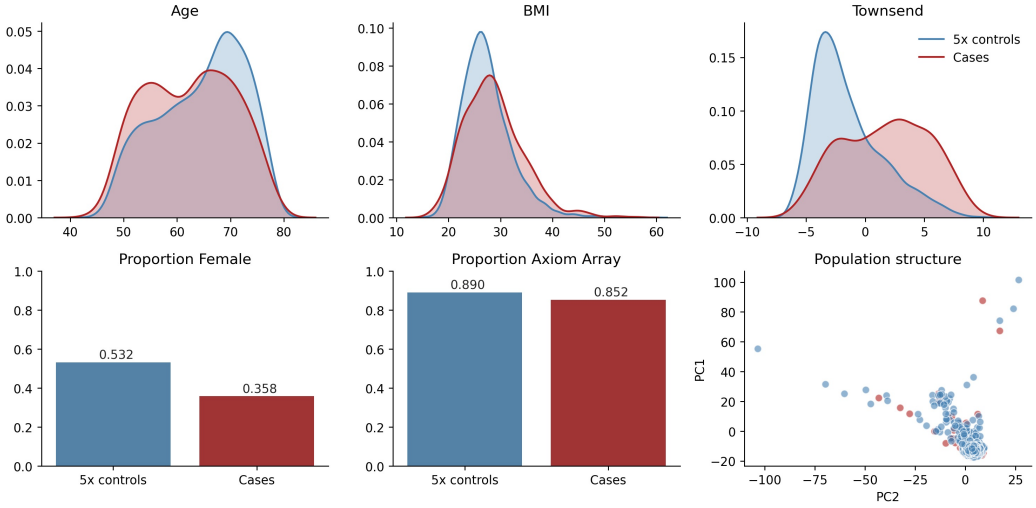


FIGURE C.7: Sample characteristics comparison between cases and controls. Age, BMI, deprivation and sex show strong differences, while genotyping aray and first two principal components are similar.

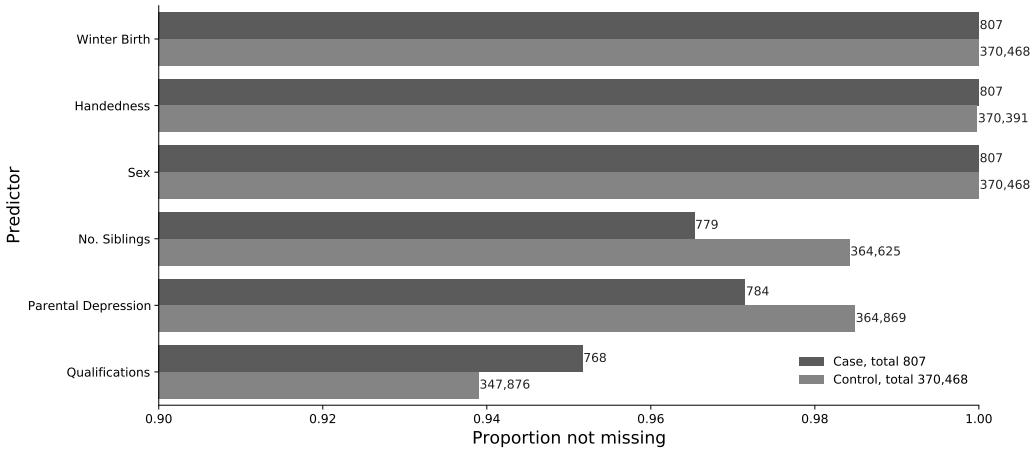


FIGURE C.8: Per-predictor missingness split by case-control status. Proportion not missing (x-axis) and total number not-missing (bar annotations) are shown.

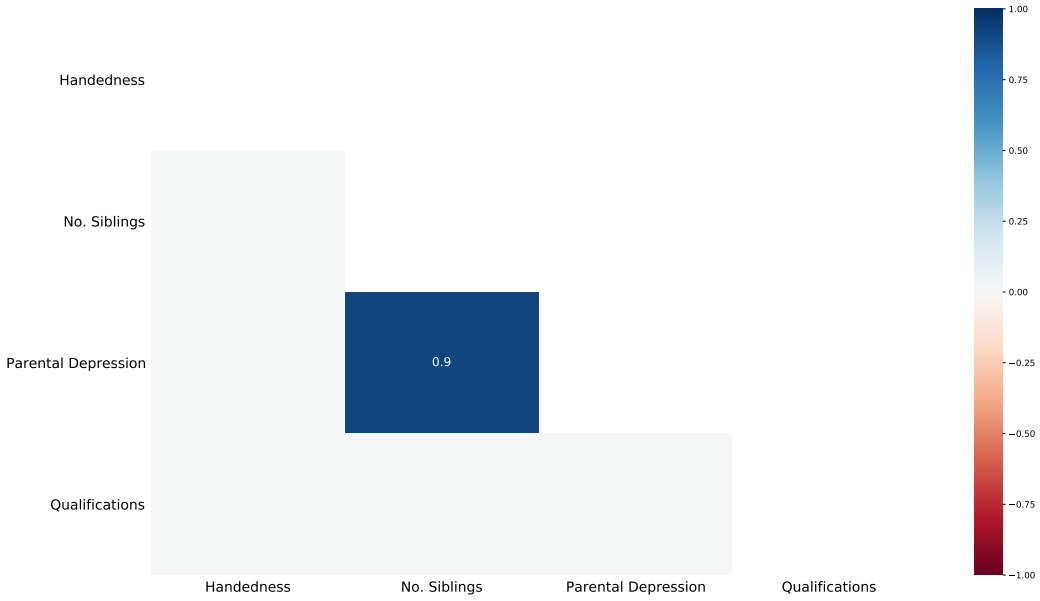


FIGURE C.9: Predictor missingness correlation. Binary-coded missingness is used to derive correlation between predictors.

C.2.4 Discrimination

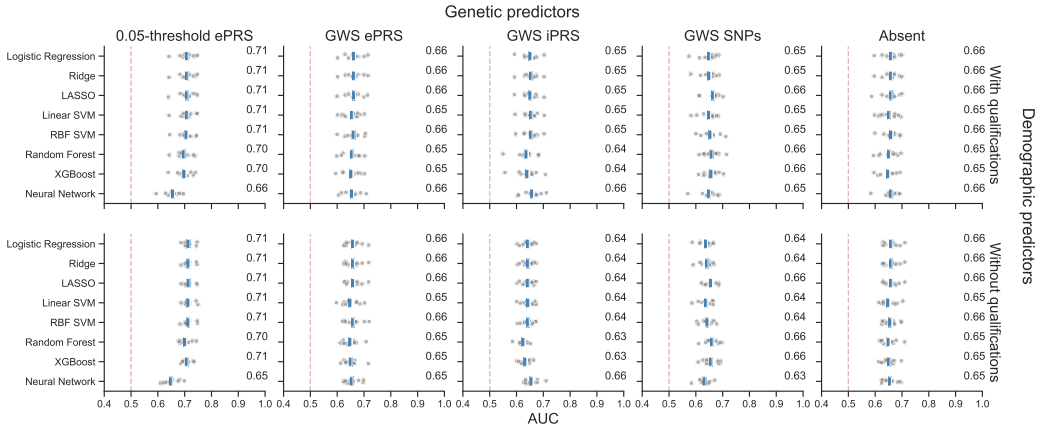


FIGURE C.10: Discrimination with and without qualifications as a demographic predictor. Mean AUC and outer-fold nested CV results are shown for all models.

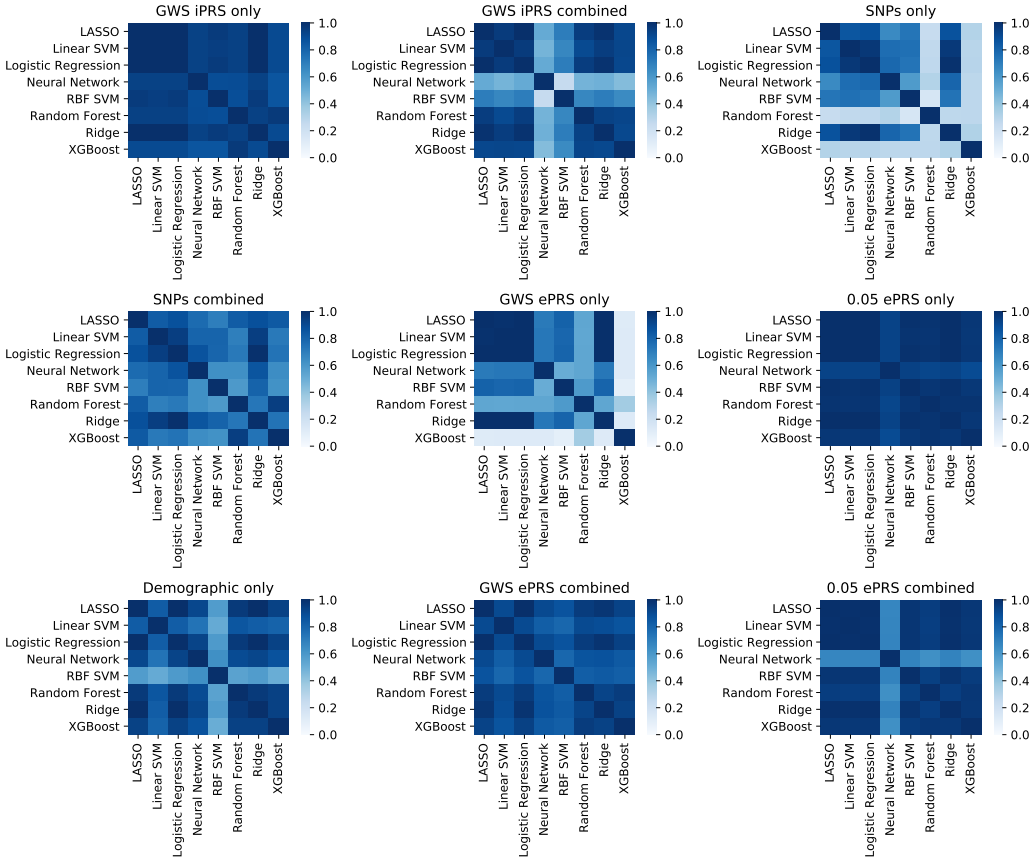


FIGURE C.11: Correlation between model predictions across nested CV folds for all models and datasets. Correlation is calculated for all corresponding test folds between model pairs and averaged over cross-validation.

Classifier	Comparison	W	p	Adjusted p	Reject H_0
XGBoost	ePRS vs. iPRS	0	0.0051	0.081	True
Random Forest	ePRS vs. iPRS	7	0.037	0.099	True
Logistic Regression	ePRS vs. SNPs	8	0.047	0.099	True
Ridge	ePRS vs. SNPs	8	0.047	0.099	True
RBF SVM	ePRS vs. SNPs	8	0.047	0.099	True
XGBoost	ePRS vs. SNPs	8	0.047	0.099	True
LASSO	ePRS vs. SNPs	9	0.059	0.099	True
Linear SVM	ePRS vs. SNPs	9	0.059	0.099	True
Logistic Regression	ePRS vs. iPRS	10	0.074	0.099	True
Ridge	ePRS vs. iPRS	10	0.074	0.099	True
LASSO	ePRS vs. iPRS	10	0.074	0.099	True
Linear SVM	ePRS vs. iPRS	10	0.074	0.099	True
RBF SVM	ePRS vs. iPRS	11	0.093	0.11	False
Neural Network	ePRS vs. iPRS	13	0.14	0.16	False
Neural Network	ePRS vs. SNPs	14	0.17	0.18	False

Classifier	Comparison	W	p	Adjusted p	Reject H_0
Random Forest	ePRS vs. SNPs	16	0.24	0.24	False

TABLE C.3: Per-model comparison of ePRS versus iPRS or SNP models of genome-wide significant SNPs only using the Wilcoxon signed-rank test in the larger sample of 807 cases before exclusion by misingness. Adjusted p-values were produced using FDR-correction at 0.1. iPRS: internal polygenic risk score, ePRS: external polygenic risk score, SNPs: single nucleotide polymorphisms.

C.2.5 Model rankings

Type	Classifier	Mean Rank	Mean AUC
Non-linear	RBF SVM	1.778	0.6301
Non-linear	Neural Network	2.333	0.6239
Non-linear	Random Forest	2.778	0.6241
Non-linear	XGBoost	3.111	0.6212
Linear	LASSO	2.056	0.633
Linear	Linear SVM	2.278	0.631
Linear	Logistic Regression	2.778	0.6317
Linear	Ridge	2.889	0.6314

TABLE C.4: Linear and non-linear models rankings across all models. Approaches are grouped by type, and sorted by mean ranking and median AUC across datasets

C.2.6 Calibration

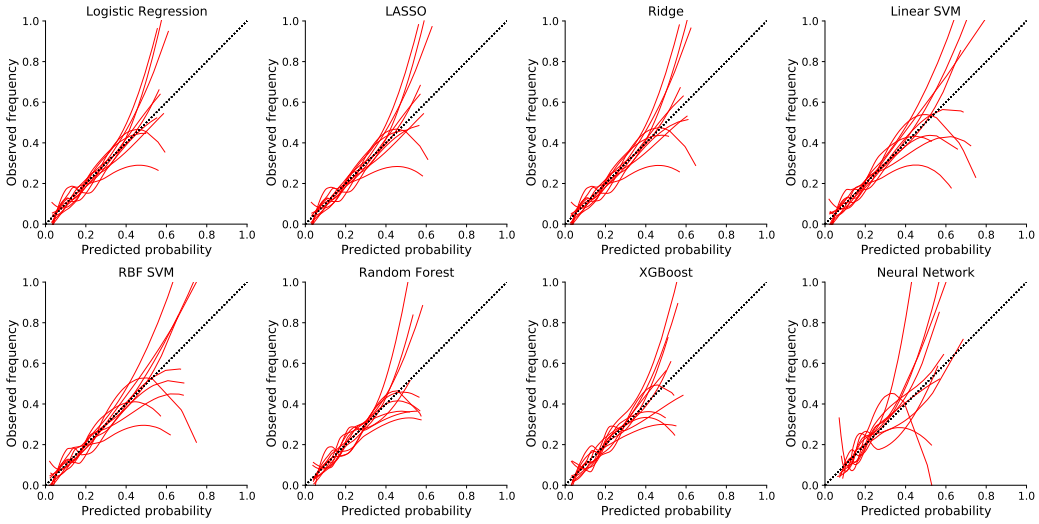


FIGURE C.12: Calibration for all models in the nested case-control sample, split by fold of cross-validation. Calibration is similar for all models across folds; variation in the upper right tail is with the expected range (Austin and Steyerberg, 2014)

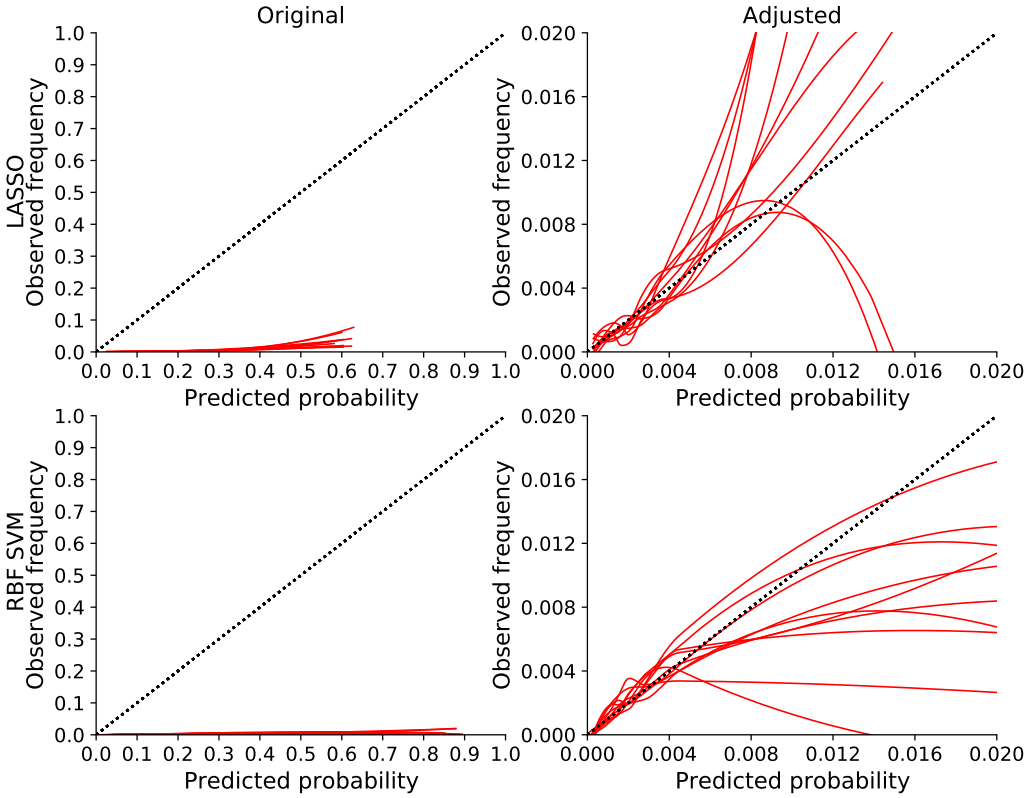


FIGURE C.13: Calibration in the whole UK Biobank sample before and after adjusting for prevalence. Due to in-memory limits in computation of loess curves in python, a subsample of 30,000 participants was used to show calibration in Figure 5.8). Here, calibration for 10 random samples of 30,000 participants is shown. Variation across samples is reasonable and demonstrates results in Figure 5.8 are representative of the whole UK Biobank sample.

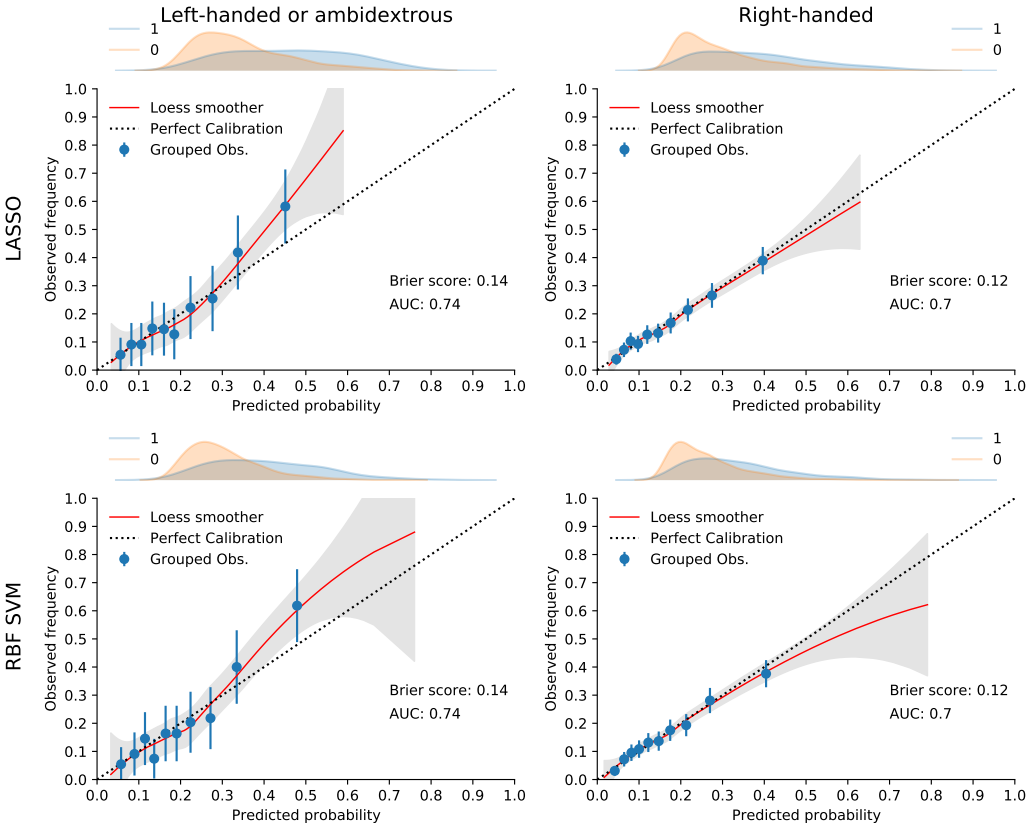


FIGURE C.14: Validation plots demonstrate calibration by handedness in the nested case-control sample.

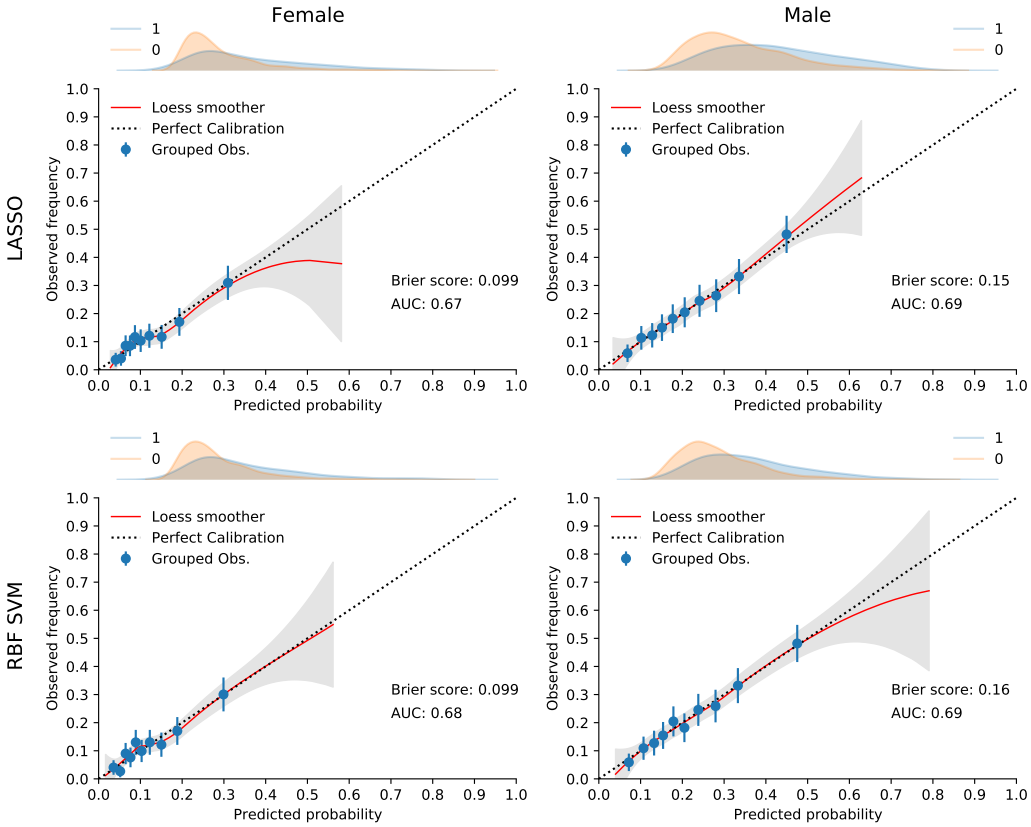


FIGURE C.15: Validation plots demonstrate calibration by sex in the nested case-control sample.

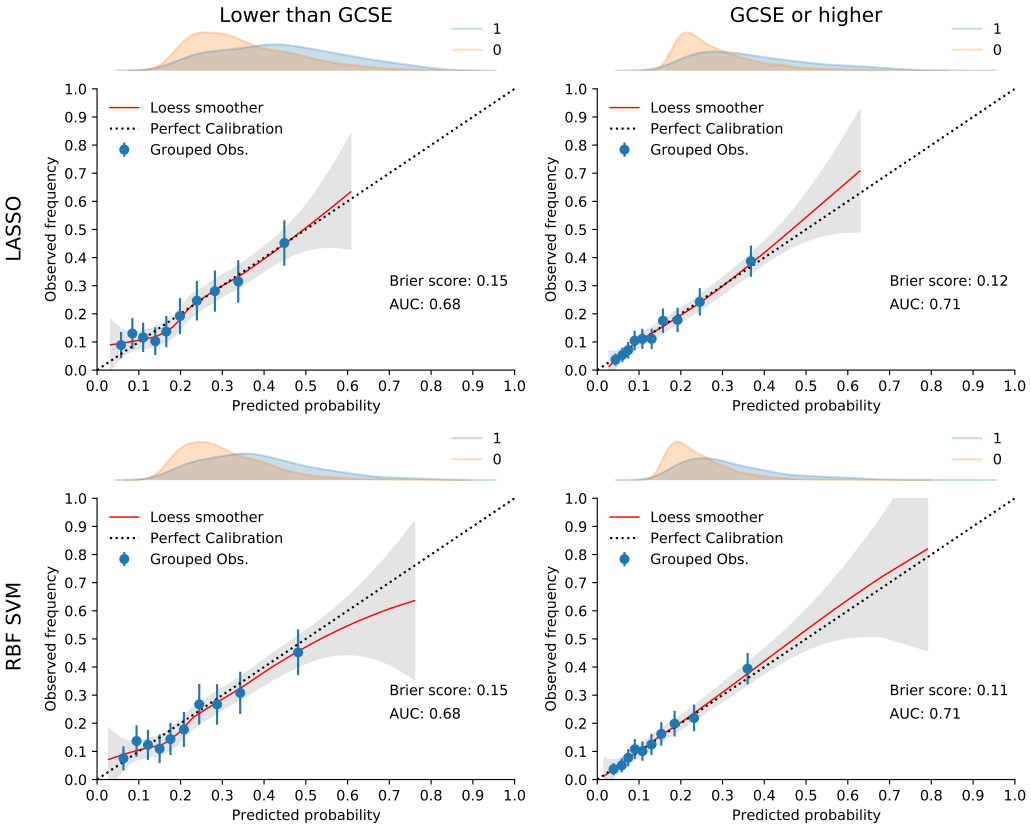


FIGURE C.16: Validation plots demonstrate calibration by educational attainment in the nested case-control sample.

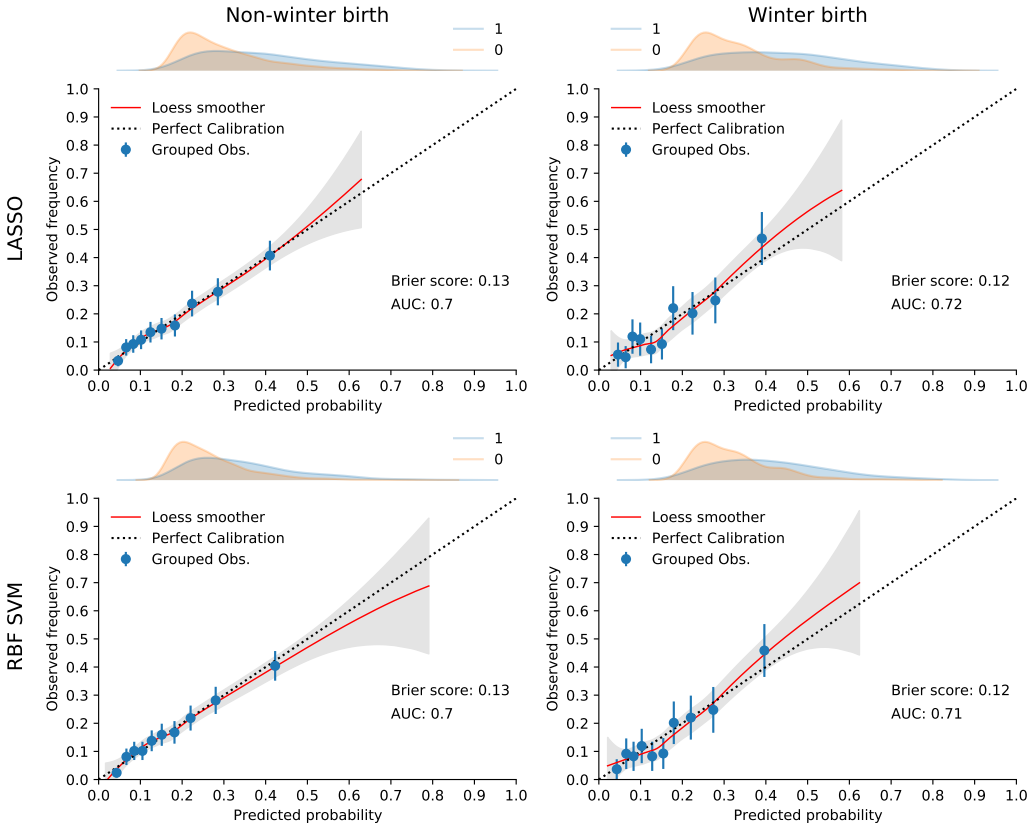


FIGURE C.17: Validation plots demonstrate calibration by season of birth in the nested case-control sample.

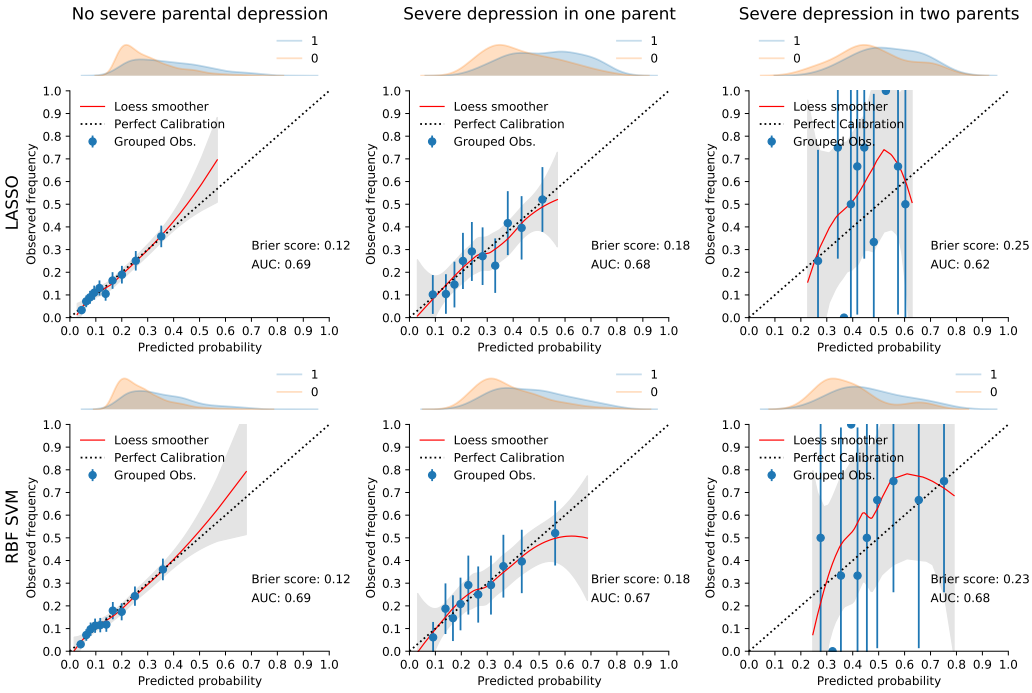


FIGURE C.18: Validation plots demonstrate calibration by severe parental depression in the nested case-control sample.

C.2.7 Importance scores

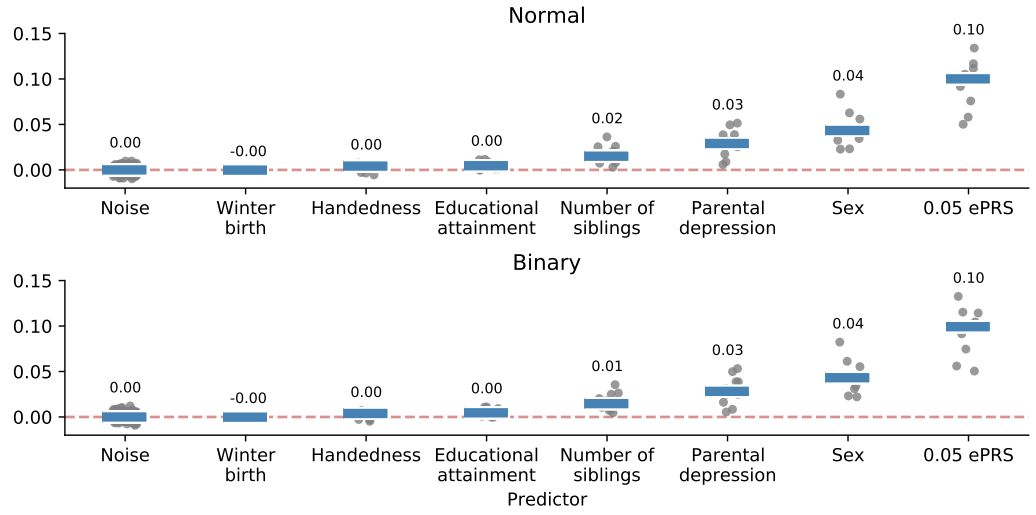


FIGURE C.19: Verification of importance score method. A logistic regression model was trained in 10-fold cross-validation using the combined set of demographic and 0.05 ePRS predictors. Either a binary random variable or a normally-distributed random variable with mean 0 and unit variance were added as noise to the dataset to assess whether permutation importance correctly identified it as the least important predictor. 100 repeats of each cross-validation were performed were both types of noise, each with a different random variable drawn. Noise variables were correctly identified as the least important predictors using the same permutation importance as for all model evaluations.

C.2.8 Generalisable association of predictions with cognitive, demographic, psychiatric and neurological outcomes

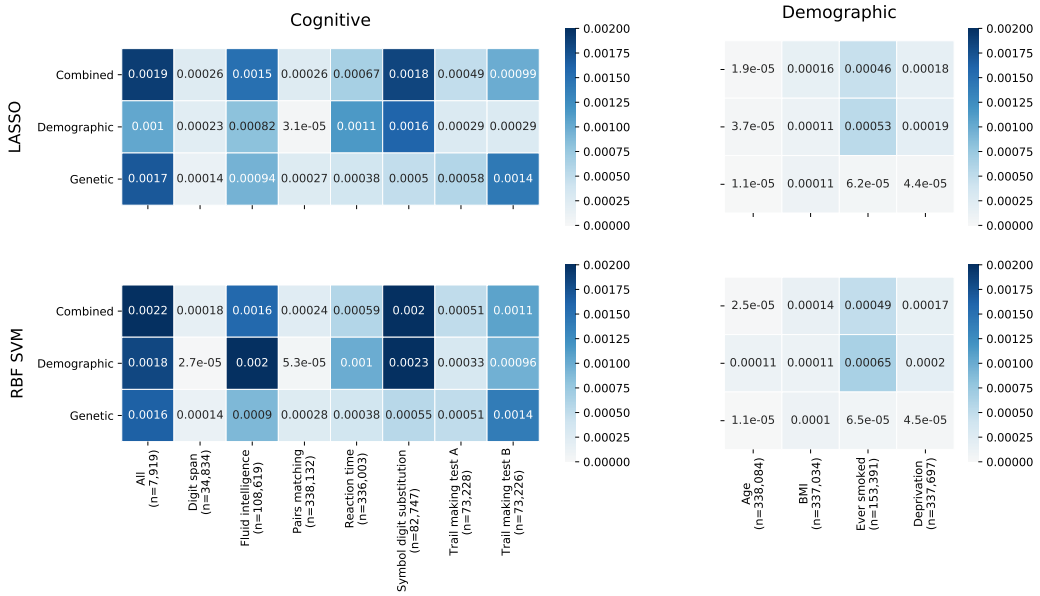


FIGURE C.20: Standard errors for prediction of risk scores using cognitive tests and widely-available demographic information in UK Biobank controls. LASSO and RBF SVM models are shown for genetic (0.05 ePRS), demographic or combined models.

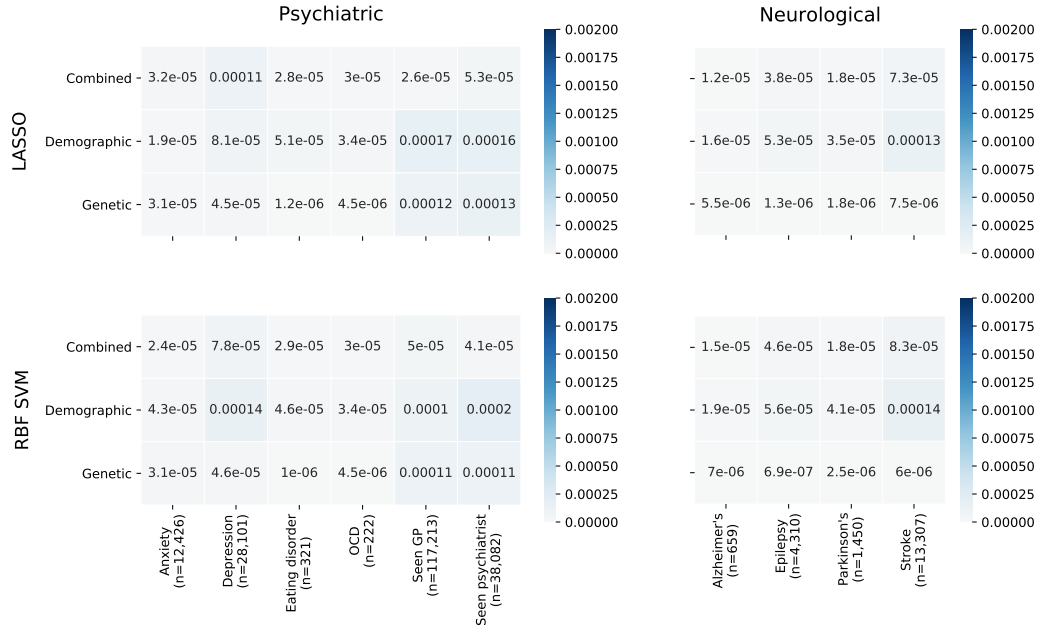


FIGURE C.21: Standard errors for prediction of risk scores using neurological and psychiatric outcomes in UK Biobank controls. LASSO and RBF SVM models are shown for genetic (0.05 ePRS), demographic or combined models.

C.2.9 Deconfounding

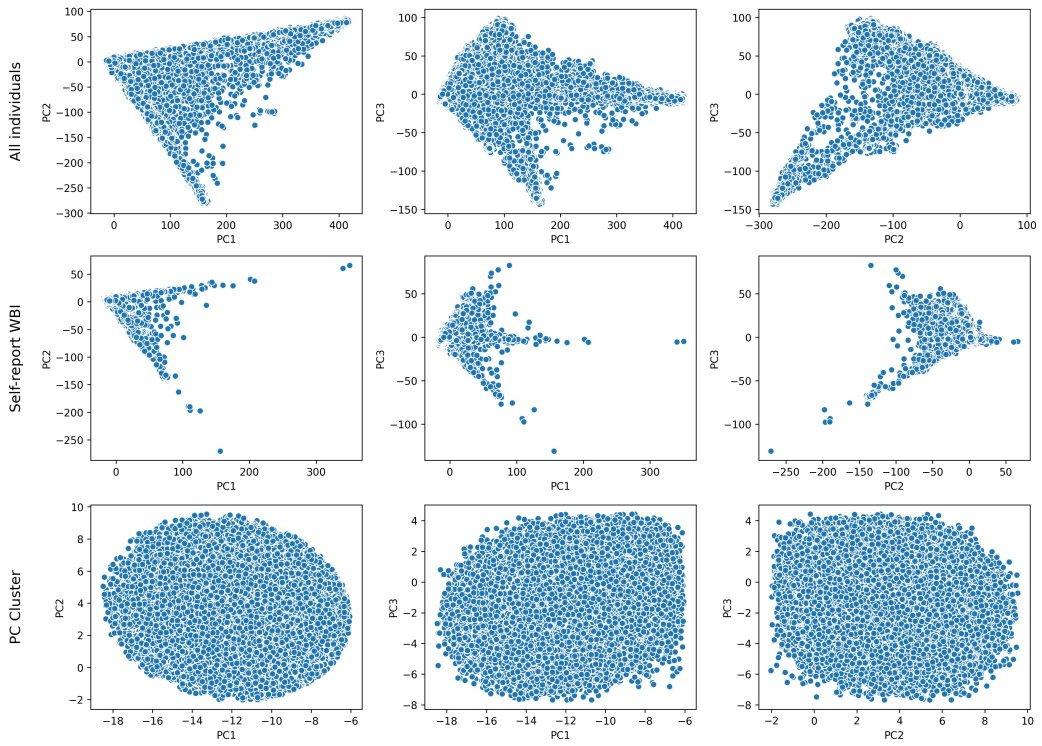


FIGURE C.22: Plots of the first 3 principal components. Plots are shown for the full UK Biobank cohort, those restricted to self-report white British or Irish, and those selected by UK Biobank as part of a more homogeneous principal components cluster.

Bibliography

Abel, Kathryn M, Richard Drake, and Jill M Goldstein (2010). "Sex differences in schizophrenia". In: *International review of psychiatry* 22.5, pp. 417–428.

Abu-Mostafa, Yaser S, Malik Magdon-Ismail, and Hsuan-Tien Lin (2012). *Learning from data*. Vol. 4. AMLBook New York, NY, USA:

Acikel, Cengizhan et al. (2016). "Evaluation of potential novel variations and their interactions related to bipolar disorders: analysis of genome-wide association study data". In: *Neuropsychiatric disease and treatment* 12, p. 2997.

Adams, Niall M and David J Hand (1999). "Comparing classifiers when the misallocation costs are uncertain". In: *Pattern Recognition* 32.7, pp. 1139–1147.

Aggarwal, Charu (2015). *Data classification : algorithms and applications*. Boca Raton: CRC Press. ISBN: 978-1466586741.

Aguiar-Pulido, Vanessa et al. (2010). "Machine Learning Techniques for Single Nucleotide Polymorphism—Disease Classification Models in Schizophrenia". In: *Molecules* 15.7, pp. 4875–4889.

Aguiar-Pulido, Vanessa et al. (2013). "Applied computational techniques on schizophrenia using genetic mutations". In: *Current topics in medicinal chemistry* 13.5, pp. 675–684.

Aleman, Andre, René S Kahn, and Jean-Paul Selten (2003). "Sex differences in the risk of schizophrenia: evidence from meta-analysis". In: *Archives of general psychiatry* 60.6, pp. 565–571.

Allardyce, Judith and Jane Boydell (2006). "Environment and schizophrenia: review: the wider social environment and schizophrenia". In: *Schizophrenia bulletin* 32.4, pp. 592–598.

Andreasen, Nancy C (1987). "Creativity and mental illness: prevalence rates in writers and their first-degree relatives." In: *The American Journal of Psychiatry*.

Andreasen, Nancy C et al. (2012). "Statistical epistasis and progressive brain change in schizophrenia: an approach for examining the relationships between multiple genes". In: *Molecular psychiatry* 17.11, pp. 1093–1102.

Andréasson, Sven et al. (1987). "Cannabis and schizophrenia a longitudinal study of Swedish conscripts". In: *The Lancet* 330.8574, pp. 1483–1486.

Anttila, Verneri et al. (2018). "Analysis of shared heritability in common disorders of the brain". In: *Science* 360.6395, eaap8757.

Arseneault, Louise et al. (2011). "Childhood trauma and children's emerging psychotic symptoms: a genetically sensitive longitudinal cohort study". In: *American Journal of Psychiatry* 168.1, pp. 65–72.

Association, American Psychiatric et al. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5)*. American Psychiatric Pub.

Austin, Peter C and Ewout W Steyerberg (2014). "Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers". In: *Statistics in medicine* 33.3, pp. 517–535.

Avramopoulos, Dimitrios (2018). "Recent advances in the genetics of schizophrenia". In: *Molecular neuropsychiatry* 4.1, pp. 35–51.

Ayano, Getinet, Getachew Tesfaw, and Shegaye Shumet (2019). "The prevalence of schizophrenia and other psychotic disorders among homeless people: a systematic review and meta-analysis". In: *BMC psychiatry* 19.1, pp. 1–14.

Bailey, Thomas et al. (2018). "Childhood trauma is associated with severity of hallucinations and delusions in psychotic disorders: a systematic review and meta-analysis". In: *Schizophrenia bulletin* 44.5, pp. 1111–1122.

Baron, Miron and Rhoda Gruen (1988). "Risk factors in schizophrenia: Season of birth and family history". In: *The British Journal of Psychiatry* 152.4, pp. 460–465.

Baselmans, Bart ML et al. (2020). "Risk in relatives, heritability, SNP-based heritability and genetic correlations in psychiatric disorders: a review". In: *Biological Psychiatry*.

Bateson, William (1909). *Mendel's principles of heredity*. Cambridge University Press.

Batty, G David et al. (2020). "Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis". In: *bmj* 368.

Bayat, Arash et al. (2020). "VariantSpark: Cloud-based machine learning for association study of complex phenotype and large-scale genomic data". In: *GigaScience* 9.8, giaa077.

Beach, Thomas G et al. (2008). "The Sun Health Research Institute Brain Donation Program: Description and Eexperience, 1987–2007". In: *Cell and tissue banking* 9.3, pp. 229–245.

Beam, Andrew L and Isaac S Kohane (2018). "Big data and machine learning in health care". In: *JAMA* 319.13, pp. 1317–1318.

Bebbington, Paul and Liz Kuipers (1994). "The predictive utility of expressed emotion in schizophrenia: an aggregate analysis". In: *Psychological medicine* 24.3, pp. 707–718.

Behravan, Hamid et al. (2018). "Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls". In: *Scientific reports* 8.1, pp. 1–13.

Belgard, T Grant et al. (2014). "Population structure confounds autism genetic classifier". In: *Molecular psychiatry* 19.4, p. 405.

Ben-Hur, Asa and Jason Weston (2010). "A user's guide to support vector machines". In: *Data mining techniques for the life sciences*. Springer, pp. 223–239.

Bergen, SE et al. (2012). "Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder". In: *Molecular psychiatry* 17.9, pp. 880–886.

Bergstra, James and Yoshua Bengio (2012). "Random search for hyper-parameter optimization". In: *The Journal of Machine Learning Research* 13.1, pp. 281–305.

Bernardo, Miguel et al. (2017). "Modelling gene-environment interaction in first episodes of psychosis". In: *Schizophrenia research* 189, pp. 181–189.

Biesheuvel, Cornelis J et al. (2008). "Advantages of the nested case-control design in diagnostic research". In: *BMC medical research methodology* 8.1, p. 48.

Bigdeli, Tim B et al. (2020). "Contributions of common genetic variants to risk of schizophrenia among individuals of African and Latino ancestry". In: *Molecular psychiatry* 25.10, pp. 2455–2467.

Bilogur, Aleksey (2018). "Missingno: a missing data visualization suite". In: *Journal of Open Source Software* 3.22, p. 547.

Boraska, Vesna et al. (2014). "A genome-wide association study of anorexia nervosa". In: *Molecular psychiatry* 19.10, pp. 1085–1094.

Boulesteix, Anne-Laure, Sabine Lauer, and Manuel JA Eugster (2013). "A plea for neutral comparison studies in computational sciences". In: *PloS one* 8.4, e61562.

Boulesteix, Anne-Laure et al. (2020). "Statistical learning approaches in the genetic epidemiology of complex diseases". In: *Human Genetics* 139.1, pp. 73–84.

Boyd, Jeffrey H, Ann E Pulver, and Walter Stewart (1986). "Season of Birth: Schizophernia and Bipolar Disorder". In: *Schizophrenia bulletin* 12.2, pp. 173–186.

Boydell, Jane et al. (2001). "Incidence of schizophrenia in ethnic minorities in London: ecological study into interactions with environment". In: *BMJ* 323.7325, p. 1336.

Bradbury, Thomas N and Gregory A Miller (1985). "Season of birth in schizophrenia: a review of evidence, methodology, and etiology." In: *Psychological bulletin* 98.3, p. 569.

Bradley, Andrew P (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7, pp. 1145–1159.

Breiman, Leo (1996). "Bagging predictors". In: *Machine learning* 24.2, pp. 123–140.

– (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Breiman, Leo et al. (1984). *Classification and regression trees*. Boca Raton: CRC press.

Breiman, Leo et al. (2001). "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical science* 16.3, pp. 199–231.

Bridges, Michael et al. (2011). "Genetic classification of populations using supervised learning". In: *PloS one* 6.5, e14802.

Brown, Alan S and Elena J Derkits (2010). "Prenatal infection and schizophrenia: a review of epidemiologic and translational studies". In: *American Journal of Psychiatry* 167.3, pp. 261–280.

Brown, Alan S et al. (1995). "Increased risk of affective disorders in males after second trimester prenatal exposure to the Dutch hunger winter of 1944–45". In: *The British Journal of Psychiatry* 166.5, pp. 601–606.

Brown, George W and James LT Birley (1968). "Crises and life changes and the onset of schizophrenia". In: *Journal of health and social behavior*, pp. 203–214.

Brown, George W, James LT Birley, and John K Wing (1972). "Influence of family life on the course of schizophrenic disorders: A replication". In: *The British Journal of Psychiatry* 121.562, pp. 241–258.

Burmeister, Margit, Melvin G McInnis, and Sebastian Zöllner (2008). "Psychiatric genetics: progress amid controversy". In: *Nature Reviews Genetics* 9.7, pp. 527–540.

Burton, Neel (2016). *Psychiatry*. Acheron Press. ISBN: 9780992912741.

Butzlaff, Ronald L and Jill M Hooley (1998). "Expressed emotion and psychiatric relapse: a meta-analysis". In: *Archives of general psychiatry* 55.6, pp. 547–552.

Bycroft, Clare et al. (2018). "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726, pp. 203–209.

Bzdok, Danilo, Naomi Altman, and Martin Krzywinski (2018). *Points of significance: statistics versus machine learning*.

Bzdok, Danilo, Denis Engemann, and Bertrand Thirion (2020). "Inference and Prediction Diverge in Biomedicine". In: *Patterns* 1.8, p. 100119.

Bzdok, Danilo, Gael Varoquaux, and Ewout W Steyerberg (2020). "Prediction, not association, paves the road to precision medicine". In: *JAMA psychiatry*.

Cannon, Mary, Peter B Jones, and Robin M Murray (2002). "Obstetric complications and schizophrenia: historical and meta-analytic review". In: *American Journal of Psychiatry* 159.7, pp. 1080–1092.

Cao, Hongbao et al. (2013). "Integrating fMRI and SNP data for biomarker identification for schizophrenia with a sparse representation based variable selection method". In: *BMC medical genomics* 6.S3, S2.

Cardno, Alastair G and Irving I Gottesman (2000). "Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics". In: *American journal of medical genetics* 97.1, pp. 12–17.

Cecile, A et al. (2011). "Strengthening the reporting of genetic risk prediction studies: the GRIPS statement". In: *European Journal of Epidemiology* 26.4, p. 255.

Chang, Chih-Chung and Chih-Jen Lin (2011). "LIBSVM: a library for support vector machines". In: *ACM transactions on intelligent systems and technology (TIST)* 2.3, pp. 1–27.

Charlson, Fiona J et al. (2018). "Global epidemiology and burden of schizophrenia: findings from the global burden of disease study 2016". In: *Schizophrenia bulletin* 44.6, pp. 1195–1203.

Chatelain, Clément et al. (2018). "Performance of epistasis detection methods in semi-simulated GWAS". In: *BMC bioinformatics* 19.1, pp. 1–17.

Chen, Jingchun et al. (2018). "Prediction of schizophrenia diagnosis by integration of genetically correlated conditions and traits". In: *Journal of Neuroimmune Pharmacology* 13.4, pp. 532–540.

Chen, Shyh-Huei et al. (2008). "A support vector machine approach for detecting gene-gene interaction". In: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 32.2, pp. 152–167.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chen, Xi and Hemant Ishwaran (2012). "Random forests for genomic data analysis". In: *Genomics* 99.6, pp. 323–329.

Cheniaux, Elie, J Landeira-Fernandez, and Marcio Versiani (2009). "The diagnoses of schizophrenia, schizoaffective disorder, bipolar disorder and unipolar depression: interrater reliability and congruence between DSM-IV and ICD-10". In: *Psychopathology* 42.5, pp. 293–298.

Chesney, Edward, Guy M Goodwin, and Seena Fazel (2014). "Risks of all-cause and suicide mortality in mental disorders: a meta-review". In: *World psychiatry* 13.2, pp. 153–160.

Christodoulou, Evangelia et al. (2019). "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models". In: *Journal of clinical epidemiology* 110, pp. 12–22.

Chumakov, Ilya et al. (2002). "Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia". In: *Proceedings of the National Academy of Sciences* 99.21, pp. 13675–13680.

Chyzhyk, Darya et al. (2018). "Controlling a confound in predictive models with a test set minimizing its effect". In: *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. IEEE, pp. 1–4.

Clarke, Mary Catherine, Michelle Harley, and Mary Cannon (2006). "The role of obstetric events in schizophrenia". In:

Collins, Gary S and Karel GM Moons (2019). "Reporting of artificial intelligence prediction models". In: *The Lancet* 393.10181, pp. 1577–1579.

Collins, Gary S et al. (2015). "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement". In: *BMC medicine* 13.1, p. 1.

Collins, Rory (2012). "What makes UK Biobank special?" In: *The Lancet* 9822.379, pp. 1173–1174.

Consortium, Wellcome Trust Case Control et al. (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". In: *Nature* 447.7145, p. 661.

Cooper, John Edward et al. (1972). "Psychiatric diagnosis in New York and London: A comparative study of mental hospital admissions." In:

Cordell, Heather J (2002). "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans". In: *Human molecular genetics* 11.20, pp. 2463–2468.

– (2009). "Detecting gene–gene interactions that underlie human diseases". In: *Nature Reviews Genetics* 10.6, p. 392.

Corey, LA et al. (1979). "Effects of type of placentation on birthweight and its variability in monozygotic and dizygotic twins". In: *Acta geneticae medicae et gemellologiae: twin research* 28.1, pp. 41–50.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

Cox, David R (1958). "Two further applications of a model for binary regression". In: *Biometrika* 45.3/4, pp. 562–565.

Craddock, Nick and Michael J Owen (2010). "The Kraepelinian dichotomy–going, going... but still not gone". In: *The British Journal of Psychiatry* 196.2, pp. 92–95.

Cribari-Neto, Francisco and Achim Zeileis (2009). "Beta regression in R". In:

Cutler, Adele, D Richard Cutler, and John R Stevens (2012). "Random forests". In: *Ensemble machine learning*. Springer, pp. 157–175.

Daneshjou, Roxana et al. (2017). "Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges". In: *Human mutation* 38.9, pp. 1182–1192.

Davies, Geoffrey et al. (2003). "A systematic review and meta-analysis of Northern Hemisphere season of birth studies in schizophrenia". In: *Schizophrenia bulletin* 29.3, pp. 587–593.

Day, Richard et al. (1987). "Stressful life events preceding the acute onset of schizophrenia: a cross-national study from the World Health Organization". In: *Culture, Medicine and Psychiatry* 11.2, pp. 123–205.

De Leon, Jose and Francisco J Diaz (2005). "A meta-analysis of worldwide studies demonstrates an association between schizophrenia and tobacco smoking behaviors". In: *Schizophrenia research* 76.2-3, pp. 135–157.

Deary, Ian J et al. (2005). "The cognitive cost of being a twin: two whole-population surveys". In: *Twin Research and Human Genetics* 8.4, pp. 376–383.

Debray, Thomas PA et al. (2019). "A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes". In: *Statistical methods in medical research* 28.9, pp. 2768–2786.

DeLong, Elizabeth R, David M DeLong, and Daniel L Clarke-Pearson (1988). "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach". In: *Biometrics*, pp. 837–845.

Demšar, Janez (2006). "Statistical comparisons of classifiers over multiple data sets". In: *Journal of Machine learning research* 7.Jan, pp. 1–30.

Deng, Jia et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Dick, Danielle M et al. (2003). "Genomewide linkage analyses of bipolar disorder: a new sample of 250 pedigrees from the National Institute of Mental Health Genetics Initiative". In: *The American Journal of Human Genetics* 73.1, pp. 107–114.

Dickson, Hannah et al. (2012). "Meta-analyses of cognitive and motor function in youth aged 16 years and younger who subsequently develop schizophrenia". In: *Psychological medicine* 42.4, p. 743.

Dietterich, Thomas G (1998). "Approximate statistical tests for comparing supervised classification learning algorithms". In: *Neural computation* 10.7, pp. 1895–1923.

Dietterich, Thomas G and Ghulum Bakiri (1994). "Solving multiclass learning problems via error-correcting output codes". In: *Journal of artificial intelligence research* 2, pp. 263–286.

Dinga, Richard et al. (2020). "Controlling for effects of confounding variables on machine learning predictions". In: *BioRxiv*.

Doan, Nhat Trung et al. (2017). "Distinct multivariate brain morphological patterns and their added predictive value with cognitive and polygenic risk scores in mental disorders". In: *NeuroImage: Clinical* 15, pp. 719–731.

Domingos, Pedro (1998). "Occam's two razors: the sharp and the blunt". In: *KDD*, pp. 37–43.

Domingos, Pedro M (2012). "A few useful things to know about machine learning." In: *Commun. acm* 55.10, pp. 78–87.

Domínguez, Eduardo et al. (2007). "Extensive linkage disequilibrium mapping at HTR2A and DRD3 for schizophrenia susceptibility genes in the Galician population". In: *Schizophrenia research* 90.1-3, pp. 123–129.

Dragovic, Milan and Geoff Hammond (2005). "Handedness in schizophrenia: a quantitative review of evidence". In: *Acta Psychiatrica Scandinavica* 111.6, pp. 410–419.

Drummond, Chris and Robert C Holte (2006). "Cost curves: An improved method for visualizing classifier performance". In: *Machine learning* 65.1, pp. 95–130.

Duchi, John, Elad Hazan, and Yoram Singer (2011). "Adaptive subgradient methods for online learning and stochastic optimization." In: *Journal of machine learning research* 12.7.

Dudbridge, Frank (2013). "Power and predictive accuracy of polygenic risk scores". In: *PLoS Genet* 9.3, e1003348.

Dudbridge, Frank and Arief Gusnanto (2008). "Estimation of significance thresholds for genomewide association scans". In: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 32.3, pp. 227–234.

Dumancas, Gerard G and Ghalib A Bello (2015). "Comparison of machine-learning techniques for handling multicollinearity in big data analytics and high-performance data mining". In:

Durstewitz, Daniel, Georgia Koppe, and Andreas Meyer-Lindenberg (2019). "Deep neural networks in psychiatry". In: *Molecular psychiatry* 24.11, pp. 1583–1598.

Eichler, Evan E et al. (2010). "Missing heritability and strategies for finding the underlying causes of complex disease". In: *Nature Reviews Genetics* 11.6, p. 446.

Elkan, Charles (2001). "The foundations of cost-sensitive learning". In: *International joint conference on artificial intelligence*. Vol. 17. 1. Lawrence Erlbaum Associates Ltd, pp. 973–978.

Elvevag, Brita and Terry E Goldberg (2000). "Cognitive impairment in schizophrenia is the core of the disorder". In: *Critical Reviews in Neurobiology* 14.1.

Endicott, J and RL Spitzer (1987). "Schedule for affective disorders and schizophrenia (SADS)". In: *Acta Psychiatrica Belgica* 87.4, pp. 361–516.

Engchuan, Worrawat et al. (2015). "Performance of case-control rare copy number variation annotation in classification of autism". In: *BMC medical genomics* 8.1, S7.

Eriksen, Willy, Jon M Sundet, and Kristian Tambs (2012). "Twin–Singleton Differences in Intelligence: A Register-Based Birth Cohort Study of Norwegian Males". In: *Twin Research and Human Genetics* 15.5, pp. 649–655.

Escott-Price, Valentina et al. (2020). "Genetic liability to schizophrenia is negatively associated with educational attainment in UK Biobank". In: *Molecular Psychiatry* 25.4, pp. 703–705.

Falconer, Douglas S (1965). "The inheritance of liability to certain diseases, estimated from the incidence among relatives". In: *Annals of human genetics* 29.1, pp. 51–76.

Falconer, DS and TFC Mackay (1996). *Introduction to quantitative genetics*. Pearson.

Fan, Rong-En et al. (2008). "LIBLINEAR: A library for large linear classification". In: *the Journal of machine Learning research* 9, pp. 1871–1874.

Fang, Yao-Hwei and Yen-Feng Chiu (2012). "SVM-Based Generalized Multifactor Dimensionality Reduction Approaches for Detecting Gene-Gene Interactions in Family Studies". In: *Genetic epidemiology* 36.2, pp. 88–98.

Fawcett, Tom (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8, pp. 861–874.

Finucane, Hilary K et al. (2018). "Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types". In: *Nature genetics* 50.4, pp. 621–629.

First, Michael B (2014). "Structured clinical interview for the DSM (SCID)". In: *The encyclopedia of clinical psychology*, pp. 1–6.

Fisher, Ronald A (1918). "The correlation between relatives on the supposition of Mendelian inheritance." In: *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 52.2, pp. 399–433.

Flagel, Lex, Yaniv Brandvain, and Daniel R Schrider (2019). "The unreasonable effectiveness of convolutional neural networks in population genetic inference". In: *Molecular biology and evolution* 36.2, pp. 220–238.

Flint, Jonathan and Marcus Munafo (2014). "Genesis of a complex disease". In: *Nature* 511.7510, pp. 412–413.

Formann-Roe, Scott (2012). "Understanding the Bias-Variance Tradeoff". In: *http://scott.fortmann-roe.com/docs/BiasVariance.html*. (Visited on 01/15/2021).

Frankel, Wayne N and Nicholas J Schork (1996). "Who's afraid of epistasis?" In: *Nature genetics* 14.4, p. 371.

Freund, Yoav (1995). "Boosting a weak learning algorithm by majority". In: *Information and computation* 121.2, pp. 256–285.

Freund, Yoav and Robert E Schapire (1997). "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1, pp. 119–139.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani, et al. (2000). "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)". In: *The annals of statistics* 28.2, pp. 337–407.

Friedman, Jerome H (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*, pp. 1189–1232.

– (2002). "Stochastic gradient boosting". In: *Computational statistics & data analysis* 38.4, pp. 367–378.

Frith, Chris and Eve C Johnstone (2003). *Schizophrenia: A very short introduction*. OUP Oxford.

Fromer, Menachem et al. (2016). "Gene expression elucidates functional impact of polygenic risk for schizophrenia". In: *Nature neuroscience* 19.11, pp. 1442–1453.

Fry, Anna et al. (2017). "Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population". In: *American journal of epidemiology* 186.9, pp. 1026–1034.

Fung, Si Ming et al. (2019). "Performance of single-nucleotide polymorphisms in breast cancer risk prediction models: a systematic review and meta-analysis". In: *Cancer Epidemiology and Prevention Biomarkers* 28.3, pp. 506–521.

Gage, Suzanne H et al. (2017). "Assessing causality in associations between cannabis use and schizophrenia risk: a two-sample Mendelian randomization study". In: *Psychological medicine* 47.5, pp. 971–980.

Geddes, John R and Stephen M Lawrie (1995). "Obstetric complications and schizophrenia: a meta-analysis". In: *The British Journal of Psychiatry* 167.6, pp. 786–793.

Gejman, Pablo V, Alan R Sanders, and Kenneth S Kendler (2011). "Genetics of schizophrenia: new findings and challenges". In: *Annual review of genomics and human genetics* 12, pp. 121–144.

Genovese, Giulio et al. (2016). "Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia". In: *Nature neuroscience* 19.11, pp. 1433–1441.

Genuer, Robin et al. (2017). "Random forests for big data". In: *Big Data Research* 9, pp. 28–46.

Géron, Aurélien (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Ghafouri-Fard, Soudeh et al. (2019). "Application of Single-Nucleotide Polymorphisms in the Diagnosis of Autism Spectrum Disorders: A Preliminary Study with Artificial Neural Networks". In: *Journal of Molecular Neuroscience*, pp. 1–7.

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio (2011). "Deep sparse rectifier neural networks". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323.

Goodfellow, Ian et al. (2016). *Deep learning*. Vol. 1. MIT press Cambridge.

Goodwin, Guy M and John R Geddes (2007). "What is the heartland of psychiatry?" In: *The British Journal of Psychiatry* 191.3, pp. 189–191.

Gottesman, Irving I (1991). *Schizophrenia genesis: The origins of madness*. WH Freeman and Company, New York.

Gottesman, Irving I and James Shields (1972). "A polygenic theory of schizophrenia". In: *International Journal of Mental Health* 1.1-2, pp. 107–115.

Green, Elaine K et al. (2005). "Operation of the schizophrenia susceptibility gene, neuregulin 1, across traditional diagnostic boundaries to increase risk for bipolar disorder". In: *Archives of general psychiatry* 62.6, pp. 642–648.

Green, Elaine K et al. (2006). "Genetic variation of brain-derived neurotrophic factor (BDNF) in bipolar disorder: case-control study of over 3000 individuals from the UK". In: *The British Journal of Psychiatry* 188.1, pp. 21–25.

Green, Michael F et al. (1992). "Wisconsin Card Sorting Test performance in schizophrenia: Remediation of a stubborn deficit." In: *The American journal of psychiatry* 149.1, pp. 62–67.

Grotzinger, Andrew D (2021). "Shared genetic architecture across psychiatric disorders". In: *Psychological Medicine*, pp. 1–7.

Guan, L et al. (2016). "Common variants on 17q25 and gene–gene interactions conferring risk of schizophrenia in Han Chinese population and regulating gene expressions in human brain". In: *Molecular psychiatry* 21.9, pp. 1244–1250.

Günther, Frauke, Nina Wawro, and Karin Bammann (2009). "Neural networks for modeling gene-gene interactions in association studies". In: *BMC genetics* 10.1, pp. 1–14.

Guo, Yiran et al. (2015). "Machine learning derived risk prediction of anorexia nervosa". In: *BMC medical genomics* 9.1, p. 4.

Guyon, Isabelle et al. (2002). "Gene selection for cancer classification using support vector machines". In: *Machine learning* 46.1, pp. 389–422.

Hafner, H et al. (1994). "The epidemiology of early schizophrenia". In: *Br J Psychiatry* 164.23, pp. 29–38.

Halevy, Alon, Peter Norvig, and Fernando Pereira (2009). "The unreasonable effectiveness of data". In: *IEEE Intelligent Systems* 24.2, pp. 8–12.

Hamshere, Marian L et al. (2013). "Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC". In: *Molecular psychiatry* 18.6, pp. 708–712.

Hanley, James A and Barbara J McNeil (1982). "The meaning and use of the area under a receiver operating characteristic (ROC) curve." In: *Radiology* 143.1, pp. 29–36.

Harrell Jr, Frank E (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Harrison, Glynn (1990). "Searching for the causes of schizophrenia: the role of migrant studies". In: *Schizophrenia bulletin* 16.4, pp. 663–672.

Harrison, Glynn et al. (1988). "A prospective study of severe mental disorder in Afro-Caribbean patients". In: *Psychological medicine* 18.3, pp. 643–657.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

He, Kaiming et al. (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034.

Heaton, Jeff (2008). *Introduction to neural networks with Java*. Heaton Research, Inc.

Heinrichs, R Walter and Konstantine K Zakzanis (1998). "Neurocognitive deficit in schizophrenia: a quantitative review of the evidence." In: *Neuropsychology* 12.3, p. 426.

Hellman, SG et al. (1998). "Monetary reinforcement and Wisconsin Card Sorting performance in schizophrenia: why show me the money?" In: *Schizophrenia Research* 34.1-2, pp. 67–75.

Henssler, Jonathan et al. (2019). "Migration and schizophrenia: meta-analysis and explanatory framework". In: *European archives of psychiatry and clinical neuroscience*, pp. 1–11.

Herrera, Victor M et al. (2019). "Random forest implementation and optimization for Big Data analytics on LexisNexis's high performance computing cluster platform". In: *Journal of Big Data* 6.1, pp. 1–36.

Heston, Leonard L (1966). "Psychiatric disorders in foster home reared children of schizophrenic mothers". In: *British journal of Psychiatry* 112.489, pp. 819–825.

Higgins, Julian PT et al. (2019). *Cochrane handbook for systematic reviews of interventions version 6.0 (updated July 2019)*. www.training.cochrane.org/handbook. Accessed: 10/07/2020.

Hinton, Geoffrey et al. (2012). "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". In: *IEEE Signal processing magazine* 29.6, pp. 82–97.

Hirnstein, Marco and Kenneth Hugdahl (2014). "Excess of non-right-handedness in schizophrenia: meta-analysis of gender effects and potential biases in handedness assessment". In: *The British Journal of Psychiatry* 205.4, pp. 260–267.

Hirsch, Steven R and Julian P Leff (1975). *Abnormalities in parents of schizophrenics*. Oxford University Press Oxford, London.

Ho, Daniel Sik Wai et al. (2019). "Machine learning SNP based prediction for precision medicine". In: *Frontiers in genetics* 10, p. 267.

Hoek, Hans W, Alan S Brown, and Ezra Susser (1998). "The Dutch famine and schizophrenia spectrum disorders". In: *Social psychiatry and psychiatric epidemiology* 33.8, pp. 373–379.

Hoerl, Arthur E and Robert W Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1, pp. 55–67.

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks". In: *Neural networks* 4.2, pp. 251–257.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis (2006). "Unbiased recursive partitioning: A conditional inference framework". In: *Journal of Computational and Graphical statistics* 15.3, pp. 651–674.

Howes, Oliver D and Shitij Kapur (2009). "The dopamine hypothesis of schizophrenia: version III—the final common pathway". In: *Schizophrenia bulletin* 35.3, pp. 549–562.

Howes, Oliver D et al. (2017). "The role of genes, stress, and dopamine in the development of schizophrenia". In: *Biological psychiatry* 81.1, pp. 9–20.

Hsu, Chih-Wei, Chih-Chung Chang, Chih-Jen Lin, et al. (2003). *A practical guide to support vector classification*. URL: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (visited on 09/20/2020).

Huang, Yingxiang et al. (2020). "A tutorial on calibration measurements and calibration models for clinical prediction models". In: *Journal of the American Medical Informatics Association* 27.4, pp. 621–633.

Hunter, John D (2007). "Matplotlib: A 2D graphics environment". In: *Computing in science & engineering* 9.3, pp. 90–95.

Iniesta, R, D Stahl, and P McGuffin (2016). "Machine learning, statistical learning and the future of biological research in psychiatry". In: *Psychological medicine* 46.12, pp. 2455–2465.

Inouye, Michael et al. (2018). "Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention". In: *Journal of the American College of Cardiology* 72.16, pp. 1883–1893.

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR, pp. 448–456.

Islam, Md Mohaimenul et al. (2020). "Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis". In: *Computer Methods and Programs in Biomedicine* 191, p. 105320.

Jääskeläinen, Erika et al. (2013). "A systematic review and meta-analysis of recovery in schizophrenia". In: *Schizophrenia bulletin* 39.6, pp. 1296–1306.

James, Gareth et al. (2013). *An introduction to statistical learning*. Springer, New York.

Jamison, Kay Redfield (1996). *Touched with Fire: Manic-Depressive Illness and the Artistic Temperament*. Simon and Schuster.

Janssen, Kristel JM et al. (2010). "Missing covariate data in medical research: to impute is better than to ignore". In: *Journal of clinical epidemiology* 63.7, pp. 721–727.

Janssens, A Cecile JW and Forike K Martens (2020). "Reflection on modern methods: revisiting the area under the ROC curve". In: *International journal of epidemiology* 49.4, pp. 1397–1403.

Janssens, A Cecile JW et al. (2011). "Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration". In: *European journal of epidemiology* 26.4, p. 313.

Jiang, Rui et al. (2009). "A random forest approach to the detection of epistatic interactions in case-control studies". In: *BMC bioinformatics* 10.1, pp. 1–12.

Johnstone, Eve C et al. (1981). "Institutionalization and the defects of schizophrenia". In: *The British Journal of Psychiatry* 139.3, pp. 195–203.

Jones, Hannah J et al. (2020). "Examining pathways between genetic liability for schizophrenia and patterns of tobacco and cannabis use in adolescence". In: *Psychological medicine*, pp. 1–8.

Jones, Peter B (2013). "Adult mental health disorders and their age at onset". In: *The British Journal of Psychiatry* 202.s54, s5–s10.

Jones, Peter B et al. (1998). "Schizophrenia as a long-term outcome of pregnancy, delivery, and perinatal complications: a 28-year follow-up of the 1966 north Finland general population birth cohort". In: *American Journal of Psychiatry* 155.3, pp. 355–364.

Kahn, René S and Richard SE Keefe (2013). "Schizophrenia is a cognitive illness: time for a change in focus". In: *JAMA psychiatry* 70.10, pp. 1107–1112.

Kallner, Anders (2018). "Bayes' theorem, the ROC diagram and reference values: Definition and use in clinical diagnosis". In: *Biochemia medica* 28.1, pp. 16–25.

Kapur, Shitij, Anthony G Phillips, and Thomas R Insel (2012). "Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it?" In: *Molecular psychiatry* 17.12, p. 1174.

Karayiorgou, Maria et al. (1995). "Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11". In: *Proceedings of the National Academy of Sciences* 92.17, pp. 7612–7616.

Kassem, Layla et al. (2006). "Familiality of polarity at illness onset in bipolar affective disorder". In: *American Journal of Psychiatry* 163.10, pp. 1754–1759.

Keefe, Richard SE and Wayne S Fenton (2007). "How should DSM-V criteria for schizophrenia include cognitive impairment?" In: *Schizophrenia bulletin* 33.4, pp. 912–920.

Kendall, Kimberley M et al. (2017). "Cognitive performance among carriers of pathogenic copy number variants: analysis of 152,000 UK Biobank subjects". In: *Biological psychiatry* 82.2, pp. 103–110.

Kendler, Kenneth S et al. (1996). "The treated incidence of psychotic and affective illness in twins compared with population expectation: a study in the Swedish Twin and Psychiatric Registries". In: *Psychological Medicine* 26.6, pp. 1135–1144.

Kety, Seymour S (1987). "The significance of genetic factors in the etiology of schizophrenia: results from the national study of adoptees in Denmark". In: *Journal of psychiatric research* 21.4, pp. 423–429.

Kety, Seymour S et al. (1994). "Mental illness in the biological and adoptive relatives of schizophrenic adoptees: replication of the Copenhagen study in the rest of Denmark". In: *Archives of general psychiatry* 51.6, pp. 442–455.

Khandaker, Golam M et al. (2011). "A quantitative meta-analysis of population-based studies of premorbid intelligence and schizophrenia". In: *Schizophrenia research* 132.2-3, pp. 220–227.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv*.

Kirov, George (2015). "CNVs in neuropsychiatric disorders". In: *Human molecular genetics* 24.R1, R45–R49.

Kirov, George et al. (2014). "The penetrance of copy number variations for schizophrenia and developmental delay". In: *Biological psychiatry* 75.5, pp. 378–385.

Knickmeyer, Rebecca C et al. (2011). "Twin-singleton differences in neonatal brain structure". In: *Twin Research and Human Genetics* 14.3, pp. 268–276.

Kohoutová, Lada et al. (2020). "Toward a unified framework for interpreting machine-learning models in neuroimaging". In: *Nature Protocols* 15.4, pp. 1399–1435.

Kokhlikyan, Narine et al. (2020). "Captum: A unified and generic model interpretability library for pytorch". In: *arXiv*.

Kong, Eun Bae and Thomas G Dietterich (1995). "Error-correcting output coding corrects bias and variance". In: *Machine learning proceedings 1995*. Elsevier, pp. 313–321.

Koo, Ching Lee et al. (2013). "A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology". In: *BioMed research international* 2013.

Kosten, Therese A and Bruce J Rounsaville (1992). "Sensitivity of psychiatric diagnosis based on the best estimate procedure." In: *The American journal of psychiatry* 149.9, 1225—1227.

Kraepelin, Emil (1919). *Dementia praecox and paraphrenia*. Livingstone.

Kringlen, Einar and Gunnar Cramer (1989). "Offspring of monozygotic twins discordant for schizophrenia". In: *Archives of general psychiatry* 46.10, pp. 873–877.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.

Kruppa, Jochen, Andreas Ziegler, and Inke R König (2012). "Risk estimation and risk prediction using machine-learning methods". In: *Human genetics* 131.10, pp. 1639–1654.

Krystal, John H et al. (2017). "Computational psychiatry and the challenge of schizophrenia". In: *Schizophrenia Bulletin* 43.3, pp. 473–475.

Laksshman, Sundaram et al. (2017). "DeepBipolar: Identifying genomic mutations for bipolar disorder via deep learning". In: *Human mutation* 38.9, pp. 1217–1224.

Lam, Max et al. (2019). "Comparative genetic architectures of schizophrenia in East Asian and European populations". In: *Nature genetics* 51.12, pp. 1670–1678.

Lan, TH et al. (2008). "Performance of a neuro-fuzzy model in predicting weight changes of chronic schizophrenic patients exposed to antipsychotics". In: *Molecular psychiatry* 13.12, pp. 1129–1137.

Le, Trang T, Weixuan Fu, and Jason H Moore (2020). "Scaling tree-based automated machine learning to biomedical big data with a feature set selector". In: *Bioinformatics* 36.1, pp. 250–256.

Le Hellard, Stéphanie et al. (2017). "Identification of gene loci that overlap between schizophrenia and educational attainment". In: *Schizophrenia bulletin* 43.3, pp. 654–664.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, pp. 436–444.

Lee, S Hong et al. (2012a). "Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs". In: *Nature genetics* 44.3, pp. 247–250.

Lee, S Hong et al. (2013). "Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs". In: *Nature genetics* 45.9, p. 984.

Lee, Sang Hong et al. (2012b). "A better coefficient of determination for genetic profile analysis". In: *Genetic epidemiology* 36.3, pp. 214–224.

Lee, Yena et al. (2018). "Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review". In: *Journal of affective disorders* 241, pp. 519–532.

Legge, Sophie E et al. (2021). "Genetic architecture of schizophrenia: a review of major advancements". In: *Psychological Medicine*, pp. 1–10.

Lencz, T et al. (2007). "Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia". In: *Molecular psychiatry* 12.6, pp. 572–580.

Leonenko, Ganna et al. (2017). "Mutation intolerant genes and targets of FMRP are enriched for nonsynonymous alleles in schizophrenia". In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 174.7, pp. 724–731.

Lewis, Cathryn M and Evangelos Vassos (2020). "Polygenic risk scores: from research tools to clinical instruments". In: *Genome Medicine* 12, pp. 1–11.

Lewis, Cathryn M et al. (2003). "Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia". In: *The American Journal of Human Genetics* 73.1, pp. 34–48.

Li, Cong et al. (2014). "Improving genetic risk prediction by leveraging pleiotropy". In: *Human genetics* 133.5, pp. 639–650.

Li, Gang et al. (2020). "Application of deep canonically correlated sparse autoencoder for the classification of schizophrenia". In: *Computer methods and programs in biomedicine* 183, p. 105073.

Li, Hao et al. (2017a). "Visualizing the loss landscape of neural nets". In: *arXiv*.

Li, Jing et al. (2016). "Detecting gene-gene interactions using a permutation-based random forest method". In: *BioData mining* 9.1, pp. 1–17.

Li, Lisha et al. (2017b). "Hyperband: A novel bandit-based approach to hyperparameter optimization". In: *The Journal of Machine Learning Research* 18.1, pp. 6765–6816.

Li, Wentian and Jens Reich (2000). "A complete enumeration and classification of two-locus disease models". In: *Human heredity* 50.6, pp. 334–349.

Librenza-Garcia, Diego et al. (2017). "The impact of machine learning techniques in the study of bipolar disorder: a systematic review". In: *Neuroscience & Biobehavioral Reviews* 80, pp. 538–554.

Lichtenstein, Paul et al. (2009). "Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study". In: *The Lancet* 373.9659, pp. 234–239.

Lin, Chao-Cheng et al. (2008). "Artificial neural network prediction of clozapine response with combined pharmacogenetic and clinical data". In: *Computer methods and programs in biomedicine* 91.2, pp. 91–99.

Liu, Chunyu, H Hoxie Ackerman, and John P Carulli (2011). "A genome-wide screen of gene–gene interactions for rheumatoid arthritis susceptibility". In: *Human genetics* 129.5, pp. 473–485.

Ludwig, Arnold M (1992). "Creative achievement and psychopathology: Comparison among professions". In: *American journal of psychotherapy* 46.3, pp. 330–354.

Lunetta, Kathryn L et al. (2004). "Screening large-scale association study data: exploiting interactions using random forests". In: *BMC genetics* 5.1, pp. 1–13.

Luo, Wei et al. (2016). "Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view". In: *Journal of medical Internet research* 18.12, e323.

Lupski, James R (1998). "Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits". In: *Trends in genetics* 14.10, pp. 417–422.

MacCabe, James H et al. (2010). "Excellent school performance at age 16 and risk of adult bipolar disorder: national cohort study". In: *The British Journal of Psychiatry* 196.2, pp. 109–115.

MacCabe, JH et al. (2008). "Scholastic achievement at age 16 and risk of schizophrenia and other psychoses: a national cohort study". In: *Psychological medicine* 38.8, pp. 1133–1140.

Mahon, Pamela Belmonte et al. (2009). "Genome-wide linkage and follow-up association study of postpartum mood symptoms". In: *American Journal of Psychiatry* 166.11, pp. 1229–1237.

Mailman, Matthew D et al. (2007). "The NCBI dbGaP database of genotypes and phenotypes". In: *Nature genetics* 39.10, pp. 1181–1186.

Mäki, Pirjo et al. (2010). "Schizophrenia in the offspring of antenatally depressed mothers in the northern Finland 1966 birth cohort: relationship to family history of psychosis". In: *American Journal of Psychiatry* 167.1, pp. 70–77.

Manchia, Mirko et al. (2020). "Challenges and future prospects of precision medicine in psychiatry". In: *Pharmacogenomics and Personalized Medicine* 13, p. 127.

Mangalore, Roshni and Martin Knapp (2007). "Cost of schizophrenia in England." In: *The journal of mental health policy and economics* 10.1, pp. 23–41.

Manolio, Teri A and Rory Collins (2010). "Enhancing the feasibility of large cohort studies". In: *Jama* 304.20, pp. 2290–2291.

Manolio, Teri A et al. (2007). "New models of collaboration in genome-wide association studies: the Genetic Association Information Network". In: *Nature genetics* 39.9, p. 1045.

Marchini, Jonathan, Peter Donnelly, and Lon R Cardon (2005). "Genome-wide strategies for detecting multiple loci that influence complex diseases". In: *Nature genetics* 37.4, p. 413.

Marchini, Jonathan et al. (2004). "The effects of human population structure on large genetic association studies". In: *Nature genetics* 36.5, p. 512.

Marshall, Christian R et al. (2017). "Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects". In: *Nature genetics* 49.1, pp. 27–35.

Martin, Alicia R et al. (2017). "Human demographic history impacts genetic risk prediction across diverse populations". In: *The American Journal of Human Genetics* 100.4, pp. 635–649.

Marwaha, Steven and Sonia Johnson (2004). "Schizophrenia and employment". In: *Social psychiatry and psychiatric epidemiology* 39.5, pp. 337–349.

Masters, Timothy (1993). *Practical neural network recipes in C++*. Morgan Kaufmann.

Matchenko-Shimko, N and Marie-Pierre Dube (2007). "Gene-gene interaction tests using SVM and neural network modeling". In: *2007 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*. IEEE, pp. 90–97.

McCarthy, Mark I et al. (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges". In: *Nature reviews genetics* 9.5, pp. 356–369.

McCulloch, Warren S and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133.

McDonald, Colm and Robin M Murray (2000). "Early and late environmental risk factors for schizophrenia". In: *Brain Research Reviews* 31.2-3, pp. 130–137.

McDonald-McGinn, Donna M et al. (2015). "22q11. 2 deletion syndrome". In: *Nature reviews Disease primers* 1.1, pp. 1–19.

McGrath, JJ and JL Welham (1999). "Season of birth and schizophrenia: a systematic review and meta-analysis of data from the Southern Hemisphere". In: *Schizophrenia research* 35.3, pp. 237–242.

McGrath, John et al. (2004). "A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology". In: *BMC medicine* 2.1, pp. 1–22.

McGrath, John et al. (2008). "Schizophrenia: a concise overview of incidence, prevalence, and mortality". In: *Epidemiologic reviews* 30.1, pp. 67–76.

McGuffin, Peter et al. (1990). "Exclusion of a schizophrenia susceptibility gene from the chromosome 5q11-q13 region: new data and a reanalysis of previous reports." In: *American journal of human genetics* 47.3, p. 524.

McInnis, Melvin G et al. (2003). "Genome-wide scan and conditional analysis in bipolar disorder: evidence for genomic interaction in the National Institute of Mental Health genetics initiative bipolar pedigrees". In: *Biological psychiatry* 54.11, pp. 1265–1273.

McKinney, Wes (2010). "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman, pp. 56 –61. DOI: 10.25080/Majora-92bf1922-00a.

McKinney, Wes et al. (2010). "Data structures for statistical computing in python". In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX, pp. 51–56.

Mednick, Sarnoff A et al. (1988). "Adult schizophrenia following prenatal exposure to an influenza epidemic". In: *Archives of general psychiatry* 45.2, pp. 189–192.

Mehrabi, Ninareh et al. (2019). "A survey on bias and fairness in machine learning". In: *arXiv*.

Mohamed, Somaia et al. (1999). "Generalized cognitive deficits in schizophrenia: a study of first-episode patients". In: *Archives of general psychiatry* 56.8, pp. 749–754.

Moher, David et al. (2009). "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement". In: *Annals of internal medicine* 151.4, pp. 264–269.

Molnar, Christoph (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/.

Monson, Eric T et al. (2017). "Assessment of whole-exome sequence data in attempted suicide within a bipolar disorder cohort". In: *Molecular neuropsychiatry* 3.1, pp. 1–11.

Monteith, Scott et al. (2015). "Big data are coming to psychiatry: a general introduction". In: *International journal of bipolar disorders* 3.1, pp. 1–11.

Moons, Karel GM et al. (2012). "Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker". In: *Heart* 98.9, pp. 683–690.

Moons, Karel GM et al. (2014). "Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist". In: *PLoS medicine* 11.10, e1001744.

Moons, Karel GM et al. (2015). "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration". In: *Annals of internal medicine* 162.1, W1–W73.

Moons, Karel GM et al. (2019). "PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration". In: *Annals of Internal Medicine* 170.1, W1–W33.

Moore, Theresa HM et al. (2007). "Cannabis use and risk of psychotic or affective mental health outcomes: a systematic review". In: *The Lancet* 370.9584, pp. 319–328.

Moreno-Küstner, Berta, Carlos Martin, and Loly Pastor (2018). "Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses". In: *PloS one* 13.4, e0195687.

Mortensen, Preben Bo et al. (1999). "Effects of family history and place and season of birth on the risk of schizophrenia". In: *New England Journal of Medicine* 340.8, pp. 603–608.

Motsinger, Alison A et al. (2006). "GPNN: Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease". In: *BMC bioinformatics* 7.1, pp. 1–10.

Motsinger-Reif, Alison A et al. (2008). "Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error". In: *BMC Research Notes* 1.1, pp. 1–8.

Murphy, Kieran C, Lisa A Jones, and Michael J Owen (1999). "High rates of schizophrenia in adults with velo-cardio-facial syndrome". In: *Archives of general psychiatry* 56.10, pp. 940–945.

Murray, Robin M and Paul Fearon (1999). "The developmental 'risk factor'model of schizophrenia". In: *Journal of psychiatric research* 33.6, pp. 497–499.

Murray, Robin M et al. (2002). *The epidemiology of schizophrenia*. Cambridge University Press.

Need, Anna C et al. (2009). "A genome-wide investigation of SNPs and CNVs in schizophrenia". In: *PLoS Genet* 5.2, e1000373.

Neuman, Rosalind J, John P Rice, and Aravinda Chakravarti (1992). "Two-locus models of disease". In: *Genetic epidemiology* 9.5, pp. 347–365.

Ng, Mandy YM et al. (2009). "Meta-analysis of 32 genome-wide linkage studies of schizophrenia". In: *Molecular psychiatry* 14.8, pp. 774–785.

Nicodemus, Kristin K et al. (2014). "Variability in working memory performance explained by epistasis vs polygenic scores in the ZNF804A pathway". In: *JAMA psychiatry* 71.7, pp. 778–785.

Niculescu-Mizil, Alexandru and Rich Caruana (2005). "Predicting good probabilities with supervised learning". In: *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632.

Niel, Clément et al. (2015). "A survey about methods dedicated to epistasis detection". In: *Frontiers in genetics* 6, p. 285.

Noble, William S (2006). "What is a support vector machine?" In: *Nature biotechnology* 24.12, pp. 1565–1567.

Norman, Ross MG and Ashok K Malla (1993a). "Stressful life events and schizophrenia: I: a review of the research". In: *The British Journal of Psychiatry* 162.2, pp. 161–166.

– (1993b). "Stressful life events and schizophrenia II: Conceptual and methodological issues". In: *The British Journal of Psychiatry* 162.2, pp. 166–174.

Nurnberger, John I et al. (1994). "Diagnostic interview for genetic studies: rationale, unique features, and training". In: *Archives of general psychiatry* 51.11, pp. 849–859.

Ødegaard, Ornulv (1932). "Emigration and insanity". In: *Acta. Psychiatr. Neurol., Suppl.*

O'donovan, Michael C et al. (2008). "Identification of loci associated with schizophrenia by genome-wide association and follow-up". In: *Nature genetics* 40.9, pp. 1053–1055.

Okser, Sebastian, Tapio Pahikkala, and Tero Aittokallio (2013). "Genetic variants and their interactions in disease risk prediction–machine learning and network perspectives". In: *BioData mining* 6.1, p. 5.

Okser, Sebastian et al. (2014). "Regularized machine learning in the genetic prediction of complex traits". In: *PLoS genetics* 10.11, e1004754.

Olfson, Mark et al. (2015). "Premature mortality among adults with schizophrenia in the United States". In: *JAMA psychiatry* 72.12, pp. 1172–1181.

Organization, World Health (2004). *ICD-10 : international statistical classification of diseases and related health problems : tenth revision*. 2nd ed. World Health Organization.

Owen, Michael J, Akira Sawa, and Preben B Mortensen (2016). "Schizophrenia." In: *Lancet* 388.10039, pp. 86–97.

O'Donoghue, Brian, Eric Roche, and Abbie Lane (2016). "Neighbourhood level social deprivation and the risk of psychotic disorders: a systematic review". In: *Social psychiatry and psychiatric epidemiology* 51.7, pp. 941–950.

Pan, Qinxin et al. (2013). "Supervising random forest using attribute interaction networks". In: *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Springer, pp. 104–116.

Pardiñas, Antonio F et al. (2018). "Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection". In: *Nature genetics* 50.3, pp. 381–389.

Parikshak, Neelroop N et al. (2016). "Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism". In: *Nature* 540.7633, pp. 423–427.

Pasman, Joëlle A et al. (2018). "GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal effect of schizophrenia liability". In: *Nature neuroscience* 21.9, pp. 1161–1170.

Patil, Shankargouda et al. (2019). "Machine learning and its potential applications to the genomic study of head and neck cancer — A systematic review". In: *Journal of Oral Pathology & Medicine* 48.9, pp. 773–779.

Pavlou, Menelaos et al. (2015). "How to develop a more accurate risk prediction model when there are few events". In: *BMJ* 351, h3868.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12, pp. 2825–2830.

Peralta, Victor and Manuel J Cuesta (1999). "Diagnostic significance of Schneider's first-rank symptoms in schizophrenia: Comparative study between schizophrenic and non-schizophrenic psychotic disorders". In: *The British Journal of Psychiatry* 174.3, pp. 243–248.

Perkins, Diana O et al. (2020). "Polygenic risk score contribution to psychosis prediction in a target population of persons at clinical high risk". In: *American Journal of Psychiatry* 177.2, pp. 155–163.

Pettersson-Yeo, William et al. (2013). "Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level". In: *Psychological medicine* 43.12, pp. 2547–2562.

Phillips, DI W (1993). "Twin studies in medical research: can they tell us whether diseases are genetically determined?" In: *Lancet* 341.8851, pp. 1008–1009.

Phillips, Patrick C (1998). "The language of gene interaction". In: *Genetics* 149.3, pp. 1167–1171.

Pinto, Dalila et al. (2010). "Functional impact of global rare copy number variation in autism spectrum disorders". In: *Nature* 466.7304, pp. 368–372.

Pinto, Dalila et al. (2011). "Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants". In: *Nature biotechnology* 29.6, pp. 512–520.

Pinto, Dalila et al. (2014). "Convergence of genes and cellular pathways dysregulated in autism spectrum disorders". In: *The American Journal of Human Genetics* 94.5, pp. 677–694.

Pirooznia, Mehdi et al. (2012). "Data mining approaches for genome-wide association of mood disorders". In: *Psychiatric genetics* 22.2, p. 55.

Platt, John et al. (1999). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3, pp. 61–74.

Ploeg, Tjeerd van der, Peter C Austin, and Ewout W Steyerberg (2014). "Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints". In: *BMC medical research methodology* 14.1, p. 137.

Plomin, Robert et al. (2013). *Behavioral Genetics, 6th Edition*. Worth Publishers, New York.

Plot digitizer. URL: http://plotdigitizer.sourceforge.net/ (visited on 02/07/2020).

Porter, Theodore M (2020). *Genetics in the madhouse: The unknown history of human heredity*. Princeton University Press.

Power, Chris and Jane Elliott (2006). "Cohort profile: 1958 British birth cohort (national child development study)". In: *International journal of epidemiology* 35.1, pp. 34–41.

Power, Robert A et al. (2015). "Polygenic risk scores for schizophrenia and bipolar disorder predict creativity". In: *Nature neuroscience* 18.7, pp. 953–955.

Price, Alkes L et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature genetics* 38.8, p. 904.

PsychENCODE Integrative Analysis. https://www.nimhgenetics.org/resources/psychencode. Accessed: 28/11/2019.

Purcell, Shaun et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American journal of human genetics* 81.3, pp. 559–575.

Purcell, Shaun M et al. (2009). "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder". In: *Nature* 460.7256, pp. 748–752.

Purcell, Shaun M et al. (2014). "A polygenic burden of rare disruptive mutations in schizophrenia". In: *Nature* 506.7487, p. 185.

Qian, Junyang et al. (2020). "A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank". In: *PLoS genetics* 16.10, e1009141.

Quinlan, J. Ross (1986). "Induction of decision trees". In: *Machine learning* 1.1, pp. 81–106.

Quinlan, J Ross (2014). *C4.5: programs for machine learning*. Elsevier.

Rabinowitz, Jonathan et al. (2000). "Cognitive and behavioural functioning in men with schizophrenia both before and shortly after first admission to hospital: Cross-sectional analysis". In: *The British Journal of Psychiatry* 177.1, pp. 26–32.

Rajji, Tarek K et al. (2013). "Cognitive performance of individuals with schizophrenia across seven decades: a study using the MATRICS consensus cognitive battery". In: *The American Journal of Geriatric Psychiatry* 21.2, pp. 108–118.

Raschka, Sebastian (2018). "Model evaluation, model selection, and algorithm selection in machine learning". In: *arXiv*.

Rees, Elliott et al. (2014). "Analysis of copy number variations at 15 schizophrenia-associated loci". In: *The British Journal of Psychiatry* 204.2, pp. 108–114.

Rees, Elliott et al. (2016). "Analysis of intellectual disability copy number variants for association with schizophrenia". In: *JAMA psychiatry* 73.9, pp. 963–969.

Rees, Elliott et al. (2020). "De novo mutations identified by exome sequencing implicate rare missense variants in SLC6A1 in schizophrenia". In: *Nature neuroscience* 23.2, pp. 179–184.

Regier, Darrel A et al. (2013). "DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses". In: *American journal of psychiatry* 170.1, pp. 59–70.

Reich, David E and Eric S Lander (2001). "On the allelic spectrum of human disease". In: *Trends in Genetics* 17.9, pp. 502–510.

Reichenberg, Abraham et al. (2002). "A population-based cohort study of premorbid intellectual, language, and behavioral functioning in patients with schizophrenia, schizoaffective disorder, and nonpsychotic bipolar disorder". In: *American Journal of Psychiatry* 159.12, pp. 2027–2035.

Reichenberg, Abraham et al. (2005). "Elaboration on premorbid intellectual performance in schizophrenia: premorbid intellectual decline and risk for schizophrenia". In: *Archives of General Psychiatry* 62.12, pp. 1297–1304.

Ripke, Stephan et al. (2011). "Genome-wide association study identifies five new schizophrenia loci". In: *Nature genetics* 43.10, p. 969.

Ripke, Stephan et al. (2013). "Genome-wide association analysis identifies 13 new risk loci for schizophrenia". In: *Nature genetics* 45.10, p. 1150.

Ripke, Stephan et al. (2014). "Biological insights from 108 schizophrenia-associated genetic loci". In: *Nature* 511.7510, p. 421.

Ripke, Stephan et al. (2020). "Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia". In: *MedRxiv*.

Ritchie, Marylyn D et al. (2003). "Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases". In: *BMC bioinformatics* 4.1, pp. 1–14.

Ritchie, Marylyn D et al. (2007). "Genetic programming neural networks: A powerful bioinformatics tool for human genetics". In: *Applied Soft Computing* 7.1, pp. 471–479.

Ronalds, Georgina A, Bianca L De Stavola, and David A Leon (2005). "The cognitive cost of being a twin: evidence from comparisons within families in the Aberdeen children of the 1950s cohort study". In: *BMJ* 331.7528, p. 1306.

Rosenthal, David (1960). "Confusion of identity and the frequency of schizophrenia in twins". In: *Archives of General Psychiatry* 3.3, pp. 297–304.

Rosenthal, David et al. (1968). "Schizophrenics' offspring reared in adoptive homes". In: *Journal of psychiatric research* 6, pp. 377–391.

Rothman, Kenneth J (2012). *Epidemiology: an introduction*. Oxford university press.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1985). *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science.

Saha, S et al. (2006). "The incidence and prevalence of schizophrenia varies with latitude". In: *Acta Psychiatrica Scandinavica* 114.1, pp. 36–39.

Saha, Sukanta et al. (2005). "A systematic review of the prevalence of schizophrenia". In: *PLoS Med* 2.5, e141.

Sanders, Alan R et al. (2008). "No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: implications for psychiatric genetics". In: *American Journal of Psychiatry* 165.4, pp. 497–506.

Schapire, Robert E (1990). "The strength of weak learnability". In: *Machine learning* 5.2, pp. 197–227.

Schnack, Hugo G (2017). "Improving individual predictions: machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases)". In: *Schizophrenia research*.

Schneider, Kurt (1959). *Clinical psychopathology*. Grune & Stratton.

Schreier, Andrea et al. (2009). "Prospective study of peer victimization in childhood and psychotic symptoms in a nonclinical population at age 12 years". In: *Archives of general psychiatry* 66.5, pp. 527–536.

Schrider, Daniel R and Andrew D Kern (2018). "Supervised machine learning for population genetics: a new paradigm". In: *Trends in Genetics* 34.4, pp. 301–312.

Schwarz, Daniel F, Inke R König, and Andreas Ziegler (2010). "On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data". In: *Bioinformatics* 26.14, pp. 1752–1758.

Seabold, Skipper and Josef Perktold (2010). "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*.

Segal, Mark and Yuanyuan Xiao (2011). "Multivariate random forests". In: *Wiley interdisciplinary reviews: Data mining and knowledge discovery* 1.1, pp. 80–87.

Selten, Jean-Paul et al. (2002). "Ødegaard's selection hypothesis revisited: schizophrenia in Surinamese immigrants to The Netherlands". In: *American Journal of Psychiatry* 159.4, pp. 669–671.

Senden Theis, Sophie van (1924). *How Foster Children Turn Out: A Study and Critical Analysis of 910 Children who Were Placed in Foster Homes by the State Charities Aid Association and who are Now Eighteen Years of Age Or Over*. 165. State Charities Aid Association.

Sharpley, Mandy et al. (2001). "Understanding the excess of psychosis among the African-Caribbean population in England: review of current hypotheses". In: *The British Journal of Psychiatry* 178.S40, s60–s68.

Shen, Yuanyuan, Zhe Liu, and Jurg Ott (2010). "Detecting gene-gene interactions using support vector machines with L 1 penalty". In: *2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*. IEEE, pp. 309–311.

Sherrington, Robin et al. (1988). "Localization of a susceptibility locus for schizophrenia on chromosome 5". In: *Nature* 336.6195, pp. 164–167.

Shi, Jianxin et al. (2009). "Common variants on chromosome 6p22. 1 are associated with schizophrenia". In: *Nature* 460.7256, pp. 753–757.

Simeone, Jason C et al. (2015). "An evaluation of variation in published estimates of schizophrenia prevalence from 1990-2013: a systematic literature review". In: *BMC psychiatry* 15.1, pp. 1–14.

Singh, Tarjinder et al. (2020). "Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia". In: *medRxiv*.

Sitskoorn, Margriet M et al. (2004). "Cognitive deficits in relatives of patients with schizophrenia: a meta-analysis". In: *Schizophrenia research* 71.2-3, pp. 285–295.

Smeland, Olav B et al. (2020). "The polygenic architecture of schizophrenia—Rethinking pathogenesis and nosology". In: *Nature Reviews Neurology* 16.7, pp. 366–379.

Smemo, Scott et al. (2014). "Obesity-associated variants within FTO form long-range functional connections with IRX3". In: *Nature* 507.7492, pp. 371–375.

Snell, Kym IE et al. (2018). "Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures?" In: *Statistical methods in medical research* 27.11, pp. 3505–3522.

So, Hon-Cheong and Pak C Sham (2010). "A unifying framework for evaluating the predictive power of genetic variants based on the level of heritability explained". In: *PLoS genetics* 6.12, e1001230.

Sommer, Iris et al. (2001). "Handedness, language lateralisation and anatomical asymmetry in schizophrenia: meta-analysis". In: *The British Journal of Psychiatry* 178.4, pp. 344–351.

Spitzer, Robert L, Joseph L Fleiss, et al. (1974). "A re-analysis of the reliability of psychiatric diagnosis". In: *British Journal of Psychiatry* 125.0, pp. 341–347.

Srivastava, Nitish et al. (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1, pp. 1929–1958.

St Clair, David et al. (2005). "Rates of adult schizophrenia following prenatal exposure to the Chinese famine of 1959-1961". In: *Jama* 294.5, pp. 557–562.

Stahl, Eli A et al. (2019). "Genome-wide association study identifies 30 loci associated with bipolar disorder". In: *Nature genetics* 51.5, pp. 793–803.

Stefansson, Hreinn et al. (2002). "Neuregulin 1 and susceptibility to schizophrenia". In: *The American Journal of Human Genetics* 71.4, pp. 877–892.

Stefansson, Hreinn et al. (2009). "Common variants conferring risk of schizophrenia". In: *Nature* 460.7256, pp. 744–747.

Stephan, Johannes, Oliver Stegle, and Andreas Beyer (2015). "A random forest approach to capture genetic effects in the presence of population structure". In: *Nature communications* 6, p. 7432.

Steyerberg, Ewout W et al. (2001). "Internal validation of predictive models: efficiency of some procedures for logistic regression analysis". In: *Journal of clinical epidemiology* 54.8, pp. 774–781.

Steyerberg, Ewout W et al. (2010). "Assessing the performance of prediction models: a framework for some traditional and novel measures". In: *Epidemiology (Cambridge, Mass.)* 21.1, p. 128.

Steyerberg, Ewout W et al. (2019). *Clinical prediction models, 2nd Edition*. Springer.

Stilo, Simona A and Robin M Murray (2019). "Non-genetic factors in schizophrenia". In: *Current psychiatry reports* 21.10, pp. 1–10.

Straub, Richard E et al. (2002). "Genetic variation in the 6p22. 3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia". In: *The American Journal of Human Genetics* 71.2, pp. 337–348.

Strobl, Carolin et al. (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8.1, pp. 1–21.

Stroup, T Scott et al. (2003). "The National Institute of Mental Health clinical antipsychotic trials of intervention effectiveness (CATIE) project: schizophrenia trial design and protocol development". In: *Schizophrenia bulletin* 29.1, pp. 15–31.

Sudlow, Cathie et al. (2015). "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLoS medicine* 12.3.

Sullivan, Patrick F, Kenneth S Kendler, and Michael C Neale (2003). "Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies". In: *Archives of general psychiatry* 60.12, pp. 1187–1192.

Sullivan, Patrick F et al. (2008). "Genomewide association for schizophrenia in the CATIE study: results of stage 1". In: *Molecular psychiatry* 13.6, pp. 570–584.

Sun, Guo-Wen, Thomas L Shook, and Gregory L Kay (1996). "Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis". In: *Journal of clinical epidemiology* 49.8, pp. 907–916.

Susser, Ezra et al. (1996). "Schizophrenia after prenatal famine: further evidence". In: *Archives of general psychiatry* 53.1, pp. 25–31.

Susser, Ezra S and Shang P Lin (1992). "Schizophrenia after prenatal exposure to the Dutch Hunger Winter of 1944-1945". In: *Archives of general psychiatry* 49.12, pp. 983–988.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*, pp. 3104–3112.

Sutskever, Ilya et al. (2013). "On the importance of initialization and momentum in deep learning". In: *International conference on machine learning*. PMLR, pp. 1139–1147.

Tandon, Neeraj and Rajiv Tandon (2018). "Will machine learning enable us to finally cut the gordian knot of schizophrenia". In: *Schizophrenia Bulletin* 44.5, pp. 939–941.

– (2019). "Machine learning in psychiatry-standards and guidelines". In: *Asian Journal of Psychiatry* 44, A1–A4.

Tandon, Rajiv (2013). "Schizophrenia and other psychotic disorders in DSM-5: clinical implications of revisions from DSM-IV". In: *Clinical schizophrenia & related psychoses* 7.1, pp. 16–19.

Tansey, Katherine E et al. (2016). "Common alleles contribute to schizophrenia in CNV carriers". In: *Molecular psychiatry* 21.8, pp. 1085–1089.

Teschendorff, Andrew E (2019). "Avoiding common pitfalls in machine learning omic data science". In: *Nature Materials* 18.5, pp. 422–427.

The pandas development team (Mar. 2020). *pandas-dev/pandas: Pandas 1.0.3*. Version v1.0.3. DOI: 10.5281/zenodo.3715232. URL: https://doi.org/10.5281/zenodo.3715232.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tieleman, Tijmen and Geoffrey Hinton (2012). "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude". In: *COURSERA: Neural networks for machine learning* 4.2, pp. 26–31.

Tienari, PEKKA et al. (1985). "The Finnish adoptive family study of schizophrenia." In: *The Yale Journal of Biology and Medicine* 58.3, p. 227.

Ting, Kai Ming and Ian H Witten (1999). "Issues in stacked generalization". In: *Journal of artificial intelligence research* 10, pp. 271–289.

Tomita, Yasuyuki et al. (2004). "Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma". In: *BMC bioinformatics* 5.1, pp. 1–13.

Torrey, E Fuller, Barbara Boyle Torrey, and Michael R Peterson (1977). "Seasonality of schizophrenic births in the United States". In: *Archives of General Psychiatry* 34.9, pp. 1065–1070.

Töscher, Andreas, Michael Jahrer, and Robert M Bell (2009). "The bigchaos solution to the netflix grand prize". In: *Netflix prize documentation*, pp. 1–52.

Townsend, Peter, Peter Phillimore, and Alastair Beattie (1988). *Health and deprivation: inequality and the North*. Routledge.

Trakadis, Yannis J et al. (2019). "Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes". In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 180.2, pp. 103–112.

Üçok, Alp and S Bıkmaz (2007). "The effects of childhood trauma in patients with first-episode schizophrenia". In: *Acta Psychiatrica Scandinavica* 116.5, pp. 371–377.

Uppu, Suneetha, Aneesh Krishna, and Raj P Gopalan (2016a). "A deep learning approach to detect SNP interactions." In: *JSW* 11.10, pp. 965–975.

– (2016b). "A review on methods for detecting SNP interactions in high-dimensional genomic data". In: *IEEE/ACM transactions on computational biology and bioinformatics* 15.2, pp. 599–612.

Vabalas, Andrius et al. (2019). "Machine learning algorithm validation with a limited sample size". In: *PloS one* 14.11, e0224365.

Van Calster, Ben et al. (2019). "Calibration: the Achilles heel of predictive analytics". In: *BMC medicine* 17.1, pp. 1–7.

Van Os, Jim, Gunter Kenis, and Bart PF Rutten (2010). "The environment and schizophrenia". In: *Nature* 468.7321, pp. 203–212.

Van Os, Jim et al. (2009). "A systematic review and meta-analysis of the psychosis continuum: evidence for a psychosis proneness-persistence-impairment model of psychotic disorder". In: *Psychological medicine* 39.2, p. 179.

Varese, Filippo et al. (2012). "Childhood adversities increase the risk of psychosis: a meta-analysis of patient-control, prospective-and cross-sectional cohort studies". In: *Schizophrenia bulletin* 38.4, pp. 661–671.

Varma, Sudhir and Richard Simon (2006). "Bias in error estimation when using cross-validation for model selection". In: *BMC bioinformatics* 7.1, p. 91.

Vassos, Evangelos et al. (2012). "Meta-analysis of the association of urbanicity with schizophrenia". In: *Schizophrenia bulletin* 38.6, pp. 1118–1123.

Vaucher, Julien et al. (2018). "Cannabis use and risk of schizophrenia: a Mendelian randomization study". In: *Molecular psychiatry* 23.5, pp. 1287–1292.

Ven, E Van der et al. (2015). "Testing Ødegaard's selective migration hypothesis: a longitudinal cohort study of risk factors for non-affective psychotic disorders among prospective emigrants". In: *Psychological Medicine* 45.4, p. 727.

Verdoux, Hélène and Anne-Laure Sutter (2002). "Perinatal risk factors for schizophrenia: diagnostic specificity and relationships with maternal psychopathology". In: *American journal of medical genetics* 114.8, pp. 898–905.

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272.

Vivian-Griffiths, Timothy et al. (2019). "Predictive modeling of schizophrenia from genomic data: Comparison of polygenic risk score with kernel support vector machines approach". In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 180.1, pp. 80–85.

Wahlbeck, K et al. (2001). "Associations between childhood living circumstances and schizophrenia: a population-based cohort study". In: *Acta Psychiatrica Scandinavica* 104.5, pp. 356–360.

Walt, Stéfan van der, S Chris Colbert, and Gael Varoquaux (2011). "The NumPy array: a structure for efficient numerical computation". In: *Computing in science & engineering* 13.2, pp. 22–30.

Wan, Xiang et al. (2010). "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies". In: *The American Journal of Human Genetics* 87.3, pp. 325–340.

Wang, Daifeng et al. (2018). "Comprehensive functional genomic resource and integrative model for the human brain". In: *Science* 362.6420, eaat8464.

Ware, Jennifer J et al. (2016). "Genome-wide meta-analysis of cotinine levels in cigarette smokers identifies locus at 4q13. 2". In: *Scientific reports* 6, p. 20092.

Waskom, Michael et al. (Apr. 2020). *mwaskom/seaborn: v0.10.1 (April 2020)*. Version v0.10.1. DOI: 10.5281/zenodo.3767070. URL: https://doi.org/10.5281/zenodo.3767070.

Wei, Changshuai and Qing Lu (2014). "GWGGI: software for genome-wide gene-gene interaction analysis". In: *BMC genetics* 15.1, pp. 1–6.

Wing, John Kenneth et al. (1990). "SCAN: schedules fonr clinical assessment in neuropsychiatry". In: *Archives of general psychiatry* 47.6, pp. 589–593.

Winham, Stacey J et al. (2012). "SNP interaction detection with random forests in high-dimensional genetic data". In: *BMC bioinformatics* 13.1, pp. 1–13.

Wirshing, Donna A (2004). "Schizophrenia and obesity: impact of antipsychotic medications." In: *The Journal of clinical psychiatry*.

Wolff, Robert F et al. (2019). "PROBAST: a tool to assess the risk of bias and applicability of prediction model studies". In: *Annals of Internal Medicine* 170.1, pp. 51–58.

Wolpert, David H (1992). "Stacked generalization". In: *Neural networks* 5.2, pp. 241–259.

Woo, Hyung Jun et al. (2017). "Large-scale interaction effects reveal missing heritability in schizophrenia, bipolar disorder and posttraumatic stress disorder". In: *Translational psychiatry* 7.4, e1089–e1089.

Wray, Naomi R et al. (2010). "The genetic interpretation of area under the ROC curve in genomic profiling". In: *PLoS genetics* 6.2, e1000864.

Wright, Marvin N, Andreas Ziegler, and Inke R König (2016). "Do little interactions get lost in dark random forests?" In: *BMC bioinformatics* 17.1, pp. 1–10.

Wynne, Lyman C and Margaret Thaler Singer (1963). "Thought disorder and family relations of schizophrenics: I. A research strategy". In: *Archives of general Psychiatry* 9.3, pp. 191–198.

Yang, Honghui et al. (2010a). "A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia". In: *Frontiers in human neuroscience* 4, p. 192.

Yang, Jian et al. (2010b). "Common SNPs explain a large proportion of the heritability for human height". In: *Nature genetics* 42.7, p. 565.

Yoshida, Makiko and Asako Koike (2011). "SNPInterForest: a new method for detecting epistatic interactions". In: *BMC bioinformatics* 12.1, pp. 1–10.

Yung, Ling Sing et al. (2011). "GBOOST: a GPU-based tool for detecting gene–gene interactions in genome–wide case control studies". In: *Bioinformatics* 27.9, pp. 1309–1310.

Zammit, Stanley et al. (2002). "Self reported cannabis use as a risk factor for schizophrenia in Swedish conscripts of 1969: historical cohort study". In: *BMJ* 325.7374, p. 1199.

Zhang, Yu and Jun S Liu (2007). "Bayesian inference of epistatic interactions in case-control studies". In: *Nature genetics* 39.9, p. 1167.

Zhao, Qingyu, Ehsan Adeli, and Kilian M Pohl (2020). "Training confounder-free deep learning models for medical applications". In: *Nature communications* 11.1, pp. 1–9.

Zhao, Yang et al. (2012). "Correction for population stratification in random forest analysis". In: *International journal of epidemiology* 41.6, pp. 1798–1806.

Zheutlin, Amanda B et al. (2018). "Multivariate Pattern Analysis of Genotype-Phenotype Relationships in Schizophrenia". In: *Schizophrenia bulletin* 44.5, pp. 1045–1052.