Prognostic research in psychiatry: towards a clinically-relevant prediction model for first episode psychosis

Samuel P Leighton

BMedSci(Hons), MBChB, MRCP(UK), MRCPsych

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Health & Wellbeing

College of Medical, Veterinary and Life Sciences

University of Glasgow

December 2023

Abstract

Background. Prognosis is the determination of risk of future health outcomes in people with a given health condition. The primary aim for my thesis was to conduct prognostic model research into first episode psychosis (FEP). The prognosis of people with FEP is poor in around half of those affected and difficult to predict in individuals. Prognostic prediction models to predict outcome in individuals could facilitate early intervention to change clinical trajectories and improve prognosis. As part of my primary aim, I sought to answer four research questions. 1) Is prediction of individual patient outcome possible in FEP using clinical variables? 2) Does prediction model performance remain robust at external validation? 3) Does prediction model performance improve with the addition of biologically relevant disease markers? 4) Does prediction model performance improve with the application of advanced machine learning classifiers compared to logistic regression? These questions are addressed in studies 1 to 3.

The secondary aim for my thesis was to test whether routinely collected electronic healthcare record data could be used for prognostic research in the National Health Service (NHS) in Greater Glasgow and Clyde (GG&C). The coronavirus pandemic delayed collection of routine data in FEP. I took the opportunity to examine this question in a more common area of psychiatric disease, delirium, in the hope that information from this would inform future prospective studies in FEP. Delirium is an important risk factor for subsequent dementia. However, the field lacks large studies with long-term follow-up of delirium in subjects initially free of dementia to clearly establish clinical trajectories. This formed study 4.

Study 1. This study aimed to conduct a systematic review of prognostic prediction models developed for predicting poor outcome in FEP. Thirteen studies reporting 31 prediction models across a range of clinical outcomes met criteria for inclusion. Eleven studies used logistic regression with clinical variables. External validation was carried out in four studies. Only one study assessed whether biologically relevant disease markers added value as predictors. Two studies used machine learning but did not provide enough information to allow comparison to logistic regression. Most studies had

methodological flaws and the potential for prediction modelling in FEP is yet to be fully realised.

Study 2. This study aimed to develop and externally validate a prognostic prediction model of symptom nonremission in FEP developed using multivariable logistic regression employing clinical variables. The development cohort consisted of 673 FEP patients and the validation cohort consisted of 191 FEP patients. The prediction model showed good discrimination C-statistic of 0.73 (0.64, 0.81) and adequate calibration with intercept alpha of -0.014 (-0.34, 0.31) and slope beta of 0.85 (0.42, 1.27). The model improved the net-benefit by 16% at a risk threshold of 50% compared to the strategy of treating all. The model could allow clinicians to intervene earlier to change trajectories and improve prognosis in first episode psychosis but first requires prospective validation and its clinical impact established in a future trial.

Study 3. This study assessed the potential for biologically relevant disease markers as predictor variables and compared advanced machine learning classifiers to logistic regression in 168 patients with FEP. The addition of a biological variable did not improve the performance of a logistic regression model built using clinical variables. It is possible that the usefulness of the biological variables for prediction was curtailed by the lack of a mechanistic link to the pathophysiology of psychosis thereby limiting their effect size. The naïve Bayes machine learning model was better than maximum likelihood estimation (MLE) but not elastic net logistic regression in terms of discrimination. However, for all models except MLE logistic regression there were problems with calibration.

Study 4. This study consisted of a retrospective cohort study of all patients over the age of 65 diagnosed with an episode of delirium who were initially dementia free at onset of delirium within NHS GG&C between 1996 and 2020 using routinely collected electronic healthcare record (EHR) data. 12949 patients with an incident episode of delirium were included and followed up for an average of 741 days. The estimated cumulative incidence of dementia was 31% by 5 years. The estimated cumulative incidence of the competing risk of death without dementia was 49.2% by 5 years. The cause-specific hazard of dementia was

increased with higher levels of deprivation and also with advancing age from 65, plateauing and decreasing from age 90.

Conclusions. Systematic review of the literature showed that there is considerable potential for prognostic prediction modelling in FEP, but that most existing models have methodological flaws. Developing on this literature, my FEP prognostic prediction model can help to identify individual patients at increased risk of nonremission at initial clinical contact and showed robust external validation. However, this approach did not benefit from the addition of biologically relevant disease markers as predictor variables or the application of machine learning methods. Finally, I demonstrated the feasibility of using routinely collected EHR data from NHS GG&C for prognostic research into delirium and the risk of subsequent dementia. This will inform future prospective prognostic modelling studies of routinely collected data in FEP. Altogether, this thesis made several contributions to the growing body of clinical prognostic research in first episode psychosis and delirium. In particular, considerable progress has been made towards the deployment of a useable and informative clinical prediction model which will improve care for people with first episode psychosis.

Table of Contents

List of Tables

List of Figures

Publications arising from this thesis

Conference proceedings

Leighton, S. P., Krishnadas, R., Upthegrove, R., Marwaha, S., Steyerberg, E. W., Gkoutos, G. V., Broome, M. R., Liddle, P. F., Everard, L., Singh, S. P., Freemantle, N., Fowler, D., Jones, P. B., Sharma, V., Murray, R., Wykes, T., Drake, R. J., Buchan, I., Rogers, S., Cavanagh, J., Lewis, S. W., Birchwood, M., Mallikarjun, P. K. (2021). Development and Validation of a Nonremission Risk Prediction Model in First-Episode Psychosis: An Analysis of 2 Longitudinal Studies. Awarded 2nd place in the rapid-fire poster presentation at RCPsych International Congress 21-24 June 2021.

Leighton, S.P., Deligianni, F., Cavanagh, J. (2023) Prediction modelling in first episode psychosis: an assessment of biological disease markers and machine learning classifiers. Rapid-fire poster presentation at RCPsych International Congress 10-13 July 2023.

Articles

Lee, R.*, Leighton, S. P.*, Thomas, L., Gkoutos, G. V., Wood, S. J., Fenton, S-J. H., Deligianni, F., Cavanagh, J., Mallikarjun, P. K. (2022). Prediction models in first-episode psychosis: systematic review and critical appraisal. The British Journal of Psychiatry, 220(4), 179-191. doi:10.1192/bjp.2021.219
* Joint first authors.

Leighton, S. P., Krishnadas, R., Upthegrove, R., Marwaha, S., Steyerberg, E. W., Gkoutos, G. V., Broome, M. R., Liddle, P. F., Everard, L., Singh, S. P., Freemantle, N., Fowler, D., Jones, P. B., Sharma, V., Murray, R., Wykes, T., Drake, R. J., Buchan, I., Rogers, S., Cavanagh, J., Lewis, S. W., Birchwood, M., Mallikarjun, P. K. (2021). Development and Validation of a Nonremission Risk Prediction Model in First-Episode Psychosis: An Analysis of 2 Longitudinal Studies. Schizophr Bull Open, 2(1), sgab041. doi:10.1093/schizbullopen/sgab041

Leighton, S. P., Herron, J. W., Jackson, E., Sheridan, M., Deligianni, F., & Cavanagh, J. (2022). Delirium and the risk of developing dementia: a cohort

study of 12 949 patients. J Neurol Neurosurg Psychiatry, 93(8), 822-827. doi:10.1136/jnnp-2022-328903

Acknowledgements

There are a great many people to whom I owe an immense debt of gratitude for helping me on my PhD journey.

In particular, I would like to thank:

- My wife Danielle and daughter Hannah who was born during my PhD.
- My family especially my parents (Jack and Kath), siblings (Alison, Emma and Steven), Jackie and Adam, cousin Matthew and my grandma Sylvia who sadly passed away during my PhD.
- My PhD supervisors Prof Jonathan Cavanagh, Dr Rajeev Krishnadas, Dr Simon Rogers and Dr Fani Deligianni.
- Dr Pavan Mallikarjun without whom my PhD would not have been possible.
- Friends and colleagues in Jonathan's lab group past and present including Dr James Herron and Dr Alison McColl.
- My clinical supervisors who helped me balance my training in psychiatry with my research – Dr Wai Lan Imrie, Dr Alex Thom, Dr Jonathan Dourish and Dr Pavan Srireddy.
- All the patients who participated in the studies that were included in my thesis and the research teams that conducted them.
- My co-authors for the articles and conference proceedings resulting from this thesis.
- My funder – the Chief Scientist Office of Scotland.

Author's Declaration

I declare that the work described in this thesis is original and, except where explicit reference is made to the contribution of others, represents entirely my own efforts. None of the data included in this thesis has been submitted for any other degree at the University of Glasgow or any other institution.

Samuel P. Leighton

Abbreviations

1H = proton

95% CI = 95% confidence interval

A&E = accident and emergency

ANOVA = analysis of variance

AUC = area under the curve

BG = basal ganglia

CHARMS = CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies

CIF = cumulative incidence function

CITL = calibration-in-the-large

Cr = creatine

CRAN = Comprehensive R Archive Network

c-statistic = concordance statistic

d = Cohen's d

DAS = Disability Assessment Schedule

Dev. = development sample

DFBETAS = difference in beta values

DSM-IV = Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition

DUP = duration of untreated psychosis

EET = employment, education or training

EHR = electronic healthcare record

EMPATH = Electronic Measures in Psychosis – Assessing Trajectory and Health-Outcomes

EPP = events per predictor parameter

EPV = events per variable

EQ-5D = EuroQol five-dimensions

F = frontal cortex

F = one-way independent ANOVA

f/u = follow-up

FEP = first episode psychosis

FOV = field-of-view

GABA = γ-aminobutyric acid

GAF = Global Assessment of Functioning

GG&C = Greater Glasgow & Clyde

GLX = combined glutamate and glutamine signal

GRADE = Grading of Recommendations Assessment, Development and Evaluation

H = hippocampus

ICD-10 = Tenth Revision of the International Classification of Diseases

IDO = indoleamine 2, 3-dioxygenase

IMD = Index of Multiple Deprivation

IQR = interquartile range

LASSO = least absolute shrinkage and selection operator

LP = linear predictor

LR = logistic regression

MLE = maximum likelihood estimation

MLR = monocyte/lymphocyte ratio

MRS = magnetic resonance spectroscopy

N/A = not applicable

NEDEN = National Evaluation of Development of Early intervention Network study

NEX = number of excitations

NHS = National Health Service

NICE = National Institute for Health and Care Excellence

NLR = neutrophil/lymphocyte ratio

NMDA = N-methyl-D-aspartate

NPV = negative predictive value

NR = not reported

OR = odds ratio

PANSS = Positive And Negative Syndrome Scale

PAS = Premorbid Adjustment Scale

PCT = primary care trust

PET = positron emission tomography

PICOTS = Population, Index, Comparator, Outcome, Timing and Setting

PPV = positive predictive value

PRESS = Point RESolved Spectroscopic

PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PROBAST = Prediction model Risk Of Bias ASsessment Tool

PROGRESS = PROGnosis RESearch Strategy

PROSPERO = International Prospective Register of Systematic Reviews

PSI = prognostic summary index

r = Pearson's correlation coefficient

RECORD = Reporting of studies Conducted using Observational Routinely-collected health Data

ROB = risk of bias

ROC = receiver operating characteristic

ROI = region of interest

RR = relative risk

SD = standard deviation

SIGN = Scottish Intercollegiate Guidelines Network

SIMD = Scottish Index of Multiple Deprivation

STROBE = Strengthening the Reporting of Observational Studies in Epidemiology

SVM = support vector machine

t = independent t-test

TE = time to echo

TR = repetition time

TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

UK = United Kingdom

Val. = validation sample

W = Wilcoxon rank sum test

WHO = World Health Organization

x^2 = Pearson's chi-squared test

Chapter 1 Introduction

1.1 Prognosis research

The central theme of my PhD thesis is prognosis. Prognosis is the determination of risk of future health outcomes in people with a given health condition (Hemingway et al., 2013). Prognostic research is of considerable importance. Globally, there are more people living with health conditions than ever before (Mathers & Loncar, 2006). Prognostic research seeks to improve the outcomes of people living with health conditions. However, there is a disparity between the potential and actual impact of prognostic research. Research studies often fall short of the high methodological standards required. Serious flaws in the design, conduct and reporting of prognosis studies have been identified. In response, the PROGnosis RESearch Strategy (PROGRESS) initiative has established standards for higher quality prognostic research.

PROGRESS centres on four themes (Figure 1-1): 1) fundamental prognosis research, which investigates the course of health conditions in the context of their current care; 2) prognostic factor research, which looks at specific factors associated with prognosis; 3) prognostic model research, which is concerned with the development, validation and impact of models incorporating multiple prognostic factors, and; 4) stratified medicine research, which focuses on the use of prognostic information to tailor treatments to individuals or groups according to their shared risk (Hemingway et al., 2013; Hingorani et al., 2013; Riley et al., 2013; Steyerberg et al., 2013).

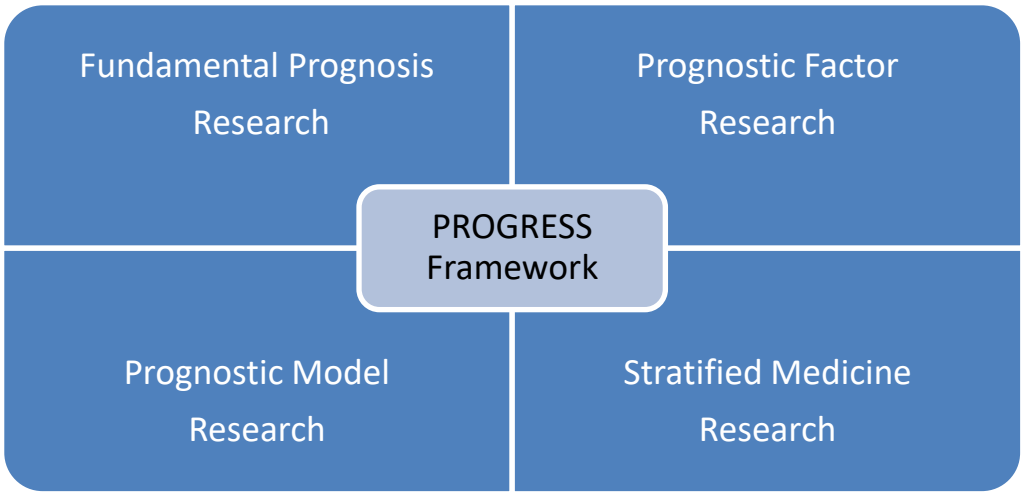


Figure 1-1 The four themes which make up the PROGRESS Framework.

My thesis will address each of these four PROGRESS themes and is laid out in two sections. The first section and main body of the work is focussed on the third PROGRESS theme, prognostic model research, and will touch on the fourth PROGRESS theme, stratified medicine research. The clinical problem being addressed is psychosis.

The second section of my PhD seeks to address the feasibility of using routinely collected clinical data for prognostic research. This section is focussed on the first PROGRESS theme, fundamental prognosis research, and the second PROGRESS theme, prognostic factor research. As I will lay out below, as a consequence of the global coronavirus pandemic the clinical problem I addressed changed to delirium and dementia.

Fundamental prognosis research encompasses studies describing and explaining future outcomes in people with a given health condition in the context of current clinical practice. Fundamental prognosis research is vital for our overall understanding of a health condition. It informs clinical care as well as broader public health policy. Public health policy makers require estimates of average prognosis to model the population burden of diseases as well as to assess the relative contribution of healthcare delivery among those with and without disease (i.e., primary and secondary prevention). Fundamental prognosis research also focuses future research goals including into disease mechanisms and potential therapeutic targets (Hemingway et al., 2013).

Prognostic factor research involves research into specific factors that are associated with prognosis. A prognostic factor is a measure recorded in people with a given health condition that is associated with subsequent clinical outcome. Measures can be at the individual level like biomarkers or at the ecological level such as social deprivation. Research into prognostic factors attempts to discover and understand factors which could be modifiable targets for interventions, components of a prognostic model or predictors of treatment response (Riley et al., 2013).

I will outline and explain prognostic model research in detail below as well as touch on its relationship with stratified medicine.

1.2 Psychosis

1.2.1 Overview

Psychosis is derived from the Greek for abnormal condition of the mind. It is a serious mental disorder characterised by positive symptoms and negative symptoms. Positive symptoms include hallucinations (perceptions in the absence of a stimulus), delusions (fixed or falsely held beliefs) and disorganised behaviour, speech and thoughts. Negative symptoms include emotional blunting, reduced speech, loss of motivation, self-neglect and social withdrawal (National Institute for Health and Care Excellence (NICE), 2021b). In the United Kingdom (UK) mental disorders are diagnosed based on the presence or absence of signs and symptoms as defined in the Tenth Revision of the International Classification of Diseases (ICD-10) Classification of Mental and Behavioural Disorders. Broadly, ICD-10 divides the psychotic disorders into three groups: idiopathic psychoses, substance induced psychoses and psychoses due to medical conditions (including neurodegenerative disorders). Idiopathic psychoses can be further divided into affective and non-affective psychoses based on their involvement of affect or mood. The archetypal non-affective psychosis is schizophrenia (World Health Organization (WHO), 1992). The peak age of onset of first episode psychosis is between 15 and 30 years (Jones, 2013).

In Scotland psychotic disorders affect between 1-2% of the population and estimates are of 1600 new presentations of psychoses per year (The Scottish Government, 2012, 2019). In Glasgow specifically, the population prevalence of idiopathic nonaffective psychoses was 0.53% between 2002 and 2005 (Srireddy et al., 2012). More granular data on the incidence of psychotic disorders is available in England. Kirkbride et al calculated the pooled incidence of new diagnoses of psychosis (including all idiopathic and substance induced psychoses) as 31.7 per 100000 person years based on data published between 1950 and 2009. The pooled incidence of non-affective psychosis was 23.2, schizophrenia 15.2 and affective psychosis 12.4 per 100000 person years. Rates of psychosis were stable over time with the exception of substance induced psychoses which while still rare, were increasing in incidence (Kirkbride et al., 2012). Unlike specific substance induced psychoses, substance misuse is common in patients diagnosed with psychosis with rates twice that of the general population

(Barnett et al., 2007). In contrast to the lack of change over time in England, Scottish data found that the prevalence of schizophrenia increased by 53.42% between 1981 and 2006 from 2.38/1000 to 3.59/1000 general population in the geographically defined area of Nithsdale in South-West Scotland. Further, in 2006 patients with schizophrenia were more likely to live in the community and to be managed in secondary care than in 1981 (Shivashankar et al., 2013).

Psychotic disorders are associated with considerable morbidity with symptom remission rates of 58% and recovery rates (incorporating both symptomatic and functional improvement for at least 2 years) of just 38% (Lally et al., 2017). In England the total societal cost of the most common psychotic disorder, schizophrenia, was estimated to be greater than £11.8 billion annually (Andrew et al., 2012).

Current Scottish Intercollegiate Guidelines Network (SIGN) and National Institute for Health and Care Excellence (NICE) guidelines recommend all patients with a first episode of psychosis receive treatment within the context of a specialist early intervention service (National Institute for Health and Care Excellence (NICE), 2014; Scottish Intercollegiate Guidelines Network (SIGN), 2013). Diagnosis at initial clinical contact is frequently difficult and it can take months or years before a final diagnosis is made so early intervention is offered to a broad spectrum of idiopathic psychoses (National Institute for Health and Care Excellence (NICE), 2016b). The key components of an early intervention service should be multidisciplinary and include the provision of psychosocial interventions, pharmacological treatment, case management involving smaller caseloads and an assertive approach to treatment (Bird et al., 2010). The core aims of the early intervention service are to reduce duration of untreated psychosis and produce effective outcomes in terms of recovery and relapse (National Institute for Health and Care Excellence (NICE), 2016b). The rationale for early intervention in psychosis is based on the ‘critical period’ hypothesis which states that the early stages of psychosis (the first two to three years) is a ‘critical period’ with major implications for secondary prevention of the impairments and disabilities that accompany psychosis (Birchwood et al., 1998). Compared to treatment as usual, early intervention is associated with better outcomes including reduced hospital admissions, relapse rates and symptom

severity and improved treatment adherence and global functioning (Correll et al., 2018).

Much of the early evidence for the effectiveness of early intervention in psychosis compared to treatment as usual stems from the influential Danish OPUS trial. In OPUS, at 2 years early intervention was superior to treatment as usual in terms of symptoms, comorbid substance use, adherence and satisfaction with treatment (Petersen et al., 2005). However, these differences were no longer significant at 5 years (Bertelsen et al., 2008). Indeed, results from the OPUS trial show that evidence for longer term benefit from early intervention is more ambiguous. While there was a beneficial effect of the use of supported housing and psychiatric hospitalisation at 5 years, this was no longer seen at 10- or 20-years follow-up (Hansen et al., 2023; Secher et al., 2015). The authors suggest that the lack of sustained treatment effect contradicts the 'critical period' hypothesis as it suggests treatment induced improvement in the early years of first episode psychosis does not alter the longer-term trajectory (Hansen et al., 2023). Results are more persuasive when comparing the real-world implementation of early intervention services to the trial conditions in OPUS. Posselt et al found that at 5 years compared to OPUS trial participants, real-world implementation of early intervention in Denmark was associated with fewer and shorter psychiatry admissions, better occupational functioning and more chance of being in a relationship. The treatment as usual group fared worse than both the trial and real-world implementation groups (Posselt et al., 2021).

1.2.2 Pathophysiology

1.2.2.1 Neurotransmission

Psychosis is characterised by altered neurotransmission in the dopamine and glutamate pathways which leads to symptoms of psychosis. Evidence for the role of dopamine stems from the observation that the clinical efficacy of all effective antipsychotics is intrinsically linked to blockade of dopamine D2 receptors (Seeman & Lee, 1975). However, dopamine blockade is not effective for negative and cognitive symptoms. Further, in a significant proportion of patients (those with treatment resistance) it does not improve positive symptoms either.

The finding that *N*-methyl-D-aspartate (NMDA) glutamate receptor antagonists induce psychotic symptoms has led to a magnitude of research implicating the glutamate system in psychosis (McCutcheon et al., 2020).

Positron emission tomography (PET) studies have consistently shown increased presynaptic dopamine function in the striatum (McCutcheon et al., 2020). The pathophysiological relevance of these findings has been supported by a direct association between striatal dopamine synthesis capacity and severity of positive psychotic symptoms (Jauhar et al., 2017). Further, striatal dopamine synthesis capacity correlates with antipsychotic treatment response with higher dopamine synthesis capacity in responders compared to non-responders (Jauhar et al., 2019). In contrast, PET studies have shown little consistent evidence for altered dopamine receptor levels in psychosis (McCutcheon et al., 2020).

Data from magnetic resonance spectroscopy (MRS) implicates excess glutamatergic neurotransmission with elevations in glutamate related metabolites across several brain regions reported on meta-analysis in schizophrenia (Merritt et al., 2016). More recent work points to elevated glutamate levels on MRS being associated with greater illness severity in schizophrenia and that glutamate levels may be reduced by effective antipsychotic treatment (Merritt et al., 2021). However, MRS is not able to distinguish between intra- and extracellular compartments. Further, glutamate does not solely act as a neurotransmitter but is also involved in protein synthesis and nitrogen metabolism, and as a precursor to *γ*-aminobutyric acid (GABA) (McCutcheon et al., 2020).

Taken together, the literature points to excess synaptic levels of dopamine and glutamate that cause increased postsynaptic stimulation with the downstream effect of symptoms of psychosis. Deficiencies in GABA interneurons and hypofunctioning NMDA receptors are thought to represent the molecular basis of these disturbances, altering the inhibitory-excitatory balance of neural systems (Howes et al., 2015; Lieberman & First, 2018).

1.2.2.2 Environmental factors

Psychosis has been viewed as a neurodevelopmental disorder. Evidence which points to environmental factors includes the fact that exposure to prenatal environmental insults (maternal infections, toxins and nutritional deficiencies), birth complication, postnatal trauma and deprivation at critical stages of development is associated with risk of psychotic disorders. These environmental factors and thought to interact with genetics and increase susceptibility to psychosis (Lieberman & First, 2018).

1.2.2.3 Genetics, gene expression and epigenetic factors

Idiopathic psychotic disorders are highly heritable. Data from large population registry studies indicate estimates of the heritability of schizophrenia and schizophrenia spectrum disorders are 79% and 73%.(Hilker et al., 2018) Further, among siblings and parents of people with idiopathic psychotic disorders, rates of the same disorder are 10-15 times as high as the general population. (Lieberman & First, 2018)

Many common genetic variants of low penetrance have been associated with psychotic disorders including schizophrenia. Large genome-wide association studies highlight the *DRD2* gene (coding for dopamine receptor D2, the target of all effective antipsychotic drugs) and many genes (for example, *GRM3, GRIN2A, SRR, GRIA1*) involved in glutamatergic neurotransmission and synaptic plasticity. In addition, research implicate genes involved in immunological function, including the major histocompatibility complex and complement (Ripke et al., 2014; Sekar et al., 2016).

The most common rare genetic variant of high penetrance associated with psychosis is the chromosome 22q11.2 microdeletion with causes DiGeorge syndrome. DiGeorge syndrome occurs in approximately 1 in 4000 live births and is characterised by cardiac, facial and limb abnormalities with 24% affected patients having symptoms indistinguishable from idiopathic schizophrenia (Murphy et al., 1999).

Recent gene ontology enrichment analysis of rare and common genetic variants identified in schizophrenia patients highlight three major functional clusters of

genes related to channel or transporter activities, neuronal components (synapse, axon, and dendrite) and chromatin or histone organisation (Nakamura & Takata, 2023).

Further downstream, gene expression data in post mortem brain tissue of patients with schizophrenia shows particular expression of genes related to inflammatory response and receptor activity, similar to the genetic findings above (Gandal et al., 2018; Nakamura & Takata, 2023). When looking at epigenetic modification post mortem studies show differential methylation of genes related to embryo development, cell fate determination and nervous system differentiation (Jaffe et al., 2016; Nakamura & Takata, 2023).

Although large scale genetic studies have uncovered a considerable amount of the genetic architecture of psychosis including schizophrenia, there remains a considerable gap between the overall heritability reported in epidemiological studies (as above, up to 80%) and that explained by identified common genetic variants of low penetrance (24%) and rare genetic variants of high penetrance (<10%) (Nakamura & Takata, 2023).

1.2.2.4 Animal and cellular models

Nakamura and Takata recently systematically reviewed the literature looking a mouse models based on rare genetic variant of high penetrance (specifically variants with an observed odds ratio of schizophrenia >10). They identified studies looking at mouse models of the following genetic variants: 22q11.2 deletion, 16p11.2 deletion/duplication, 3q29 deletion, 15q11.2–13.1 duplication, 2p16.3 (NRXN1) deletion, GRIN2A loss of function variant (LOF), GRIA3 LOF, and SETD1A LOF. The commonly dysregulated molecular pathways identified in these mouse models included neural transmission and regulation of transcription, similar to those highlighted in human genetic studies. Morphological analysis of the neuronal cells in the models showed reduced axonal and dendritic complexity together with abnormal spine morphology. Common electrophysiological phenotypes included altered synaptic transmission and deficits in long term potentiation. When considering mouse behavioural readouts, deficits in sociability, cognitive performance and prepulse inhibition

were identified, in line with those found in human patients with schizophrenia (Nakamura & Takata, 2023).

Nakamura and Takata also reviewed cellular models of the above identified genetic variants. Cellular models enable the reproduction of pathological conditions *in vitro* by creating patient-derived or mutation-carrying induced pluripotent stem cells and then differentiating them into central nervous system cells or miniature brains. In line with the mouse models, cellular models showed dysregulation of genes related to neural transmission and transcriptional regulators. Further, there were commonly morphological alterations of the cell soma and dendrites. The cellular models also suggested that imbalanced excitatory and inhibitory neuronal activity is important to the pathophysiology of schizophrenia (Nakamura & Takata, 2023).

1.2.2.5 Inflammation

Increasingly compelling data point to an association between inflammation and psychosis. Autoimmune disorders are associated with higher rates of psychosis (Bergink et al., 2014). As discussed above, the strongest genetic relationship in schizophrenia is with the major histocompatibility complex, a region with a key role in immunity (Ripke et al., 2014). The association between maternal infection and schizophrenia is supported by the animal maternal immune activation model of psychotic-like behaviour (Brown & Patterson, 2011). Circulating inflammatory biomarkers are associated with psychosis including raised neutrophil/lymphocyte ratio (NLR) and monocyte/lymphocyte ratio (MLR), abnormalities in circulating proinflammatory cytokines and their receptors, reduced nerve growth factor, raised S100 calcium binding protein and raised C-reactive protein (Mazza et al., 2020; Yuan et al., 2019). Inflammation is linked to increased glutamate expression in the brain via the upregulation of the enzyme indoleamine 2, 3-dioxygenase (IDO). Quinolinic acid is a downstream product which is glutamatergic (Schwarcz et al., 2012).

1.2.3 Risk factors for psychosis from epidemiology

Epidemiological studies show that psychotic disorders are more common in the young with peak incidence in people in their 20s, in males, in racial or ethnic

minorities especially black people, in urban areas and in areas of with higher levels of deprivation as indicated by lower levels of owner-occupancy (Jongsma et al., 2018; Kirkbride et al., 2012). Further, an epidemiological link between cannabis and psychosis was first identified more than 35 years ago (Andreasson et al., 1987), with findings replicated in more recent studies (van Os et al., 2002; Zammit et al., 2002).

Locally in Glasgow, epidemiological work found that psychosis was most prevalent amongst black and minority ethnic groups. The prevalence of psychosis was higher in deprived areas in white but not black and minority ethnic populations. The authors postulate that belonging to an ethnic minority group may be a risk factor for being diagnosed with psychosis independent of socioeconomic deprivation. Among the white population, psychosis was also twice as likely to be diagnosed in males compared with females (Srireddy et al., 2012).

1.2.4 Prognostic factors in psychosis

1.2.4.1 Definitions

The Remission in Schizophrenia Working Group defined symptom remission as scores of less than or equal to three in Positive And Negative Syndrome Scale (PANSS) items P1 Delusions, P2 Conceptual Disorganization, P3 Hallucinatory Behavior, N1 Blunted Affect, N4 Apathetic Social Withdrawal, N6 Lack of Spontaneity and G9 Unusual Thought Content, present for a period of at least 6 months (Andreasen et al., 2005).

Treatment resistance is defined as failing to respond to two standard antipsychotic medications of adequate dose and duration.

The concept of recovery from psychosis is less consistently defined but Lally et al's definition includes both symptomatic and functional improvement for at least 2 years (Lally et al., 2017). Functional recovery specifically is the treatment outcome most valued by patients (Iyer et al., 2011).

1.2.4.2 Psychosis overall

Systematic review and meta-analysis of observational studies provide evidence for prognostic factors for outcomes in people diagnosed with psychosis. Duration of untreated psychosis is the most replicated predictor of outcome in psychosis. Several meta-analyses have linked longer duration of untreated psychosis poor outcomes in psychosis including with lower rates of symptom remission and poorer functional recovery (Diana O. Perkins et al., 2005; Farooq et al., 2009; Howes et al., 2021; Marshall et al., 2005; Penttilä et al., 2014). Reducing duration of untreated psychosis is the key rationale behind early intervention services for psychosis and its principal aim (National Institute for Health and Care Excellence (NICE), 2016b).

1.2.4.3 First episode psychosis

The literature on prognostic factors for first episode psychosis is inconsistent. For example, while Lally's 2017 meta-analysis provided the first robust evidence of remission and recovery outcomes in first episode psychosis at 58% and 38% respectively, it was unable to establish any key clinical or demographic factors which discriminated between patients. In particular, it did not replicate earlier findings of longer duration of untreated psychosis and worse outcomes in first episode psychosis (Lally et al., 2017). Another more recent 2021 meta-analysis by Catalan et al established very similar rates for remission and recovery in first episode psychosis, at 54% and 32%. However, yet again this study failed to establish any predictors that had a significant effect on remission or recovery. Recovery was associated with male sex and positive symptoms, but these associations did not survive multiple comparison corrections. As with Lally et al, Catalan et al did not find any relationship between duration of untreated psychosis and outcome in first episode psychosis (Catalan et al., 2021).

A comprehensive 2017 meta-analysed identified several factors that were associated with better long-term functioning in first episode psychosis. These included better cognitive functioning, female sex, education, work history, lower positive, negative and joint symptoms at baseline, premorbid adjustment, shorter duration of untreated psychosis and duration of illness and remission of

positive, negative and joint symptoms. The strongest association was with symptom remission (Santesteban-Echarri et al., 2017).

Finally, when considering prognostic factors for treatment resistance in the early stages of first episode psychosis, male sex was associated with the outcome (Siskind et al., 2022).

1.2.4.4 Early-onset psychosis

When restricting the population to those with early-onset psychosis (onset before the age of 18) findings are more consistent. A 2015 systematic review by Díaz-Caneja et al identified history of developmental delay, poor premorbid adjustment, greater symptom severity at baseline (especially negative symptoms) and longer duration of untreated psychosis as the most replicated predictors of poor symptomatic, functional and cognitive outcomes. Diagnosis of schizophrenia was a significant predictor of greater disability, worse global functioning and poorer quality of life at follow-up. This study was one of the few systematic reviews to consider the evidence for biological predictors of outcome in psychosis. The review highlighted preliminary evidence for neuroimaging markers including regional cortical thickness and grey matter volume at baseline which predicted remission (Díaz-Caneja et al., 2015). However, in general, evidence for putative biological predictors of outcome in psychosis, such as those implicated in the pathophysiology as discussed above, has not been consistently identified at systematic review.

1.2.4.5 Schizophrenia

Looking at prognostic factors for treatment resistance in schizophrenia specifically, systematic review highlighted evidence for younger age of onset, schizophrenia diagnosis, level of functioning, male gender and autumn/winter season of birth (Smart et al., 2021).

1.2.5 Prediction of outcome in psychosis

As discussed above, outcomes in first episode psychosis are heterogeneous with remission rates of 58% and recovery rates of 38% (Lally et al., 2017). In the early stages of treatment for first episode psychosis, 23% of patients develop

treatment resistant schizophrenia (Siskind et al., 2022). Although effective interventions exist to combat nonremission, inadequate recovery and treatment resistance, there is often a delay in providing these interventions. For example, clozapine is more effective than other antipsychotics for alleviating symptoms in treatment resistance schizophrenia and is the recommended therapy. However, evidence suggests that there is an average delay of nearly four years in initiating clozapine (Howes et al., 2012). This is in spite of the fact that according to current treatment guidelines, treatment resistance status could be diagnosed in as little as 12 weeks (National Institute for Health and Care Excellence (NICE), 2014). Longer delays before clozapine initiation have been shown to result in a worse symptomatic response (Yoshimura et al., 2017). Clinicians have identified the difficulties in early identification of patients who are likely to become treatment resistant as a barrier to preventing the initiation of effective phase-specific treatments like clozapine at the optimal time (Farooq et al., 2009).

Yet, despite the growing body of evidence for prognostic factors from systematic review and meta-analysis of observational studies as discussed above, at present clinicians still struggle to predict outcome in individuals with psychosis. Difficulties are compounded by the fact that group level differences identified in observational studies cannot be readily extrapolated to individuals – the ecological fallacy (Sedgwick, 2015). Further, insights drawn from explanatory research do not necessarily equate to accurate predictions (Shmueli, 2010).

1.3 Prognostic model research

1.3.1 Overview

A prognostic prediction model combines two or more prognostic risk factors (also known as predictors, features or independent variables) into an algorithm which is used to predict the probability of a future event (the dependent variable or outcome). An algorithm is a sequence of statistical, mathematical or programmatic rules usually conducted by a computer to achieve a goal (e.g. predicting disease) (Dwyer & Krishnadas, 2022). As outlined above, prognostic risk factors range from demographic characteristics and clinical features to biological disease markers like imaging, genetics, blood or tissue measurements.

Prognostic model research, the third PROGRESS theme, has the potential to revolutionise medicine by the prediction of *individual* patient outcome. Early identification of patients at a higher risk of poor outcome in first episode psychosis via a prognostic model could help facilitate personalised interventions according to their individual profile of prognostic factors. This is the basis of stratified medicine, the fourth PROGRESS theme. Stratified medicine is contrasted with the current approach in most areas of medicine, including psychiatry, of "all comer" or "empirical" medicine (Hingorani et al., 2013). A stratified medicine approach has had some initial successes. Specific examples include the Predict tool which is recommended to guide adjuvant therapy in individuals with invasive breast cancer (National Institute for Health and Care Excellence (NICE), 2018; Wishart et al., 2010), and, the QRISK tool which is recommended to guide lipid lowering treatment based on an individual's cardiovascular risk (Hippisley-Cox et al., 2017; National Institute for Health and Care Excellence (NICE), 2016a). These prognostic models aim to assist, not replace, clinicians with their prediction of a patient's future outcome in order to enhance informed decision making together with the patient (Steyerberg et al., 2013). However, such examples of gold standard practice in prognostic model research are very much the exception to the rule and the vast majority of models are never applied into clinical practice (Riley et al., 2013).

1.3.2 Prediction modelling versus explanatory modelling

Prognostic model research relies on prediction modelling. Prediction modelling is defined as the process of applying a statistical model or data mining algorithm to data for the purpose of predicting outcomes in individuals. The focus is on the predictive performance of the model on new (unseen) data, i.e., model generalisability. Theoretical models are not necessary and causal interpretation is not typically of interest. Model transparency is often of secondary importance and so the range of plausible models includes not only regression models but also machine learning algorithms. Indeed, many models are often criticised as being like a "black box" as they are only viewed in terms of the inputs and outputs without any knowledge of the internal workings. Prediction modelling is contrasted with explanatory modelling which refers to the application of statistical models to data for testing hypotheses. In explanatory research we are interested in an average group response of a population. The focus is on how

well the explanatory model fits existing data. Explanatory modelling requires interpretable statistical models that are easily linked to the underlying theoretical model. This explains the popularity of statistical models, especially regression type models. Machine learning methods like artificial neural networks, which are less amenable to interpretation, may be considered ill-suited for explanatory modelling (Shmueli, 2010).

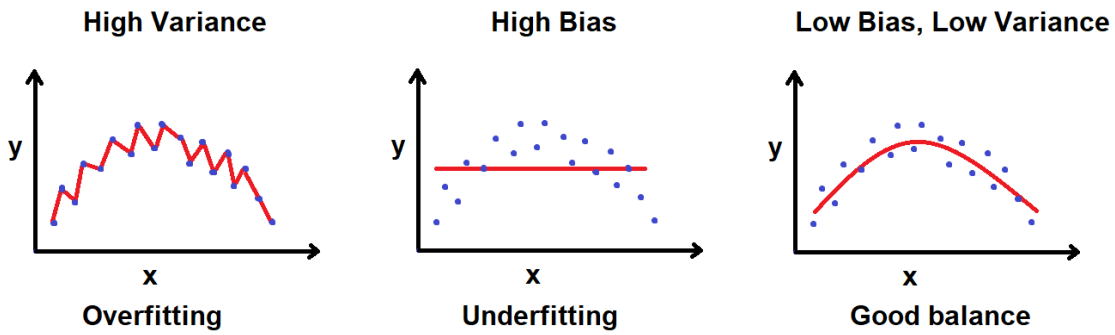


Figure 1-2 The trade-off between bias and variance.

Prediction modelling involves a trade-off between bias and variance to optimise out of sample performance (Figure 1-2). In explanatory modelling, however, the key criteria for model selection is to minimise bias on existing data. Model variance is the spread of our predictions. High variance results from overfitting training data and leads to a lack of generalisability to new data. Variance can also be conceptualised as the difference in fits between datasets. Bias is the difference between the prediction of our model and the correct value. Models with a high bias pay little attention to the training data and underfit. This results in high error on training and test dataset (James et al., 2021).

Prediction modelling can be used for both prognostic prediction models which estimate the probability that a future outcome will occur and also for diagnostic prediction models which estimate the probability that a certain outcome is present (Moons et al., 2019). This thesis is concerned with the former. In subsequent chapters where “prediction models” are discussed, it is always prognostic prediction models that are being referred to.

1.3.3 Machine learning versus regression for prediction

Machine learning is operationally defined as “models that directly and automatically learn from data”(Christodoulou et al., 2019). This is in contrast to regression models which “are based on theory and assumptions, and benefit from human intervention and subject knowledge for model specification” (Christodoulou et al., 2019). Clinical prediction models more often involve regression modelling techniques (for example, logistic regression with or without regularisation) despite the availability of machine learning techniques (for example, artificial neural networks, random forest or support vector machines) (Moons et al., 2019). For example, Meehan *et al* reviewed all the published evidence for prediction models in psychiatry and found that the majority used regression-based methods as opposed to machine learning (Meehan et al., 2022).

Machine learning algorithms can improve the accuracy of prediction over conventional regression models by capturing complex nonlinear relationships in the data (Chen & Asch, 2017). Nonlinear interactions can, however, be modelled in regression using regression splines. There are specific issues associated with the use of machine learning for prediction in a clinical setting. Despite not being the primary goal in prediction, a lack of transparency for a model can adversely affect applicability and usability especially in a clinical setting. Machine learning models are often criticised as being a “black box” which offer no explanation for their decisions. For example, they often lack a clear estimate of the importance of different features or how they interact to predict the outcome. This is contrasted with regression models which are easier to explain and interpret but as a consequence are not always capable of modelling the inherent complexities in the data. Further, there are potential ethical implications; if doctors cannot understand and explain why an algorithm made a decision, how can patients give informed consent for the proposed management (Watson et al., 2019)?

There are additional issues in clinical settings with rarer conditions. Machine learning models can be more flexible but as a consequence tend to overfit data when sample sizes are small and data is sparse (Moons et al., 2019). Prediction models developed using machine learning techniques have been demonstrated to require substantially higher events per variable (often >200) to mitigate overfitting and optimism in model performance (van der Ploeg et al., 2014).

Regression models can be the most sensible choice in datasets with lower events per variable. In contrast, machine learning performs best for problems with a high signal-to-noise ratio with higher events per variable (Christodoulou et al., 2019).

In spite of the above, the distinction between regression models and machine learning has been viewed to be artificial. Indeed, many consider logistic regression as a type of machine learning model. Alternatively, algorithms may be considered to exist "along a continuum between fully human-guided to fully machine-guided data analysis" (Beam & Kohane, 2018).

1.3.4 Further ethical considerations

There are additional ethical considerations which are common to both regression and machine learning prediction models. For example, prediction models may disempower patients by undermining their sense of agency (the belief that one can shape one's own life). This may result if patients misinterpret prediction to mean that the outcomes they predict (for example, relapse or recovery) are predetermined and beyond their ability to influence. As a consequence, patients may disengage with their care and treatment or place limits on their goals or ambitions. Moreover, if third parties like employers or insurance companies gain access to negative predictions patients may suffer discrimination. Finally, prediction models may exacerbate existing inequalities in healthcare. For example, if a prediction model is trained on a predominantly white population (as is typical of research cohorts in the developed countries), it will be less accurate in individuals from ethnic minorities and disadvantage these groups (Lane & Broome, 2022).

1.3.5 Stages of prognostic model development

Prognostic model research proceeds through three main stages: model development (including internal validation), external validation and assessment of impact in clinical practice (Figure 1-3) (Steyerberg et al., 2013). A development study typically involves the identification of important predictors, assigning relative weights to each predictor and assessing the model's performance adjusted for overfitting by an internal validation procedure such as

cross-validation or bootstrapping. An external validation study tests the model's predictive performance in new subjects from different sites (geographical validation), time points (temporal validation) or clinical context (for example, from primary to secondary care). Finally, an impact study assesses whether the use of a prognostic model in daily practice improves clinical decision making and patient outcomes. An impact study is ideally performed as a cluster-based randomised control trial with centres randomised to care with or without the benefit of a prognostic model. Only a clinical impact study can assess whether use of the model is better than usual care. Impact studies can also be used to assess other factors which affect implementation such as acceptability of the model to clinicians and ease of use (Moons et al., 2009; Steyerberg et al., 2013). Further, it needs to be determined whether an assistive or directive approach is adopted when presenting a model to clinicians. In an assistive approach predictions are simply presented as numerical probabilities without corresponding decision recommendations, while a directive approach presents decision recommendations with or without corresponding numerical probabilities (Kappen et al., 2018; Moons et al., 2009). Prognostic models only influence patient outcome when changes in clinical management result from their predictions. It is also important to remember that prognostic models have a cost in their implementation and could even have adverse consequences on clinical outcomes if their use leads to a clinician withholding potential beneficial treatment (for example, from a patient whom the model estimates to have low risk) (Steyerberg et al., 2013).

Figure 1-3 Steps required in order to translate a prognostic prediction model into clinical practice.

Most publications on prognostic models describe model development, a minority report external validation and only very few consider clinical impact. Meehan et

al recently reviewed all published prediction models in psychiatry (both prognostic and diagnostic models in this case). They found that of the 308 prediction models published only 20% had undergone any form of external validation and only one model had undergone a preliminary assessment of clinical implementation (Meehan et al., 2022). Similar findings have been reported for prognostic models across all of clinical medicine (Steyerberg et al., 2013).

1.3.6 Methodological issues in the field

Despite the existence of clear guidance for the reporting of prediction models, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement (Collins et al., 2015), and for the assessment of bias in prediction modelling studies, Prediction model Risk Of Bias ASsessment Tool (PROBAST) (Wolff et al., 2019), methodological issues in prognostic modelling studies are common. In psychiatry, Meenan et al identified risk of bias in 94.5% of model development analyses and in 68.6% of external validation analyses. Risk of bias occurs when shortcomings in study design, conduct or analysis lead to systematically distorted estimates of a model's predictive performance. The most common concern was inappropriately low numbers of events per variable with only 26.9% of studies meeting the widely adopted benchmark of ≥10. This highlights the importance of sample size. Predictor selection was often problematic too with only 8.6% of studies adopting the recommended practice to minimise risk of bias by selecting candidate predictors based on existing evidence and expert knowledge. Instead, most studies adopted data-driven methods of variable selection. Moreover, key prediction performance metrics were poorly or inconsistently reported. Steyerberg et al outline four key measures of predictive performance that should be assessed in any prediction-modelling study: two measures of calibration (the calibration-in-the-large Alpha (A) and the calibration slope Beta (B)), discrimination via a Concordance statistic (C) and clinical usefulness with Decision-curve analysis (D) (Steyerberg & Vergouwe, 2014). Model calibration is the level of agreement between the observed outcomes and the predictions. Discrimination is the ability of a model to distinguish between a patient with the outcome and one without. Yet, Meehan et al found that while 88% of models reported discrimination, just 22.1% of models assessed calibration and only two

models considered clinical usefulness. Finally, only 11.4% of development studies were judged to have performed internal validation to a sufficient statistical standard. The main issue was the use of random split sampling instead of the recommend cross-validation or bootstrapping. However, even when this was performed, only 23% appropriately nested feature selection and tuning procedures (Meehan et al., 2022). While not specifically raised by Meehan et al, inappropriate handling of missing data (which should ideally be via multiple imputation) is another methodological issue common across the field (Moons et al., 2019; Steyerberg et al., 2013). However, despite these shortcomings there remains great potential for clinical benefit from prognostic model research and stratified medicine in psychiatry and across medicine.

1.4 Routinely collected clinical data for prognostic research

In recent years, prognosis research has benefitted from the exponential growth in the availability of routinely collected health data. Its use is a key recommendation from PROGRESS (Hemingway et al., 2013). Routinely collected data are data collected without a specific *a priori* research question. Sources of routinely collected data include disease registries and electronic healthcare records (Benchimol et al., 2015). It is extremely valuable for prognosis research because it allows pragmatic cost-effective research to be conducted in an entirely naturalistic clinical setting, with much larger numbers of participants. Linkage of electronic health records across different sources enables the possibility of examining the patient journey with repeated measures of care in larger populations than would be feasible in traditional observational studies (Hemingway et al., 2013). In addition, routinely collected data enables the researcher to circumvent any issues with selection bias.

A principal aim of my PhD fellowship was to be the establishment of the Electronic Measures in Psychosis – Assessing Trajectory and Health-Outcomes (EMPATH) platform for the collection of outcome data for first episode psychosis patients treated withing Esteem National Health Service (NHS) Greater Glasgow & Clyde (GG&C) early intervention in psychosis service. EMPATH aims to operationalise the routine collection of standard outcome measures within the service.

Data collected within EMPATH will be securely deposited within West of Scotland Safe Haven which allows research use of linked unconsented routinely collected datasets. Safe Haven has ethical approval (17/WS/0237) to create a research database using routinely collected, un-consented patient data. Safe Haven provide a secure environment for hosting and analysing routinely collected patient data from NHS GG&C. Researchers benefit from accelerated ethical approval via the Safe Haven Local Privacy Advisory Committee. Safe Haven have access to many large datasets of NHS encounters in secondary care, demographics and death data for Glasgow populations, prescribing information, and many more specialised thematic datasets drawn from various NHS and patient research sources (NHS Greater Glasgow & Clyde, 2023).

I had planned to use EMPATH as a source of routinely collected data to undertake prognosis research into patients with a first episode of psychosis as part of my PhD. Unfortunately, there have been significant delays in developing and deploying the EMPATH platform as a consequence of the global coronavirus pandemic such that deployment was postponed until 2023. Routinely collected data should be available for prognosis research from 2024.

Given this, in order to establish the feasibility of using routinely collected data from NHS GG&C for prognosis research using Safe Haven, in advance of accessing routinely collected data from EMPATH, I looked at another clinical area with unanswered prognosis questions: delirium and the risk of subsequent dementia. This work falls under the first PROGRESS theme, fundamental prognosis research, and second PROGRESS theme, prognostic factor research. The data for this study came from the West of Scotland Safe Haven.

1.5 Delirium and the risk of subsequent dementia

With an aging population in industrialised countries, cognitive impairment is an increasingly frequent problem. Delirium and dementia are among the leading causes of cognitive impairment (Fong et al., 2015). Dementia is characterised by an irreversible progressive global cognitive decline. Delirium is characterised by an acute and fluctuating disturbance in attention and awareness with associated disturbance in cognition (for example, memory deficit, disorientation, language, visuospatial ability or perception), which cannot be explained by another

neurocognitive disorder and does not occur in the context of a severely reduced level of arousal, such as coma. It is a serious and life-threatening neuropsychiatric syndrome, which is a direct physiological consequence of another medical condition, substance intoxication or withdrawal, toxins or multiple aetiologies (Slooter et al., 2020).

Delirium affects as much as 50% of those over 65s in hospital but is less common in the community with a prevalence of 1-2% (Inouye et al., 2014). In comparison, the prevalence of dementia is 7.1% in the over 65s (Prince et al., 2014). The rates are particularly high in hospital settings with 1 in 4 UK hospital beds occupied by people with dementia at any one time (Royal College of Psychiatrists, 2019).

Delirium and dementia are clinical diagnoses which can be distinguished by several features. In delirium the onset is typically abrupt over hours to days, whereas dementia is insidious and progressive over months to years. In delirium, attention and consciousness are reduced and fluctuate, while in dementia these cognitive domains remain intact until the advanced stages (Fong et al., 2015).

There are a number of possible pathways linking delirium to subsequent dementia. Delirium may represent an epiphenomenon which simply exposes pre-existing cognitive impairment. The effect of delirium on dementia may be related to its precipitating factors. Alternatively, delirium itself may cause permanent neuronal damage and precipitate dementia. If delirium causes dementia this has important implications. Delirium is estimated to be preventable in 30 to 40% of cases (Inouye et al., 1999; Marcantonio et al., 2001).This suggests that many cases of dementia itself could therefore also be prevented.

Evidence linking delirium to subsequent cognitive impairment and dementia has emerged from a number of meta-analyses of observational case-control studies. A 2010 meta-analysis of two studies and 241 patients aged 65 or older by Witlox et al found that delirium is associated with an increased risk of dementia independent of age, sex, comorbid illness and illness severity with an odds ratio of 12.52 (95% CI 1.86 to 84.21) (Witlox et al., 2010). A 2021 meta-analysis of six studies and 901 patients by Pereira et al looked specifically at the relationship

between delirium and new dementia in inpatients aged 65 or older. They showed inpatients who developed delirium had 11.9 times the odds (95% CI 7.29 to 19.6) of subsequent dementia (Pereira et al., 2021). A much larger 2020 meta-analysis of 24 studies and 10549 patients by Goldberg et al showed that delirium was associated with 2.3 times the odds (95% CI 1.85 to 2.86) of cognitive decline including dementia compared to controls. This study also tried to disentangle whether delirium simply unmasked pre-existing cognitive decline or whether it was causative by performing a series of sensitivity analyses. First, they included only patients with no baseline cognitive impairment and tested the hypothesis that neither group should show cognitive decline if delirium was not causative. However, they showed that the delirium group had worse cognitive decline. Second, they assessed whether cognitive decline was unrelated to delirium by explicitly examining patients with cognitive decline, hypothesising that the delirium and non-delirium groups should decline equally if delirium was simply an epiphenomenon. However, the delirium group experienced greater decline. Finally, they compared delirious and non-delirious patients from studies that matched for baseline cognition and from studies that did not match for baseline cognition. They showed that effect sizes were larger in those studies that matched for baseline cognition suggesting that baseline cognitive compromises were not a major driver of the effect of delirium. Taken together, these sensitivity analyses indicated that delirium was causative (Goldberg et al., 2020).

Evidence that delirium is a strong risk factor for dementia also comes from a well-designed longitudinal study of 553 people over the age of 85. The Vantaa 85+ study showed that delirium increased the risk of incident dementia with an odds ratio of 8.7 (95% CI 2.1 to 35). Intriguingly, this study also showed that compared to those who developed dementia without delirium, those that developed dementia after delirium had different pathological changes from those normally associated with dementia (for example, as in Alzheimer’s, vascular or Lewy body dementia). This suggests that the acceleration of cognitive decline after delirium might result from alternative mechanisms of neuronal damage (Davis et al., 2012).

Altogether, increasing evidence suggests that delirium causes cognitive decline. However, leading authors in the field suggest that long-term studies of patients with delirium initially free of dementia are required to help clarify whether incident delirium leads to new-onset dementia (Fong et al., 2015; Inouye et al., 2014).

1.6 Summary

In this chapter, I have provided an overview of prognostic research with a focus on prognostic model research. I explored psychosis as a clinical problem. Despite growing evidence for prognostic factors, the prediction of individual patient outcomes in first episode psychosis is still a challenge. Finally, I outlined another clinical area with unanswered prognostic questions – delirium and the risk of subsequent dementia.

1.7 Thesis aims and outline

The primary aim for this thesis is to conduct prognostic model research into first episode psychosis. As part of my primary aim, I will attempt to answer four questions:

1) Is prediction of individual patient outcome possible in first episode psychosis using clinical variables?

2) Does prediction model performance remain robust at external validation?

3) Does prediction model performance improve with the addition of biologically relevant disease markers as predictors?

4) Does prediction model performance improve with the application of advanced machine learning classifiers compare to logistic regression?

Chapter 2 examines questions 1 to 4 in the context of a systematic review of existing literature. Chapter 3 explores questions 1 and 2 in two large naturalistic cohorts of first episode psychosis patients from NHS England. Chapter 4 examines question 3 and 4 in a cohort of first episode psychosis patients from a randomised controlled trial.

The secondary aim for my thesis is to answer a final question:

5) Can routinely collected electronic healthcare record data be used for prognostic research in the National Health Service in Greater Glasgow and Clyde?

As a consequence of the global coronavirus pandemic, the prognostic area I addressed changed to delirium and the risk of subsequent dementia. Chapter 5 seeks to answer this in a large cohort of patients with an incident episode of delirium derived from electronic healthcare records.

Chapter 2 Prediction models in first episode psychosis: a systematic review and critical appraisal

2.1 Overview of this chapter

This chapter presents a systematic review of prediction models in first episode psychosis. The review was published in the British Journal of Psychiatry on 24th January 2022 and is presented as published in this thesis. I am a joint first author with Rebecca Lee, University of Birmingham. We both collected and analysed the data presented herein. Rebecca Lee contributed to an earlier draft of this chapter which I substantially changed and redrafted.

2.2 Introduction

Psychosis is a mental illness characterised by hallucinations, delusions and thought disorder. The median lifetime prevalence of psychosis is around eight per 1000 of the global population (Moreno-Kustner et al., 2018). Psychotic disorders, including schizophrenia, are in the top 20 leading causes of disability worldwide (Institute for Health Metrics and Evaluation (IHME), 2020). People with psychosis have heterogeneous outcomes. More than 40% fail to achieve symptomatic remission (Lally et al., 2017). At present, clinicians struggle to predict long term outcome in individuals with first episode psychosis (FEP).

Prediction modelling has the potential to revolutionise medicine by predicting individual patient outcome (Darcy et al., 2016). Early identification of those with good and poor outcomes would allow for a more personalised approach to care, matching interventions and resources to those most at need. This is the basis of precision medicine. Risk prediction models have been successfully employed clinically in many areas of medicine; for example, the QRISK tool predicts cardiovascular risk in individual patients (Hippisley-Cox et al., 2017). However, within psychiatry, precision medicine is not yet established within clinical practice. In FEP, precision medicine could enable rapid stratification and targeted intervention thereby decreasing patient suffering and limiting treatment associated risks such as medication side effects and intrusive monitoring.

Salazar de Pablo et al recently undertook a broad systematic review of individualised prediction models in psychiatry. They found clear evidence that precision psychiatry has developed into an important area of research, with the greatest number of prediction models focussing on outcomes in psychosis. However, the field is hindered by methodological flaws, for example lack of validation. Further, there is a translation gap with only one study considering implementation into clinical practice. Systematic guidance for the development, validation and presentation of prediction models is available (Steyerberg & Vergouwe, 2014). Further, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement sets standards for reporting (Collins et al., 2015). Models that do not adhere to these guidelines result in unreliable predictions, which may cause more harm than good in guiding clinical decisions (Wynants et al., 2020). Salazar de Pablo et al 's review was impressive in scope but necessarily limited in detailed analysis of the specific models included (Salazar de Pablo et al., 2021). Systematic reviews focussing on the predicting the transition to psychosis (Rosen et al., 2021; Studerus et al., 2017), and predicting relapse in psychosis have also been published (Sullivan et al., 2017). In our present review, we focus on FEP with the aim to systematically review and critically appraise the prediction models for the prediction of poor outcomes.

2.3 Methods

We designed this systematic review in accordance with the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) (Moons et al., 2014). A protocol for this study was published on the International Prospective Register of Systematic Reviews (PROSPERO), registration number CRD42019156897.

We developed the eligibility criteria under the Population, Index, Comparator, Outcome, Timing and Setting (PICOTS) guidance (see Table 2-1). A study was eligible for inclusion if it utilised a prospective design, including patients diagnosed with FEP, and developed, updated, or validated prognostic prediction models for any possible outcome, in any setting. We excluded non-English language studies, those where the full text was not available, those involving diagnostic prediction models, and those where the outcome predicted was less

than or equal to 3 months from baseline because we were interested in longer term prediction.

Table 2-1 Population, Index, Comparator, Outcome, Timing and Setting (PICOTS)

Population	Patients with a first episode of psychosis
Intervention (model)	Any prognostic prediction model
Comparator	N/A
Outcome(s)	Any outcome
Timing	Greater than 3 months from baseline
Setting	Any setting

We searched PubMed, PsychINFO, EMBASE, CINAHL Plus, Web of Science Core Collection and Google Scholar from inception up to 28th January 2021. In addition, we manually checked references cited in the systematically searched articles. The search terms were based around three themes – 'Prediction', 'Outcome' and 'First Episode Psychosis' terms. The full search strategy is available in Appendix 1. Two reviewers (RL and LT) independently screened the titles and abstracts. Full text screening was completed by three independent reviewers (RL, PM and SPL). Disagreements were resolved by consensus.

Data extraction was conducted independently by two reviewers (RL and SPL) following recommendations in the CHARMS checklist (Moons et al., 2014). From all eligible studies, we collected information on study characteristics, methodology and performance. Study characteristics collected included first author name, year, region, whether multicentre, study type, setting, participant description, outcome, outcome timing, predictor categories and number of models presented. Methodology considered sample size, events per variable (EPV), number of events in validation dataset, number of candidate and retained predictors, methods of variable selection, presence and handling of missing data, modelling strategies, shrinkage, validation strategies (see below), whether models were recalibrated, if clinical utility was assessed and whether the full models were presented. Steyerberg and Harrell outline a hierarchy of validation strategies from apparent (which assesses model performance on the data used to develop it and will be severely optimistic), to internal (via cross validation or bootstrapping), internal-external (for example, validation across centres in the same study) and external validation (to assess if models generalise to related

populations in different settings) (Steyerberg & Harrell, 2016). Apparent, internal and internal-external validation use the derivation dataset only, while external validation requires the addition of a validation dataset. Performance for the best performing model per outcome in each article was considered by model validation strategy, including model discrimination (reported as the c-statistic which is equal to the area under the receiver operating characteristic (ROCAUC) curve for binary outcomes), calibration, other global performance measures, and classification metrics. If not reported, where possible, the balanced accuracy (sensitivity + specificity / 2) and the prognostic summary index (positive + negative predictive value - 1) were calculated.

Two reviewers (RL and SPL) independently assessed the risk of bias (ROB) in included studies using the Prediction model Risk Of Bias Assessment Tool (PROBAST), a risk of bias assessment tool designed for systematic reviews of diagnostic or prognostic prediction models (Moons et al., 2019; Wolff et al., 2019). We considered all models reported in each article and assigned to the article an overall rating. PROBAST uses a structured approach with signalling questions across four domains: ‘participants’, ‘predictors’, ‘outcome’ and ‘statistical analysis’. Signalling questions are answered ‘yes’, ‘probably yes’, ‘no’, ‘probably no’ or ‘no information’. Answering ‘yes’ indicates a low ROB, while ‘no’ indicates high ROB. A domain where all signalling questions are answered as ‘yes’ or ‘probably yes’ indicates low ROB. Answering ‘no’ or ‘probably no’ flags the potential for the presence of bias and reviewers should use their personal judgement to determine whether issues identified have introduced bias. Applicability of included studies to the review question is also considered in PROBAST.

We reported our results according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021).

2.4 Results

Systematic review of the literature yielded 2353 records from database searches and 67 from additional sources. After removal of duplicates, 1543 records were screened. Of these, 82 full texts were reviewed, which resulted in 13 studies

meeting criteria for inclusion in our qualitative synthesis (Figure 2-1) (Ajnakina et al., 2020; Bhattacharyya et al., 2021; Chua et al., 2019; de Nijs, 2019; Demjaha et al., 2017; Derks et al., 2010; Flyckt et al., 2006; Gonzalez-Blanch et al., 2010; Koutsouleris et al., 2016; Leighton, Krishnadas, et al., 2019; Leighton et al., 2021; Leighton, Upthegrove, et al., 2019; Puntis et al., 2021).

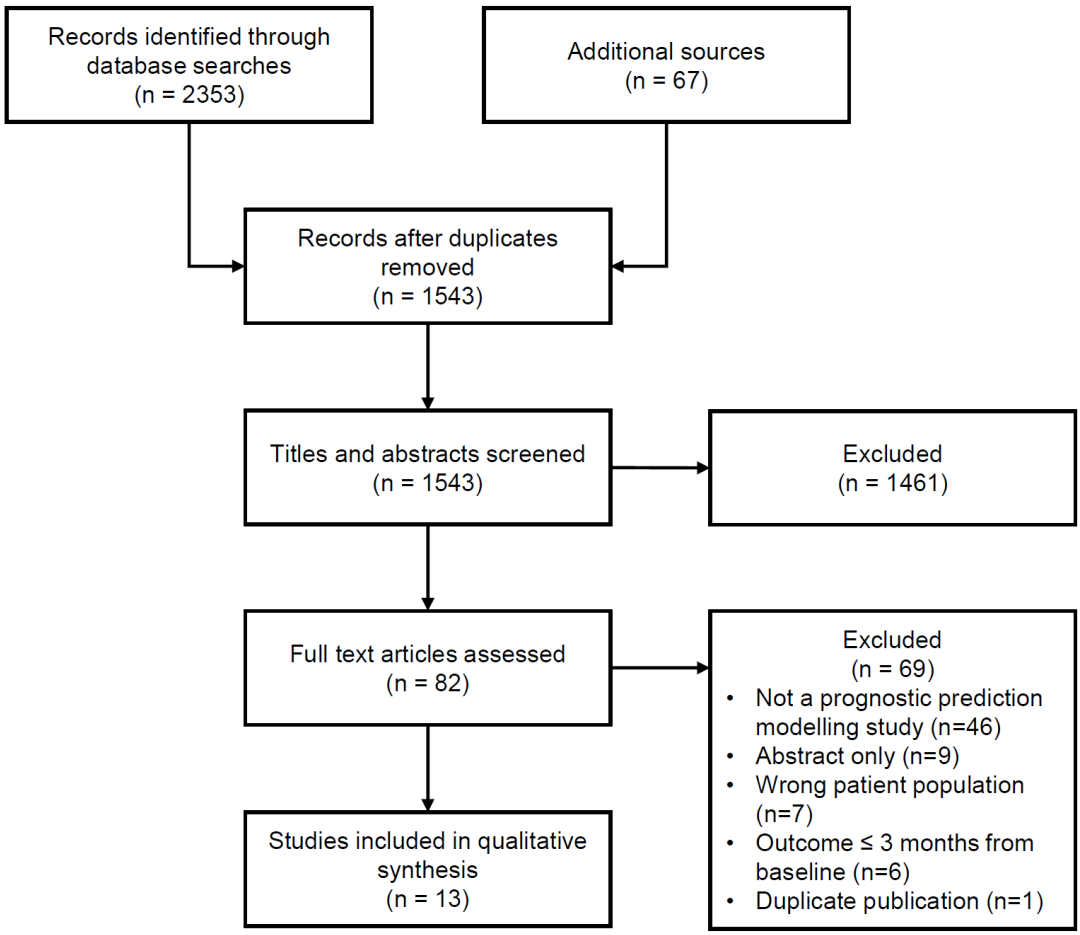


Figure 2-1 Prisma flow diagram

Study characteristics are summarised in Table 2-2. The 13 included studies, comprising a total of 19 different patient cohorts, reported 31 different prediction models. Dates of publication ranged from 2006 and 2021. Twelve studies (92%) recruited participants from Europe, with two studies (15%) also recruiting participants from Israel and one study (8%) from Singapore. Over two-thirds (n=9) of studies were multicentre. Ten studies (77%) included participants from cohort studies, three studies (23%) included participants from randomised controlled trials and two studies (15%) included participants from case registries. Two studies (15%) included only out-patients, four (31%) included in-patients and out-patients and the rest did not specify their setting. Cohort sample size ranged

from 47 to 1663 patients. The average age of patients ranged from 21 to 28 years, and 49% to 77% of the cohorts were male. Where specified, the average duration of untreated psychosis ranged from 34 to 106 weeks. Ethnicity was reported in eight studies (62%) with the percentage non-white patients in the cohorts ranging from 4% to greater than 75%. The definition of FEP was primarily non-affective psychosis in the majority of patient cohorts, with the minority also including affective psychosis and two cohorts also including drug-induced psychosis patients. All but one study (92%) considered solely sociodemographic and clinical predictors. A wide range of outcomes were assessed across the 13 included studies including symptom remission in five studies (38%), global functioning in five studies (38%), vocational functioning in three studies (23%), treatment resistance in two studies (15%), rehospitalisation in two studies (15%), and quality of life in one study (8%). All the outcomes were binary. The follow-up period of included studies ranged from 1 to 10 years.

Table 2-2 Study characteristics

						Participants included in modelling					Outcome			
Study ID	Country	Multi-centre	Recruitment Dates	Type of Study	Setting	Sex (% male)	Age (mean)	Ethnicity	DUP (mean weeks)	FEP Definition	Definition	Timing	Predictor Categories	No. of Models
AJNAKINA 2020	UK	No	Dec 2005 to Oct 2010	Cohort	In-patients & out-patients	67.5%	27.2 (at baseline)	39.9% white, 60.1% black	34.3	Non-affective	Early treatment resistance from illness onset Later treatment resistance	f/u for 5 years	Socio-demographic, Clinical	4
BHATTACHARYYA 2021	UK	No	Sample 1 - 1st Apr 2006 to 31st Mar 2012 Sample 2 - 12th Apr 2002 to 26th Jul 2013	Sample 1 - Case Registry Sample 2 - Cohort	Sample 1 - out-patients Sample 2 - out-patients	Sample 1 - 63.9% Sample 2 - 60%	Sample 1 - 24.4 (at onset) Sample 2 - 28.1 (at onset)	Sample 1 - 31.1% white, 50.6% black Sample 2 - 34.2% white, 54.2% black	NR	Sample 1 - Non-affective & affective Sample 2 - Non-affective & affective	Psychiatric rehospitalisation	f/u for 2 years	Socio-demographic, Clinical	3
CHUA 2019	Singapore	No	2001 to 2012	Cohort	NR	49.2%	27.5 (at baseline)	76.7% Chinese	65.4	Non-affective	EET status	At 2 years	Socio-demographic, Clinical	2
DEMJAHA 2017	UK	Yes	Sep 1997 to Aug 1999	Cohort	NR	58.4%	28.9 (at onset)	48.2% white, 39.8% black	NR	Non-affective & affective	Early treatment resistance from illness onset	f/u for 10 years	Socio-demographic, Clinical	1
DENIJS 2019	Netherlands & Belgium	Yes	8th Jan 2004 to 6th Feb 2008	Cohort	In-patients & out-patients	76.9%	27.6 (at baseline)	85.9% white	NR	Non-affective	Andreasen symptom remission (6 months duration) GAF ≥65	At 3 years & at 6 years	Socio-demographic, Clinical, Genetic, Environmental	8
DERKS 2010	Austria, Belgium, Bulgaria, Czech Republic, Germany, France, Israel, Italy, Netherlands, Poland, Rumania, Spain, Sweden & Switzerland	Yes	23rd Dec 2002 to 14th Jan 2006	Randomised Controlled Trial	NR	56.5%	26.0 (at baseline)	NR	NR	Non-affective	Andreasen symptom remission (6 months duration)	f/u for 1 year	Socio-demographic, Clinical	1

FLYCKT 2006	Sweden	Yes	1st Jan 1996 to 31st Dec 1997	Cohort	NR	52.9%	28.8 (at baseline)	NR	62.4	Non-affective & affective (with mood-incongruent delusions)	Global functioning (independent living, EET status & GAF ≥60)	At mean of 5.4 years	Socio-demographic, Clinical	1
GONZALEZ-BLANCH 2010	Spain	No	Feb 2001 to Feb 2005	Cohort	NR	62%	26.6 (at baseline)	NR	66.6	Non-affective	Global functioning (EET status & DAS ≤1)	At 1 year	Socio-demographic, Clinical	1
KOUTSOULERIS 2016	Austria, Belgium, Bulgaria, Czech Republic, Germany, France, Israel, Italy, Netherlands, Poland, Rumania, Spain, Sweden & Switzerland	Yes	23rd Dec 2002 to 14th Jan 2006	Randomised Controlled Trial	NR	56%	26.1 (at baseline)	NR	NR	Non-affective	GAF ≥65	At 1 year	Socio-demographic, Clinical	1
LEIGHTON 2019 (1)	UK	Yes	Dev. - 2011 to 2014 Val. - 1st Sep 2006 to 31st Aug 2009	Dev. – Cohort Val. - Cohort	Dev. - In-patients & out-patients Val. - In-patients & out-patients	Dev. - 66% Val. - 68%	Dev. - 25.2 (at baseline) Val. - 24.6 (at baseline)	Dev. - 81% white Val. - 96% white	NR	Dev. - Non-affective & affective Val. - Non-affective & affective	EET Status Andreasen symptom remission (no duration criteria) Andreasen symptom remission (6 months duration)	At 1 year	Socio-demographic, Clinical	3
LEIGHTON 2019 (2)	UK & Denmark	Yes	Dev. - Aug 2005 to Apr 2009 Val. UK - 1st Sep 2006 to 31st Aug 2009 & 2011 to 2014 Val Denmark - Jan 1998 to Dec 2000	Dev. - Cohort Val. UK - 2 Cohort studies Val. Denmark - Randomised Controlled Trial	Dev. - NR Val. UK - In-patients & out-patients Val. Denmark - In-patients & out-patients	Dev. - 69% Val. UK - 67% Val. Denmark - 59%	Dev. - 21.3 (at baseline) Val. UK - 24.9 (at baseline) Val. Denmark - 26.6 (at baseline)	Dev. - 73% white Val. UK - 88% white Val. Denmark - 94% white	Dev. - 44 Val. UK - 44.4 Val. Denmark - 106	Dev. - Non-affective, affective & drug induced Val. UK - Non-affective & affective Val. Denmark - Non-affective	EET Status GAF ≥65 Andreasen Symptom Remission (6 months duration) Quality of Life	At 1 year	Socio-demographic, Clinical	4

LEIGHTON 2021	UK	Yes	Dev. - Aug 2005 to Apr 2009 Val. - Apr 2006 to Feb 2009	Dev – Cohort Val - Cohort	NR	Dev. - 68.8% Val. - 61.8%	Dev - 22.6 (at baseline) Val. - 25.0 (at baseline)	NR	Dev. - 41.3 Val. - 48.9	Dev. - Non-affective, affective & drug induced Val. - Non-affective, affective & drug induced	Andreasen Symptom Remission (6 months duration)	At 1 year	Socio-demographic, Clinical	1
PUNTIS 2021	UK	Yes	Dev. - 1st Jan 2011 to 8th Oct 2019 Val. - 31st Jan 2006 to 18th Jun 2019	Dev. - Case Registry Val. - Case Registry	Dev. - out-patients Val. - out-patients	Dev. - 63% Val. - 63%	Dev. - 25.6 (at baseline) Val. - 26.7 (at baseline)	Dev. - 74.8% white Val. - 35.4% white	NR	NR	Psychiatric hospitalisation after discharge from early intervention	f/u for 1 year	Socio-demographic, Clinical	1

FEP – first episode psychosis; NR – not reported; DUP – duration of untreated psychosis; Dev. – development sample; Val. – validation sample; EET – employment, education or training; f/u – follow-up; GAF – Global Assessment of Functioning; DAS – Disability Assessment Schedule

Study prediction modelling methodologies are outlined in Table 2-3. Nine (69%) studies pertained solely to model development with the highest level of validation reported being apparent validity in four of the studies, internal validity in three of the studies and internal-external validity (via leave one-site out cross-validation) in two of the studies. The remaining four (31%) studies also included a validation cohort and reported external validity. High dimensionality was common across the study cohorts, with the majority having a very low EPV ratio and up to 258 candidate predictors considered. Some form of variable selection was employed in the majority (62%) of studies. The number of events in the external validation cohort ranged from 23 to 173. All the studies had missing data. Six studies (46%) used complete case analysis, five (38%) used single imputation and the remaining two (15%) applied multiple imputation.

Table 2-3 Study methodology

Study ID	Sample Size	EPV	No. Events in Validation Dataset	No. Candidate Predictors	No. Retained Predictors	Variable Selection	Missing Data Per Predictor	Handling of Missing Data	Modelling Method	Shrinkage	Validation Method Reported	Re-calibration Performed	Full Model Presented	Clinical Usefulness Assessed
AJNAKINA 2020	Recruited – 283; Included in modelling - 190 to 222	2 to 4	No external validation	13	12 to 13	Full model approach or LASSO	up to 59.9%	Single imputation	Logistic regression via ridge & LASSO	Penalised estimation & then uniform	Internal	Yes	Yes	No
BHATTACHARYYA 2021	Sample 1 - Recruited - 1738; Included in modelling - 1663 Sample 2 - Recruited - 240; Included in modelling - 240	4 to 62	No external validation	10 to 21	10 to 21	Full model approach	Sample 1 - up to 4.3% Sample 2 - none	Complete case analysis	Logistic regression via MLE	None	Apparent & internal	No	Yes	No
CHUA 2019	Recruited - 1724; Included in modelling - 1177	16	No external validation	22	22	Full model approach	Yes but NR	Complete case analysis	Logistic regression via MLE	None	Apparent	No	No	No
DEMJAHA 2017	Recruited - 557; Included in modelling - 286	8	No external validation	8	6	LASSO	Yes but NR	Complete case analysis	Logistic regression via LASSO	Penalised estimation	Internal	No	Yes	No
DENIJS 2019	Recruited - 1100; Included in modelling - 442 to 523	2	No external validation	258	119 to 152	Recursive feature elimination	up to 20%	Single imputation	Linear Support Vector Machine	None	Internal & internal-external	No	No	No
DERKS 2010	Recruited - 498; Included in modelling - 297	9 to 18	No external validation	10 to 20	10 to 20	Full model approach	Yes but NR	Complete case analysis	Logistic regression via MLE	None	Apparent	No	No	No
FLYCKT2006	Recruited 175; Included in modelling - 111	2	No external validation	32	5	Forward selection	Yes but NR	Complete case analysis	Logistic regression via MLE	None	Apparent	No	Yes	No
GONZALEZ-BLANCH 2010	Recruited - 174; Included in modelling – 92	4	No external validation	23	2	Univariate significance testing (p<0.1) then forward selection	Yes but NR	Complete case analysis	Logistic regression via MLE	None	Apparent	No	Yes	No

KOUTSOULERIS 2016	Recruited - 498; Included in modelling - 334	<1	No external validation	189	NR	Forward selection	up to 20%	Single imputation	Nonlinear Support Vector Machine	None	Internal & internal-external	No	No	No
LEIGHTON 2019 (1)	Dev. - Recruited - 83; Included in modelling - 67 to 75 Val. - Recruited - 79; Included - 64 to 67	<1	27 to 46	56	5 to 13	Elastic net	Dev. - up to 13% Val. - up to 37%	Single imputation	Logistic regression via elastic net	Penalised estimation	External	No	No	No
LEIGHTON 2019 (2)	Dev. - Recruited - 1027; Included in modelling - 673 to 829 Val. UK - Recruited - 162; Included - 47 to 142 Val. Denmark - Recruited - 578; Included - 226 to 553	1 to 2	23 to 173	163	17 to 26	Elastic net	Dev. - up to 20% Val. - Yes but NR	Single imputation	Internal Validation - Logistic regression via elastic net External Validation - Logistic regression via MLE	Internal-external validation - penalised estimation External validation - none	Internal-external & external	No	No	No
LEIGHTON 2021	Dev. - Recruited - 1027; Included in modelling – 673 Val. - Recruited - 399; Included - 191	25	103	14	14	Full model approach	Dev.- up to 14.9% Val. - up to 56.5%	Multiple imputation	Logistic regression via MLE	Uniform	Internal & external	Yes	Yes	Yes

PUNTIS 2021	Dev. - Recruited - NR; Included in modelling - 831 Val. - Recruited - NR; Included - 1393	10	162	8	8	Full model approach	Dev. - up to 15.4% Val. - up to 5.5%	Multiple imputation	Logistic regression via MLE	Uniform	Internal & external	Yes	Yes	Yes

NR – not reported; Dev. – development sample; Val. – validation sample; EPV – events per variable; LASSO – least absolute shrinkage and selection operator; MLE – maximum likelihood estimation

The most common modelling methodology was logistic regression fitted by maximum likelihood estimation, then logistic regression with regularisation. Only two studies employed machine learning based methods, both via support vector machines. Just over half of studies (54%) did not use any variable shrinkage and only three studies (23%) recalibrated their models based on validation to improve performance. The full model was presented in seven (54%) studies. Only two studies (15%) assessed clinical utility.

The performance of the best model per study outcome grouped by method of validation to allow for appropriate comparisons is reported in Table 2-4. For the five studies (38%) reporting only apparent validity, two reported a measure of discrimination and only one considered calibration. For the seven studies (54%) reporting internal validation performance, four reported discrimination with a c-statistic ranging from 0.66 to 0.77 and four reported calibration. For the three studies (23%) reporting internal-external validation only one study considered discrimination with a c-statistic which ranged from 0.703 to 0.736 across each of its four models. None of the studies reporting internal-external validation considered any measure of calibration. All four studies (31%) reporting external validation considered model discrimination with c-statistics ranging from 0.556 to 0.876. However, only two of these studies considered calibration. Table 2-4 also records any global performance metrics which included the Brier score and McFadden's pseudo-R2, both of which incorporate aspects of discrimination and calibration. Various classification metrics were reported across the study models, but it is difficult to make any meaningful comparisons between these alone, without considering the models' corresponding discrimination and calibration metrics which were not universally reported.

Table 2-4 Performance metrics for best model per outcome in each study

Study ID	Outcome	Discrimination C-Statistic	Calibration	Other Global Performance Metrics	Classification Metrics
Studies Reporting Apparent Validity					
BHATTACHARYYA 2021	Psychiatric rehospitalisation	0.749	Calibration plot only; No α or ß	Brier score - 0.192	NR
CHUA 2019	EET Status at 2 years	0.759 (95%CI: 0.728, 0.790)	NR	NR	Classification Accuracy - 0.759; PPV - 0.64; NPV - 0.78; PSI - 0.42
DERKS 2010	Andreasen symptom remission (6 months duration) with 1 year f/u	NR	NR	NR	Classification Accuracy - 0.63; Balanced Accuracy - 0.665; Sensitivity - 0.73; Specificity - 0.60; PPV - 0.73; NPV - 0.61; PSI - 0.34
FLYCKT 2006	Global functioning (Independent living, EET status, GAF ≥60) at mean 5.4 years	NR	NR	NR	Classification Accuracy - 0.81; Balanced Accuracy - 0.805; Sensitivity - 0.84; Specificity - 0.77
GONZALEZ-BLANCH 2010	Global functioning (EET status, DAS ≤1) at 1 year	NR	Hosmer–Lemeshow test - p = >0.05	NR	Classification Accuracy - 0.750; Balanced Accuracy - 0.587; Sensitivity - 0.261; Specificity - 0.913; PPV - 0.500; NPV - 0.788; PSI - 0.288
Studies Reporting Internal Validity					
AJNAKINA 2020	Early treatment resistance from illness onset with 5 years f/u	0.77	α - 0.028; ß - 1.264; No calibration plot	NR	Balanced Accuracy - 0.5; Sensitivity - 0; Specificity - 1.00; PPV - 0.48, NPV - 0.84; PSI - 0.32
	Later treatment resistance with 5 years f/u	0.77	α - 0.504; ß - 1.838; No calibration plot	NR	Balanced Accuracy - 0.81; Sensitivity - 0.62; Specificity - 1.00; PPV - 0.42; NPV - 1.00; PSI - 0.42
BHATTACHARYYA 2021	Psychiatric rehospitalisation	0.66	Calibration plot only; No α or ß	Brier score - 0.232	NR
DEMJAHA 2017	Early treatment resistance from illness onset with 10 years f/u	NR	NR	Brier score - 0.146; McFadden pseudo R^2 - 0.1	NR
DENIJS 2019	Andreasen symptom remission (6 months duration) at 3 years	NR	NR	NR	Balanced Accuracy - 0.644; Sensitivity - 0.76; Specificity - 0.50; PPV - 0.722; NPV - 0.548; PSI - 0.27
	GAF ≥65 at 3 years	NR	NR	NR	Balanced Accuracy - 0.676; Sensitivity - 0.749; Specificity - 0.584; PPV - 0.701; NPV - 0.642; PSI - 0.343
	Andreasen symptom remission (6 months duration) at 6 years	NR	NR	NR	Balanced Accuracy - 0.647; Sensitivity - 0.787; Specificity - 0.465; PPV - 0.690; NPV - 0.590; PSI - 0.28
	GAF ≥65 at 6 years	NR	NR	NR	Balanced Accuracy - 0.676; Sensitivity - 0.818; Specificity - 0.477; PPV - 0.718; NPV - 0.616; PSI - 0.334
KOUTSOULERIS 2016	GAF ≥65 at 1 year	NR	NR	NR	Balanced Accuracy - 0.738; Sensitivity - 0.667; Specificity - 0.809; PPV - 0.515; NPV - 0.888; PSI - 0.403
LEIGHTON 2021	Andreasen symptom remission (6 months duration) at 1 year	0.74 (0.72, 0.76)	ß - 0.84 (95%CI: 0.76, 0.92); No calibration plot	NR	NR

PUNTIS 2021	Psychiatric hospitalisation after discharge from early intervention	0.76 (0.75, 0.77)	α - 0.01 (95%CI: -0.25, 0.24); β - 0.89 (95%CI: 0.88, 0.89); Calibration plot	Brier score - 0.078	NR
Studies Reporting Internal-External Validity					
	Andreasen symptom remission (6 months duration) at 3 years	NR	NR	NR	Balanced Accuracy - 0.638; Sensitivity - 0.629; Specificity - 0.647; PPV - 0.758; NPV - 0.485; PSI - 0.243
	GAF ≥65 at 3 years	NR	NR	NR	Balanced Accuracy - 0.648; Sensitivity - 0.658; Specificity - 0.638; PPV - 0.727; NPV - 0.565; PSI - 0.292
	Andreasen symptom remission (6 months duration) at 6 years	NR	NR	NR	Balanced Accuracy - 0.625; Sensitivity - 0.685; Specificity - 0.565; PPV - 0.743; NPV - 0.493; PSI - 0.236
DENIJS 2019	GAF ≥65 at 6 years	NR	NR	NR	Balanced Accuracy - 0.640; Sensitivity - 0.718; Specificity - 0.561; PPV - 0.732; NPV - 0.553; PSI - 0.285
KOUTSOULERIS 2016	GAF ≥65 at 1 year	NR	NR	NR	Balanced Accuracy - 0.711; Sensitivity - 0.641; Specificity - 0.781; PPV - 0.472; NPV - 0.877; PSI - 0.349
	EET Status at 1 year	0.736 (95%CI: 0.702 - 0.771)	NR	NR	Classification Accuracy - 0.693 (95%CI: 0.660, 0.725); Balanced Accuracy - 0.694 (95%CI: 0.562, 0.812); Sensitivity - 0.722 (95%CI: 0.573, 0.821); Specificity - 0.666 (95%CI: 0.550, 0.803); PPV - 0.719 (95%CI: 0.673, 0.785); NPV - 0.668 (95%CI: 0.606, 0.736); PSI - 0.387 (95%CI: 0.279, 0.521)
	GAF ≥65 at 1 year	0.731 (95%CI: 0.697, 0.765)	NR	NR	Classification Accuracy - 0.687 (95%CI: 0.657, 0.718); Balanced Accuracy - 0.691 (95%CI: 0.541, 0.825); Sensitivity - 0.722 (95%CI: 0.487, 0.778); Specificity - 0.660 (95%CI: 0.594, 0.871); PPV - 0.650 (95%CI: 0.616, 0.769); NPV - 0.726 (95%CI: 0.655, 0.766); PSI - 0.376 (95%CI: 0.271 - 0.535)
	Andreasen symptom remission (6 months duration) at 1 year	0.703 (95%CI: 0.664, 0.742)	NR	NR	Classification Accuracy - 0.670 (95%CI: 0.636, 0.703); Balanced Accuracy - 0.668 (95%CI: 0.518, 0.827); Sensitivity - 0.584 (95%CI: 0.491, 0.827); Specificity - 0.751 (95%CI: 0.544, 0.827); PPV - 0.679 (95%CI: 0.601, 0.739); NPV - 0.667 (95%CI: 0.631, 0.734); PSI - 0.346 (95%CI: 0.232, 0.473)
LEIGHTON 2019 (2)	Quality of life at 1 year	0.704 (95%CI: 0.667, 0.742)	NR	NR	Classification Accuracy - 0.668 (95%CI: 0.632, 0.704); Balanced Accuracy - 0.667 (95%CI: 0.532, 0.789); Sensitivity - 0.623 (95%CI: 0.512, 0.774); Specificity - 0.711 (95%CI: 0.551, 0.803); PPV - 0.633 (95%CI: 0.575, 0.701); NPV 0.700 (95%CI: 0.659, 0.759); PSI - 0.333 (95%CI: 0.234, 0.460)
Studies Reporting External Validity					
	EET status at 1 year	0.876 (95%CI: 0.864, 0.887)	NR	NR	Classification Accuracy - 0.851; Balanced Accuracy - 0.845; Sensitivity - 0.815; Specificity - 0.875; PPV - 0.815; NPV - 0.875; PSI - 0.690
	Andreasen symptom remission (no duration criteria) at 1 year	0.652 (95%CI: 0.635, 0.670)	NR	NR	Classification Accuracy - 0.612; Balanced Accuracy - 0.623; Sensitivity - 0.578; Specificity - 0.667; PPV - 0.794; NPV - 0.424; PSI - 0.218
LEIGHTON 2019 (1)	Andreasen symptom remission (6 months duration) at 1 year	0.630 (95%CI: 0.612, 0.647)	NR	NR	Classification Accuracy - 0.625; Balanced Accuracy - 0.626; Sensitivity - 0.606; Specificity - 0.645; PPV - 0.645; NPV - 0.606; PSI - 0.251
LEIGHTON 2019 (2) - Validated in UK	EET Status at 1 year	0.867 (95%CI: 0.805, 0.930)	NR	NR	Classification Accuracy - 0.838 (95%CI: 0.775, 0.894); Balanced Accuracy - 0.853 (95%CI: 0.740, 0.935); Sensitivity - 0.898 (95%CI: 0.780, 0.966); Specificity - 0.807 (95%CI: 0.699, 0.904); PPV - 0.766 (95%CI: 0.679, 0.867); NPV - 0.911 (95%CI: 0.840, 0.971); PSI - 0.677 (95%CI: 0.519, 0.838)

	Andreasen symptom remission (6 months duration) at 1 year	0.680 (95%CI: 0.587, 0.773)	NR	NR	Classification Accuracy - 0.695 (95%CI: 0.618, 0.771); Balanced Accuracy - 0.695 (95%CI: 0.535, 0.841); Sensitivity - 0.621 (95%CI: 0.455, 0.773); Specificity - 0.769 (95%CI: 0.615, 0.908); PPV - 0.729 (95%CI: 0.636, 0.854); NPV - 0.667 (95%CI: 0.593, 0.759); PSI - 0.396 (95%CI: 0.229, 0.613)
	Quality of life at 1 year	0.679 (95%CI: 0.522, 0.836)	NR	NR	Classification Accuracy - 0.702 (95%CI: 0.596, 0.809); Balanced Accuracy - 0.729 (95%CI: 0.407, 0.917); Sensitivity - 0.957 (95%CI: 0.564, 1.000); Specificity - 0.500 (95%CI: 0.250, 0.833); PPV - 0.640 (95%CI: 0.561, 0.800); NPV - 0.900 (95%CI: 0.643, 1.000); PSI - 0.540 (95%CI: 0.204, 0.800)
LEIGHTON 2019 (2) - Validated in Denmark	EET Status at 1 year	0.660 (95%CI: 0.610, 0.710)	NR	NR	Classification Accuracy - 0.680 (95%CI: 0.609, 0.725); Balanced Accuracy - 0.655 (95%CI: 0.516, 0.774); Sensitivity - 0.584 (95%CI: 0.457, 0.723); Specificity - 0.726 (95%CI: 0.574, 0.824); PPV - 0.490 (95%CI: 0.421, 0.563); NPV - 0.793 (95%CI: 0.760, 0.831); PSI - 0.283 (95%CI: 0.181, 0.394)
	GAF ≥65 at 1 year	0.573 (95%CI: 0.504, 0.643)	NR	NR	Classification Accuracy - 0.456 (95%CI: 0.328, 0.817); Balanced Accuracy - 0.589 (95%CI: 0.234, 0.926); Sensitivity - 0.781 (95%CI: 0.233, 0.945); Specificity - 0.396 (95%CI: 0.234, 0.906); PPV - 0.179 (95%CI: 0.158, 0.333); NPV - 0.914 (95%CI: 0.876, 0.967); PSI - 0.093 (95%CI: 0.034, 0.300)
	Andreasen symptom remission (6 months duration) at 1 year	0.616 (95%CI: 0.553, 0.679)	NR	NR	Classification Accuracy - 0.618 (95%CI: 0.524, 0.704); Balanced Accuracy - 0.621 (95%CI: 0.342, 0.864); Sensitivity - 0.612 (95%CI: 0.306, 0.843); Specificity - 0.629 (95%CI: 0.378, 0.885); PPV - 0.476 (95%CI: 0.412, 0.636); NPV - 0.742 (95%CI: 0.687, 0.829); PSI - 0.217 (95%CI: 0.099, 0.465)
	Quality of life at 1 year	0.556 (95%CI: 0.481, 0.631)	NR	NR	Classification Accuracy - 0.589 (95%CI: 0.540, 0.637); Balanced Accuracy - 0.589 (95%CI: 0.312, 0.845); Sensitivity - 0.876 (95%CI: 0.419, 0.947); Specificity - 0.301 (95%CI: 0.204, 0.743); PPV - 0.559 (95%CI: 0.527, 0.642); NPV - 0.706 (95%CI: 0.555, 0.841); PSI - 0.265 (95%CI: 0.081, 0.483)
LEIGHTON 2021	Andreasen symptom remission (6 months duration)	0.73 (95%CI: 0.64, 0.81)	α - -0.014 (95%CI: -0.34, 0.31); β - 0.85 (95%CI: 0.42, 1.27); Calibration plot	NR	NR
PUNTIS 2021	Psychiatric hospitalisation after discharge from early intervention	0.70 (95%CI: 0.66, 0.75)	α - -0.01 (95%CI: -0.17, 0.167); β - 1.00 (95%CI: 0.78, 1.22); Calibration plot	Brier score - 0.094	NR

NR – not reported; EET – employment, education or training; GAF – Global Assessment of Functioning; DAS – Disability Assessment Schedule; PPV – positive predictive value; NPV – negative predictive value; PSI – prognostic summary index; f/u – follow-up

We applied the PROBAST tool to the 31 different prediction models across the 13 studies in our systematic review and determined an overall risk of bias rating for each study as summarised in Table 2-5. The majority (85%) of studies had an overall ‘high’ ROB. In each of these studies, the ROB was rated ‘high’ in the analysis domain with one study also having a ‘high’ ROB in the predictors domain. The main reasons for the ‘high’ ROB in the analysis domain were insufficient participant numbers and consequently low EPV, inappropriate methods of variable selection including via univariable analysis, a lack of appropriate validation with only apparent validation, an absence of reported measures of discrimination and calibration, and inappropriate handling of missing data by either complete case analysis or single imputation. Two studies, Leighton et al 2021 (Leighton et al., 2021) and Puntis et al 2021,(Puntis et al., 2021) were rated overall ‘low’ ROB. These studies considered symptom remission and psychiatric rehospitalisation outcomes, respectively. Both studies externally validated their prediction model and considered its clinical utility. However, neither study considered the implementation of the prediction model into actual clinical practice. When we assessed the 13 included studies according to PROBAST applicability concerns, all the studies were considered overall ‘low’ concern. This is indicative of the broad scope of our systematic review.

Table 2-5 PROBAST risk of bias for each study

Study ID	Participants	Predictors	Outcome	Analysis	Overall
AJNAKINA 2020	Low	Low	Low	High	High
BHATTACHARYYA 2021	Low	Low	Low	High	High
CHUA 2019	Low	High	Low	High	High
DEMJAHA 2017	Low	Low	Low	High	High
DENIJS 2019	Low	Low	Low	High	High
DERKS 2010	Low	Low	Low	High	High
FLYCKT 2006	Low	Low	Low	High	High
GONZALEZ-BLANCH 2010	Low	Low	Low	High	High
KOUTSOULERIS 2016	Low	Low	Low	High	High
LEIGHTON 2019 (1)	Low	Low	Low	High	High
LEIGHTON 2019 (2)	Low	Low	Low	High	High
LEIGHTON 2021	Low	Low	Low	Low	Low
PUNTIS 2021	Low	Low	Low	Low	Low

2.5 Discussion

Our systematic review identified 13 studies reporting 31 prognostic prediction models for the prediction of a wide range of clinical outcomes. The majority of models were developed via logistic regression. There were several methodological limitations identified including a lack of appropriate validation, issues with handling missing data and a lack of reporting of calibration and discrimination measures. We identified two studies with models at low risk of bias as assessed with PROBAST, both of which externally validated their models.

2.5.1 Principal findings in context

Our systematic review found no consistent definition of FEP across the different cohorts used for developing and validating prediction models. A lack of an operational definition for FEP within clinical and research settings has previously been identified as major a barrier to progress (Breitborde et al., 2009). The majority of cohorts in our systematic review included only individuals with non-affective psychosis with a minority also including affective psychosis. In contrast, early intervention services typically do not make a distinction between affective and non-affective psychosis in those whom they accept into their service (National Institute for Health and Care Excellence (NICE), 2016b). As such, there may be issues with generalisability of prediction models developed in cohorts with solely non-affective psychosis to real-world clinical practice.

A wide range of different outcomes were predicted by the FEP models including symptom remission, global functioning, vocational functioning, treatment resistance, rehospitalisation and quality of life outcomes. This is reflective of the fact that recovery from FEP is not readily distilled down to a single factor like symptom remission. Meaningful recovery is represented by a constellation of multidimensional outcomes unique to each individual (Jaaskelainen et al., 2013). We should engage people with lived experience, to ensure that prediction models are welcomed and are predicting outcomes most relevant to the people they are for.

All the prediction models were developed in populations from high-income developed countries and only three studies included participants from countries

outside of Europe, an issue not unique to FEP research. Consequently, it is currently unknown how prediction models for FEP would generalise to low-income developing countries. Prediction models may have considerable benefit in developing countries where almost 80% of patients with FEP live but where mental health support is often scarce (Singh et al., 2020). Prediction models could help prioritise the appropriate utilisation of limited healthcare resources.

Only one study considered predictor variables other than clinical or sociodemographic factors. In this study, the additional predictors did not add significant value (de Nijs, 2019). In recent years substantial progress has been made in elucidating the pathophysiological mechanisms underpinning the development of psychosis. We now recognise important roles for genetic factors, neurodevelopmental factors, dopamine and glutamate (Lieberman & First, 2018). Prediction model performance may be improved by the incorporation of these biological relevant disease markers as predictor variables. However, the cost-benefit of adding more expensive and less accessible disease markers must be carefully considered, especially if models are to be utilised in settings where resources are more limited.

Machine learning can be operationally defined as “models that directly and automatically learn from data” (Christodoulou et al., 2019). This is to be contrasted with regression models which “are based on theory and assumptions, and benefit from human intervention and subject knowledge for model specification”(Christodoulou et al., 2019). Just two studies employed machine learning techniques for their modelling (de Nijs, 2019; Koutsouleris et al., 2016). The rest of the studies employed logistic regression. We were unable to make any comparison between the discrimination and calibration ability of the two studies employing machine learning and the other studies because these metrics were not provided. However, a recent systematic review found no evidence of superior performance of clinical prediction models using machine learning methods over logistic regression (Christodoulou et al., 2019). In any case, the distinction between regression models and machine learning has been viewed to be artificial. Instead, algorithms may exist “along a continuum between fully human-guided to fully machine-guided data analysis” (Beam & Kohane, 2018). An alternative comparison may be between linear and non-linear classifiers. Only one study employed a non-linear classifier (Koutsouleris et al., 2016), but again

we were unable to gain meaningful insights into its relative performance because appropriate metrics were not provided.

A principal finding from our systematic review is the presence of methodological limitations across the majority of studies. Steyerberg et al outline four key measures of predictive performance that should be assessed in any prediction modelling study – two measures of calibration (the model intercept (A) and the calibration slope (B)), discrimination via a concordance statistic (C), and clinical usefulness with decision-curve analysis (D) (Steyerberg & Vergouwe, 2014). Model calibration is the level of agreement between the observed outcomes and the predictions. For example, if a model predicts a 5% risk of cancer, then, according to such a prediction, the observed proportion should be five cancers per 100 people. Discrimination is the ability of a model to distinguish between a patient with the outcome and one without (Steyerberg & Vergouwe, 2014). Our review found that only seven studies (54%) reported discrimination and just five (38%) reported any measure of calibration. The remaining studies reported only classification metrics, such as accuracy or balanced accuracy. The problem with solely reporting classification metrics is that they vary both across models and across different probability thresholds for the same model. This renders the comparison between models less meaningful. It is further argued that setting a classification threshold for a probability generating model is premature. Rather, a clinician may choose to set different probability thresholds for the same prediction model depending on the situation at hand in order to optimise the balance between false positives and false negatives. For example, in the case of a model predicting cancer, a clinician may choose a lower probability threshold to offer a non-invasive screening test and a higher probability threshold to suggest an invasive and potentially harmful biopsy. Further, without any measure of model calibration we are unable to assess if the model can make unbiased estimates of outcome (Harrell, 2015). The final key step in assessing the performance of a prediction model is to determine its clinical usefulness – that is, can better decisions be made with the model than without? Decision-curve analysis considers the net-benefit (the treatment threshold weighted sum of true- minus false-positive classifications) for a prediction model in comparison the default strategy of treating all or no patients, across an entire range of treatment thresholds (Vickers et al., 2019). Only two studies (15%) included in

our review considered whether the model was clinically useful. Without proper validation of prediction models, the reported performances are likely to be overly optimistic. Four studies (31%) report only apparent validity. Just four studies (31%) reported external validation, considered essential before applying a prediction model to clinical practice (Steyerberg & Harrell, 2016).

Altogether, just two studies (15%) had an overall 'low' risk of bias according to PROBAST, reflecting these methodological limitations. Neither study considered real-world implementation. To progress with implementation, impact studies are required. These would involve a cluster randomised trial comparing patient outcomes between a group with treatment informed by a clinical prediction model and a control group (Moons et al., 2012). We are not aware of any such study having been carried out within the field of psychiatry. However, Salazar de Pablo et al suggest that PROBAST thresholds for considering a study to be a 'low' risk of bias may be too strict (Salazar de Pablo et al., 2021). Indeed, in the field of machine learning multiple imputation is frequently computationally infeasible and single imputation may be viewed as sufficient. This is especially true in larger datasets or in the presence of relatively few missing values (Steyerberg, 2019).

2.5.2 Strengths and limitations

Our review had a number of strengths. We provide the first systematic overview of prediction modelling studies for use in patients with FEP. We offer a detailed critique of the study characteristics, their methodologies and model performance metrics. Further, our review adheres to gold standard guidance for extracting data from prediction models and for assessing bias, namely the CHARMS checklist and PROBAST.

There were several limitations. Our initial aim was to perform a meta-analysis of any prediction model which was validated across different settings and populations. However, no meta-analysis was possible because no single prediction model was validated more than once. In addition, as a consequence of poor reporting of discrimination and calibration performance across the studies, it was often difficult to make meaningful comparison between the prediction models. Also, the lack of consensus as to the most important outcome

measure in FEP, with six different outcomes considered across only 13 included studies, further hindered efforts at drawing meaningful comparisons between the included studies and their respective prediction models. Likewise, if more studies had considered the same outcome measures, this may have afforded the opportunity to validate existing prediction models rather than necessitating the creation of additional new models. All published prediction modelling studies in FEP reported significant positive findings. It is possible that studies which had negative findings were held back from publication reflecting the possibility of publication bias. We originally intended to evaluate the overall certainty in the body of evidence using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework (Schunemann et al., 2008). GRADE was originally designed for reviews of intervention studies but has not yet been adapted for use in systematic reviews of prediction models. Consequently, in its current form we did not find GRADE to be a suitable tool for our review and decided not to use it. Future research should consider how to adapt GRADE for use in systematic reviews of prediction models.

2.5.3 Implications for future research

It is clear that there is a growing trend for the development of prediction models in FEP (Salazar de Pablo et al., 2021). FEP is an illness which responds best to an early intervention paradigm (Birchwood et al., 1998). Prediction models have the potential to optimise the allocation of time-critical interventions, like clozapine for treatment resistance (Farooq et al., 2019). However, prior to meaningful implementation into real-world clinical practice several steps are necessary. The field must prioritise external validation and replication of existing prediction models in larger sample sizes to increase the EPV. This is best accomplished by an emphasis on data-sharing and open collaboration. Prediction studies should include FEP cohorts from low-income countries where there is considerable potential for benefit by helping to prioritise limited resources to those most in need. Harmonisation of data collection across the field both in terms of predictors and outcomes measured would facilitate validation efforts. There should be a greater consideration of biologically relevant and cognitive predictors based on our growing understanding of disease mechanisms, which could optimise prediction model performance. Finally, our review highlights considerable methodological pitfalls in much of the current literature. Future

prediction modelling studies should focus on methodological rigour with adherence to accepted best practice guidance (Harrell, 2015; Steyerberg & Harrell, 2016; Steyerberg & Vergouwe, 2014). Our goal in psychiatry should be to develop an innovative approach to care using prediction models. Application of these approaches into clinical practice would enable rapid and targeted intervention thereby limiting treatment associated risks and reducing patient suffering.

Chapter 3 Development and validation of a nonremission risk prediction model in first-episode psychosis: an analysis of two longitudinal studies

3.1 Overview of this chapter

This chapter presents the development and validation of a nonremission risk prediction model in first-episode psychosis (FEP). The article was published in Schizophrenia Bulletin Open on 31st August 2021.

A corrigendum was published on 20th November 2022 which corrected an error which had resulted when calculating the confidence intervals for the results after combining the data across multiple imputations. This had resulted in the estimates of the values with narrower confidence intervals than if we had correctly applied Rubin's Rules (Rubin, 1987).

A further corrigendum was published on 3rd April 2023 which corrected an error resulting from the fact we had originally standardised (centred and scaled) the development (NEDEN) and validation cohort (Outlook) data separately. This is not best practice and instead we should have used the means and standard deviations from the development cohort to standardise the validation cohort (Hastie et al., 2009; Matthew Drury (https://stats.stackexchange.com/users/74500/matthew-drury)). The results are very similar to the original published results and the interpretation is largely unchanged. The only change is that the external validation calibration slope, although its confidence intervals still overlap the ideal, does suggest a degree of overfitting.

This chapter presents the corrected version of the original article.

3.2 Introduction

Psychotic disorders, including schizophrenia, are among the 20 leading causes of disability worldwide in 2017. People with psychosis have heterogeneous outcomes with more than 40% not achieving symptomatic remission (Lally et al.,

2017). Symptom remission after the FEP is associated with long-term functional outcome (Jordan et al., 2014). The main modifiable reasons for nonremission include treatment resistance (Smart et al., 2021), medication nonadherence (Kane et al., 2013), and comorbid substance misuse (Weibell et al., 2017). Although there are effective interventions to ameliorate the reasons for nonremission, there is often a delay in providing these interventions. For people with treatment-resistant psychosis, delays of around four years in initiating effective interventions have been reported – for example, clozapine for treatment resistance (Howes et al., 2012). Delay is associated with poorer outcomes. Clinicians have identified the difficulty of early identification of patients likely to become treatment-resistant as a barrier preventing the initiation of effective phase-specific treatments like clozapine at the optimal time (Farooq et al., 2019).

Early identification of individuals with a higher risk of nonremission at initial clinical contact may facilitate personalized interventions, reduce time to their initiation and improve utilization of resources. Although there have been recent attempts to develop models to predict the individual risk of poor outcome in FEP (Koutsouleris et al., 2016; Leighton, Krishnadas, et al., 2019; Leighton, Upthegrove, et al., 2019), these are affected by suboptimal study design and reporting, lack of external validation (Koutsouleris et al., 2016), small sample sizes (Leighton, Krishnadas, et al., 2019), and no measures of calibration or clinical utility (Koutsouleris et al., 2016; Leighton, Krishnadas, et al., 2019; Leighton, Upthegrove, et al., 2019). This study aimed to develop and externally validate a new prediction model to predict the individual risk of nonremission at one year for individuals with first-episode psychosis.

3.3 Methods

3.3.1 Data sources and study population

We used data from the National Evaluation of Development of Early intervention Network study (NEDEN) for model development and internal validation. We used data from the Outlook study for external validation. Written informed consent was obtained from all participants. Both studies had NHS Research Ethics Committee approval.

3.3.2 Development cohort

NEDEN is a longitudinal naturalistic study of 1027 patients aged 14–35 with FEP recruited from 14 early intervention services across the National Health Service (NHS) in England (2005–2010); the methods and baseline characteristics have been outlined previously (Birchwood et al., 2014). An analysis of the potential of prediction modelling in FEP using this dataset has been published. We conducted a reanalysis to address methodological issues with our previous analysis (Leighton, Upthegrove, et al., 2019) (including the lack of any measures of calibration or clinical utility) and to take advantage of an external validation dataset that was similar to the development dataset (in terms of patients, geography, and clinical service they were drawn from) allowing for better assessment of generalizability. Models in the previous analysis did not inform the present study.

3.3.3 Validation cohort

The Outlook study (which was part of the PsyGrid study) is a longitudinal naturalistic study of 399 patients recruited from a further 11 NHS England early intervention services, throughout April 2006–February 2009 (Drake et al., 2020). Inclusion criteria: age 16–35, International Classification of Diseases 10th Revision (ICD-10) diagnosis of schizophrenia, schizoaffective disorder, delusional disorder, mania or severe depression with psychosis, acute and transient psychoses, drug-induced psychoses and psychosis not otherwise specified; those with organic brain disorders were excluded.

In both cohorts, participants were recruited as soon after the first contact with the early intervention services as possible. Baseline assessment occurred as soon as a referral was received by a participating service, regardless of whether the potential participant was in the hospital or the community.

3.3.4 Outcome measure

Our outcome measure was symptom nonremission at one year. Nonremission was defined as failing to meet the Remission in Schizophrenia Working Group criteria using the Positive And Negative Syndrome Scale (PANSS) at six and 12 months, a reliable and valid scale in clinical and research settings. The Remission in

Schizophrenia Working Group defined remission as scores of less than or equal to 3 in PANSS items P1 Delusions, P2 Conceptual Disorganization, P3 Hallucinatory Behavior, N1 Blunted Affect, N4 Apathetic Social Withdrawal, N6 Lack of Spontaneity and G9 Unusual Thought Content, present for a period of at least six months (Andreasen et al., 2005).

3.3.5 Candidate predictors

In both cohorts, psychologists not directly involved in clinical care trained in the use of the rating scales assessed participants at baseline, six- and 12-month follow-up. Both studies collected candidate predictors based on existing literature and expert knowledge using standardized assessment instruments. These included sociodemographic and clinical variables, the Premorbid Adjustment Scale, PANSS, Young Mania Rating Scale, Birchwood Insight Scale, Calgary Depression Scale for Schizophrenia, Global Assessment of Functioning, and EQ-5D. In addition, participant UK postcode outward code was mapped to primary care trust (PCT). Summary PCT level UK Government Index of Multiple Deprivation (IMD) data (collected between 2001 and 2005, released 2007) was then linked to each patient.

Fourteen predictors were chosen based on previous research and consensus between five psychiatrists working in the field of Early intervention in Psychosis. The list of predictors is provided in Table 3-1. As outlined above, similar research involving feature selection was performed using the NEDEN dataset (Leighton, Upthegrove, et al., 2019). This did not influence the choice of predictors for the present study.

Table 3-1 The final logistic regression nonremission prediction model specification. We now provide mean and standard deviation values to allow the transformation of the predictor variables to Z-scores for their use in the model. This was omitted from the original published paper but is required to apply the model to new patients.

Variable	Values to transform to Z-score	Unadjusted Final Model		Adjusted Final Model (Shrinkage Factor = 0.84)	
	Mean (SD)	β Coefficient (95% CI)	Odds Ratio (95% CI)	β Coefficient	Odds Ratio
Intercept		0.022 (-0.334, 0.379)		0.029	

Male Sex (1 or 0)	N/A	0.259 (-0.129, 0.646)	1.295 (0.879, 1.908)	0.217	1.242
Age at Study Entry	22.51 (4.887)	-0.037 (-0.210, 0.137)	0.964 (0.810, 1.147)	-0.031	0.970
Past Drug Use (1 or 0)	N/A	-0.101 (-0.478, 0.277)	0.904 (0.620, 1.319)	-0.084	0.919
DUP (days)	307.5 (632.3)	0.546 (0.255, 0.838)	1.727 (1.291, 2.311)	0.460	1.581
PAS Highest Functioning Achieved	1.745 (1.446)	0.427 (0.241, 0.613)	1.533 (1.273, 1.847)	0.358	1.431
PANSS P1 Delusions	2.828 (1.683)	0.060 (-0.166, 0.287)	1.062 (0.847, 1.332)	0.051	1.052
PANSS P2 Conceptual Disorganization	1.945 (1.254)	-0.359 (-0.568, -0.151)	0.698 (0.567, 0.860)	-0.301	0.740
PANSS P3 Hallucinatory Behavior	2.931 (1.686)	0.543 (0.334, 0.753)	1.722 (1.396, 2.123)	0.455	1.577
PANSS N4 Passive Social Withdrawal	2.68 (1.576)	0.346 (0.146, 0.545)	1.413 (1.157, 1.725)	0.290	1.336
PANSS G6 Depression	3.229 (1.681)	-0.198 (-0.398, 0.002)	0.820 (0.672, 1.002)	-0.166	0.847
Insight Scale – Nervous or Mental Illness	1.288 (0.7951)	-0.075 (-0.263, 0.114)	0.928 (0.768, 1.121)	-0.062	0.940
GAF Symptoms	51.48 (16.72)	-0.272 (-0.540, -0.005)	0.762 (0.583, 0.995)	-0.228	0.780
GAF Disability	53.27 (15.58)	-0.019 (-0.267, 0.229)	0.981 (0.765, 1.257)	-0.016	0.984
Average Deprivation Score in Patient's PCT	27.27 (12.22)	0.221 (0.029, 0.414)	1.248 (1.029, 1.513)	0.185	1.204

DUP = duration of untreated psychosis; PAS = premorbid adjustment scale; PANSS = Positive and Negative Syndrome Scale; GAF = Global Assessment of Functioning; PCT = Primary Care Trust; N/A = not applicable

3.3.6 Sample size calculation

Using Riley et al's (Riley et al., 2019) criteria for multivariable prediction model development for binary outcomes, the minimum sample size required given a

50% prevalence of nonremission with 14 predictor parameters (meeting the assumptions of global shrinkage factor of ≥0.90, an absolute difference of ≤0.05 between apparent and adjusted R-squared, and a 0.05 margin of error in the estimation of intercept) is 431 with 216 nonremitters (Events per Predictor Parameters [EPP] = 15). Our development cohort included 673 FEP individuals with 353 individuals meeting criteria for nonremission at one year. This provides an EPP of 25, which is above requirements. Further, the number of nonremission events in both the development and validation cohort was >100, which is in line with suggested criteria (Steyerberg, 2019). The number of nonevents in the validation cohort was 88, just below suggested criteria.

3.3.7 Missing data

Missing data were multiply imputed (m = 10) by chained equations using all predictors, auxiliary variables, and outcomes based on the assumption that data was missing at random. Imputed outcome data were then deleted (von Hippel, 2007). It is proposed that this strategy leads to more efficient estimates than an ordinary multiple imputation strategy while also protecting the estimates from problematic imputations in the outcome variable (Steyerberg, 2019). This multiple imputation strategy was carried out separately for the development and validation datasets. Ordinal variables were treated as continuous and binary variables were dummy coded. All predictor variables were standardized prior to model construction.

3.3.8 Statistical analysis for model development

We followed the TRIPOD (Transparent Reporting of a multivariable model for Individual Prognosis Or Diagnosis) guidance for development and reporting of multivariable prediction models (Moons et al., 2015).

3.3.9 Model development, internal, and external validation

A logistic regression model was fitted by maximum likelihood estimation on the 14 chosen predictors. Internal validation performance was assessed by 10-fold cross-validation repeated five times on the 10 imputed datasets. The model performance was considered using discrimination and calibration measures. Discrimination, or the ability of our model to distinguish a patient with the

outcome (nonremission) from a patient without (remission), was assessed via the c-statistic (with 95% CIs were established via U-statistic theory and permutation testing to confirm significance). Calibration is the level of agreement between the observed outcomes and the model's predictions. Two measures of model calibration were calculated: calibration-in-the-large (alpha) which is the intercept on the calibration plot and compares mean observed to mean predicted, and, the calibration slope (beta) which relates to the shrinkage of the regression coefficients. A perfectly calibrated model would show an ideal line with intercept alpha of 0 and a slope beta of 1. For internal validation, only the slope beta is of value and corresponds to the shrinkage factor or measure of overfitting (Steyerberg & Vergouwe, 2014). This uniform shrinkage factor was applied to the final logistic regression model and the intercept was re-estimated prior to external validation on the Outlook dataset.

3.3.10 Clinical utility

Clinical utility was assessed in the external validation cohort, in addition to discrimination and calibration. We assessed the clinical usefulness of using a treatment strategy based on the prediction model compared with treating all, treating none, or treating based on the duration of untreated psychosis (DUP) alone (DUP is the most researched and consistent predictor of poor outcome in FEP). Hereto, a decision curve analysis was performed (Vickers & Elkin, 2006). Clinical usefulness is considered in terms of net-benefit (the treatment threshold weighted sum of true- minus false-positive classifications for each strategy – Equation 3-1) plotted against an entire range of treatment thresholds. A treatment threshold is defined as the point where the likelihood of benefit, in our case, reduced rates of nonremission, exactly balances the likelihood of harm. Treatment thresholds vary between individual clinicians and patients depending on their context-specific weighting of relative harms and benefits.

Equation 3-1 Net benefit

$$Net\ Benefit = \frac{True\ Positives}{N} - \frac{False\ Positives}{N}\ x\ \frac{Threshold\ Probabilty}{1 - Threshold\ Probability}$$

N = total sample size

The intervention ("treatment") proposed on the prediction of a high risk of nonremission is "enhanced monitoring" over routine care leading to early identification and intervention for treatment resistance, substance misuse, or nonconcordance. To use a prediction model for such treatment decisions, we require to specify a probability threshold above which we would consider the treatment.

We consulted NHS early-intervention specialists (eight NHS Consultant Psychiatrists) to ascertain the probability threshold at which they would consider treatment. The range of thresholds varied between 40% and 60% That is; they would adopt an assertive monitoring and intervention approach when an individual's probability of nonremission is above 40%–60% to balance the likelihood of benefits versus the harms/costs (in this case, the benefit of reduced rates of nonremission against the probability of harm via intrusive monitoring, side-effects, and increased costs).

Net-benefit is calculated across the range of threshold probabilities of the outcome (nonremission) at which further intervention would be warranted. Net-benefit differs from other performance metrics such as discrimination and calibration because it incorporates the consequences of the decisions made based on a model.

All analyses were performed using R, Comprehensive R Archive Network (CRAN) version 4.1.0 (R Core Team, 2021) (with the "mice"(van Buuren & Groothuis-Oudshoorn, 2011), "caret"(Kuhn, 2008), "pROC"(Robin et al., 2011), "CalibrationCurves," and "dca" packages) and code is available in Appendix 2. The analysis pipeline is provided in Figure 3-1.

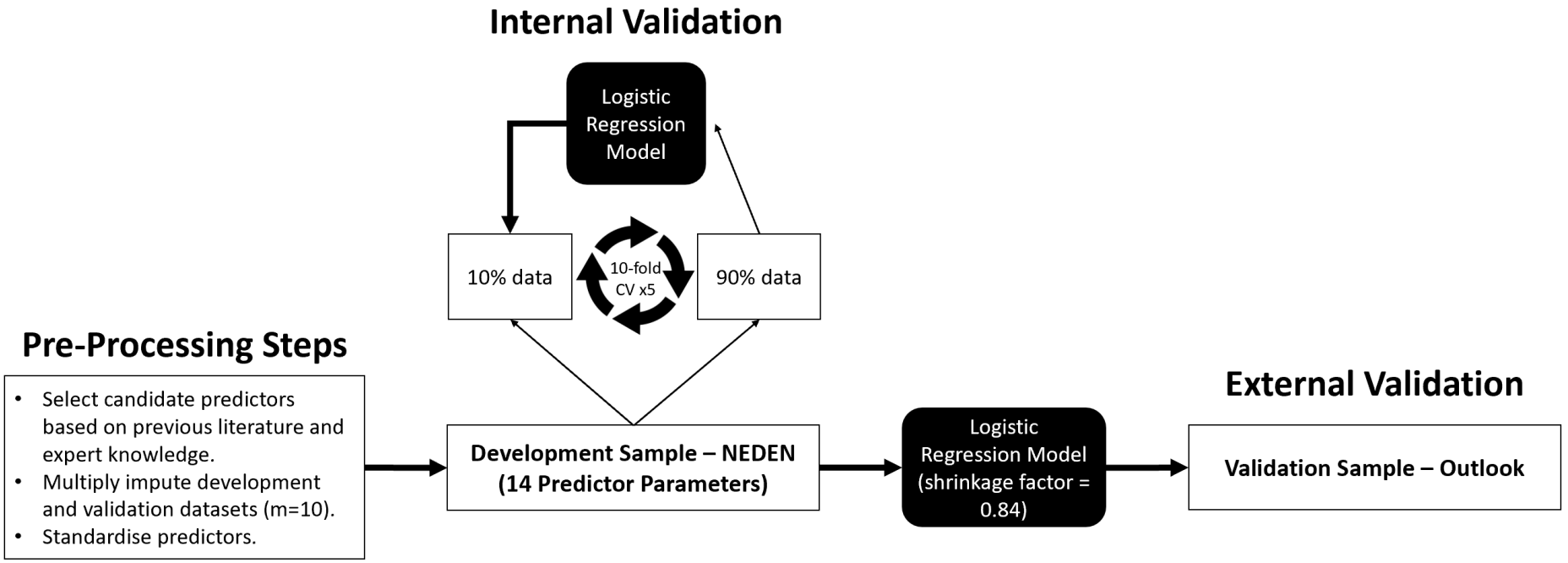


Figure 3-1 Analysis pipeline

3.4 Results

3.4.1 Study populations

In the NEDEN study, 673 (66%) of 1027 participants had outcome data, of which 353 (52%) met criteria for 1-year symptom nonremission. In the Outlook study, 191 (48%) of 399 participants had outcome data, of which 103 (54%) met criteria for nonremission. The baseline characteristics of the development (NEDEN) and validation (Outlook) cohorts are summarized in Table 3-2.

Table 3-2 Baseline characteristics of the development (NEDEN) and validation (Outlook) cohorts.

Baseline Characteristic	**Development Cohort (NEDEN)** (n=1027)		**Validation Cohort (Outlook)** (n=399)		**P-Value**
	All	**With Outcome Data** (n=673; 66%)	**All**	**With Outcome Data** (n=191; 48%)	
Age (Years)					ANOVA
Mean (SD)	22.5 (4.89)	22.6 (5.01)	25.6 (7.91)	25.0 (7.84)	*F*(3,2284)
Missing	1 (0.1%)	1 (0.1%)	0 (0%)	0 (0%)	= 35.7, *p* = <0.001*

Sex					Chi-squared test $x^2(3) = 10.4$, $p = 0.015$*
Male	709 (69.0%)	463 (68.8%)	246 (61.7%)	118 (61.8%)	
Female	318 (31.0%)	210 (31.2%)	153 (38.3%)	73 (38.2%)	
Missing	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
In Employment, Education or Training	284 (32.5%)	190 (33.2%)	174 (43.6%)	85 (44.5%)	Chi-squared test $x^2(3) = 22.6$, $p = <0.001$*
Missing	154 (15.0%)	100 (14.9%)	0 (0%)	0 (0%)	
Highest Qualification					Chi-squared test $x^2(9) = 26.2$, $p = 0.002$*
None	245 (24.4%)	156 (23.7%)	89 (23.4%)	40 (21.3%)	
GCSE/NVQ level 1 or 2	399 (39.7%)	255 (38.8%)	130 (34.2%)	67 (35.6%)	
A-level/GNVQ/ BTEC/NVQ level 3	262 (26.1%)	173 (26.3%)	92 (24.2%)	46 (24.5%)	
Degree/HND/NVQ level 4 or above	98 (9.8%)	74 (11.2%)	69 (18.2%)	35 (18.6%)	
Missing	23 (2.2%)	15 (2.2%)	19 (4.8%)	3 (1.6%)	

Adjusted Duration of Untreated Psychosis (Days)					ANOVA $F(3, 2162) = 0.291$, $p = 0.832$
Mean (SD)	308 (633)	289 (589)	293 (839)	342 (1000)	
Missing	16 (1.6%)	7 (1.0%)	62 (15.5%)	39 (20.4%)	
Average Deprivation Score in Patient's Primary Care Trust					ANOVA $F(3, 1959) = 5.0$, $p = 0.002$*
Mean (SD)	27.3 (12.2)	26.6 (11.6)	30.0 (8.0)	28.9 (7.9)	
Missing	6 (0.6%)	4 (0.6%)	209 (52.4%)	108 (56.5%)	
Positive and Negative Syndrome Scale Total					ANOVA $F(3, 2054) = 1.64$, $p = 0.178$
Mean (SD)	62.7 (18.8)	64.1 (19.0)	63.0 (15.5)	60.9 (14.5)	
Missing	105 (10.2%)	42 (6.2%)	67 (16.8%)	18 (9.4%)	

* indicates significance after Bonferroni-Holm correction (n = 7)

3.4.2 Model development and internal validation

The 14 variable logistic regression prediction model is specified in Table 3-1.

At internal validation, the discrimination c-statistic was 0.74 (0.72, 0.76), while the calibration slope beta of 0.84 (0.76, 0.92). This shrinkage factor was applied to the final model coefficients and the intercept was re-estimated (see Appendix 3 for note on internal validation method).

3.4.3 External validation

At external validation, the model showed fair discrimination with a c-statistic of 0.73 (0.64–0.81). There was a good spread of risk, with good correspondence between observed proportions with psychosis for subjects grouped by similar predicted risk (Figure 3-2).

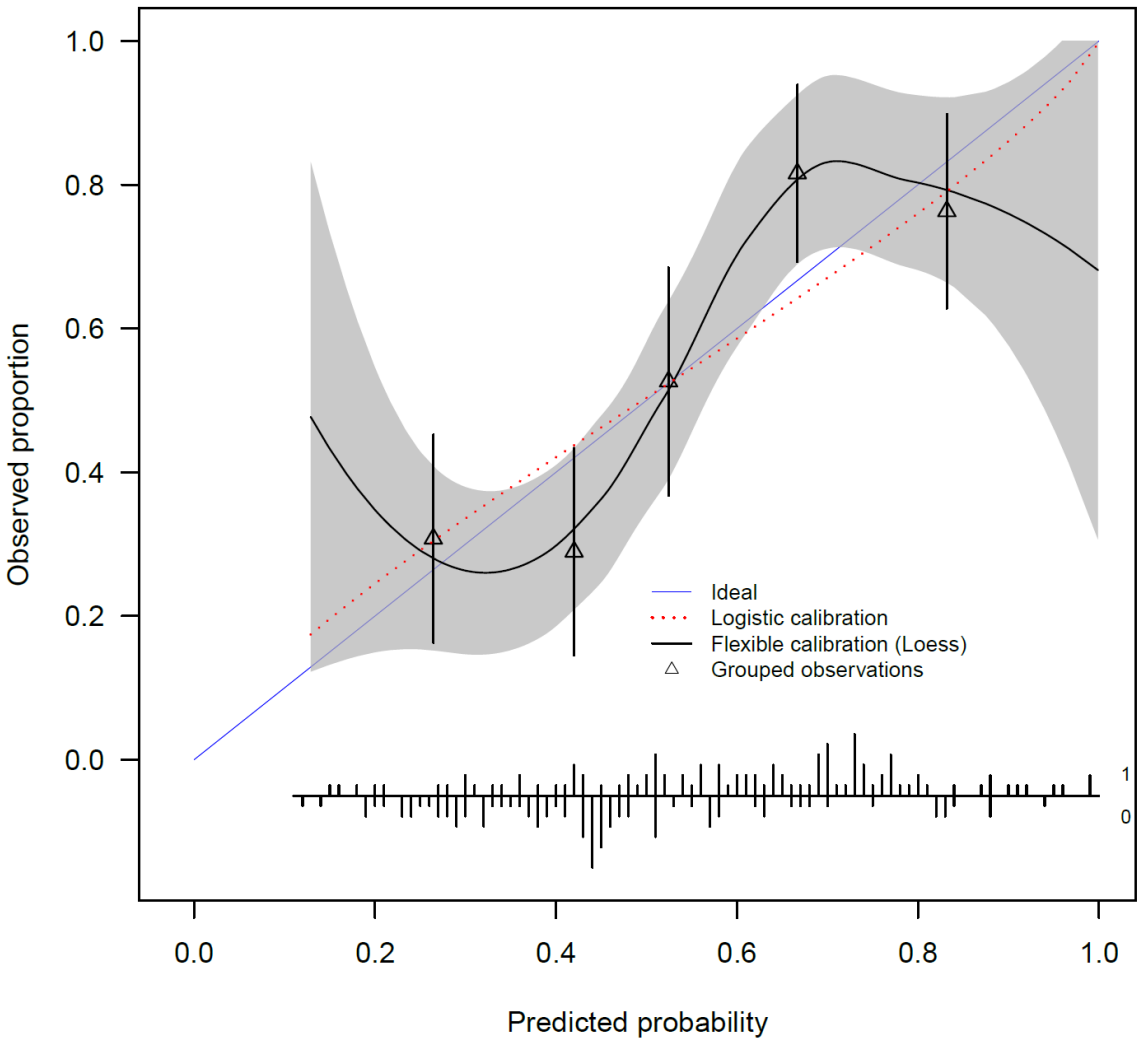


Figure 3-2 External validation calibration plot (first imputed dataset). The calibration intercept of -0.014 (-0.34, 0.31) and slope 0.85 (0.42, 1.27). Triangles represent quintiles of subjects grouped by similar predicted risk. The distribution of subjects is indicated with spikes at the bottom of the graph, stratified by endpoint (nonremitters above the x-axis, remitters below the x-axis). Although both sets of confidence intervals overlapped the ideal values, the calibration slope point estimate is smaller than 1 indicating that the predicted risks were too extreme in the sense of overestimating for patients at high risk while underestimating for patients at low risk and is indicative of overfitting of the model. The calibration intercept point estimate was close to ideal suggesting no general over- or underestimation of predicted risks.

For the Outlook external validation, the 54% overall rate of nonremission at one year implies a maximal net-benefit of 54% at a decision threshold for treatment of 0%. Figure 3-3 shows that, between thresholds of 35%–70%, treating based on our model is better than treating all, treating none or treating using DUP alone. At a probability threshold of 50% (midpoint of the range of clinician chosen thresholds), treating based on our model has an increased net-benefit of 16% compared the strategy of treating all, equivalent to 16 more detected nonremitted FEP individuals per 100 FEP individuals without an increase in incorrect classification of remitted FEP individuals as high risk.

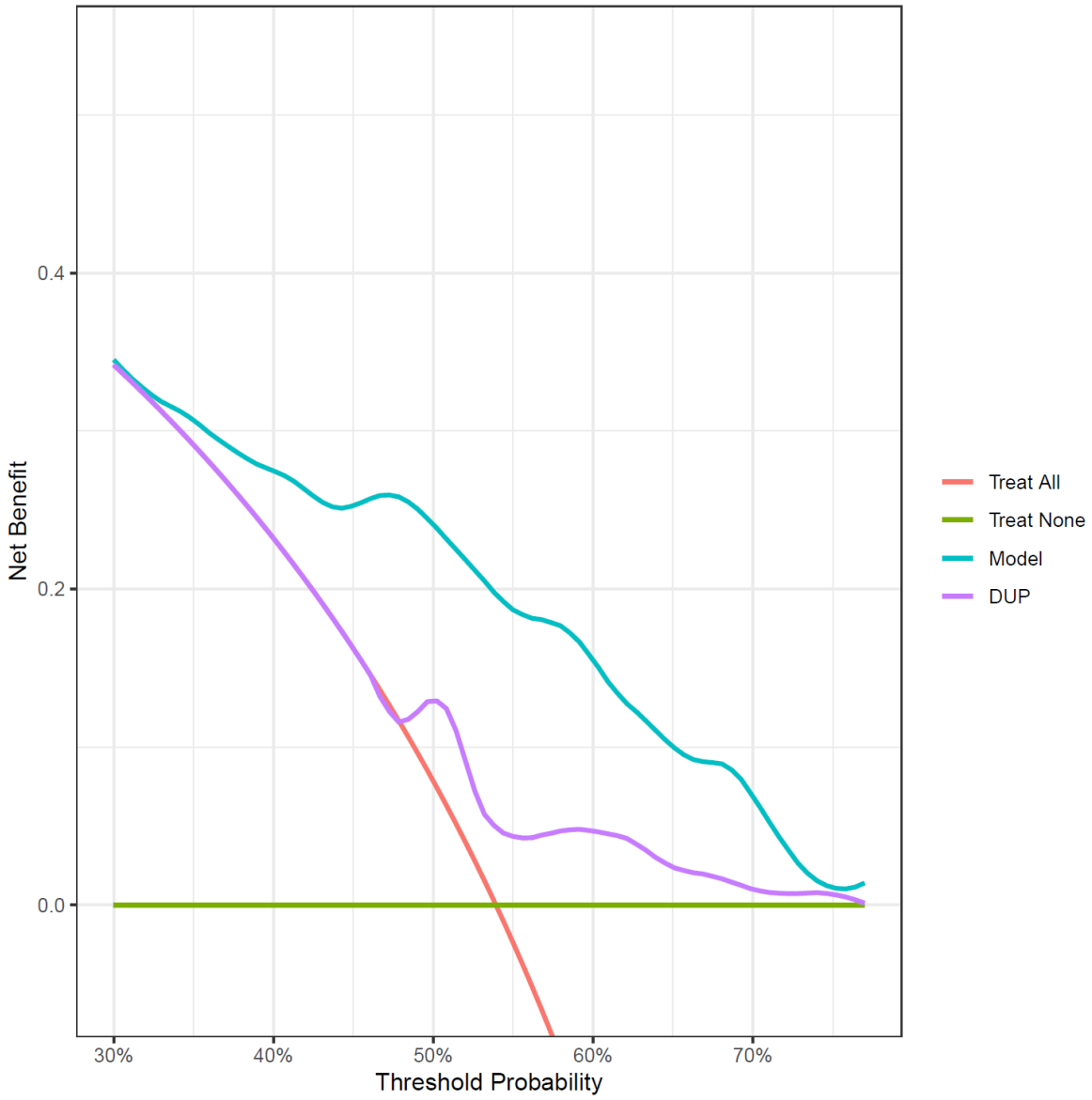


Figure 3-3 External validation decision curve analysis plot. Net-benefit is the treatment threshold weighted sum of true- minus false-positive classifications for each strategy plotted against an entire range of treatment thresholds. Green line: no patients are treated, net-benefit is zero (no true-positive and no false-positive classifications); red line: all patients are treated; purple and cyan lines: patients are treated if predictions exceed a threshold, with nonremission predictions based on adjusted DUP only, or on our prediction model. Between thresholds of 35% to 70%, treating based on our model is better than treating all, treating none or treating using DUP alone.

3.5 Discussion

We have developed a new risk prediction model based on baseline demographic and clinical variables to predict the risk of nonremission at one year after the onset of first-episode psychosis in a large sample of FEP individuals. The model was validated across services in the development population and externally validated on an independent cohort. The prediction model had fair discrimination and was fairly well-calibrated. The model showed an increase in net benefit.

3.5.1 Strengths and weakness of the study

Our study has some strengths. Both our development and validation cohorts included a representative sample of FEP participants from early intervention services in England, who were prospectively followed up for a year. Both the cohorts were assembled in similar services (early intervention) and periods in England which improves generalizability to patients within these services, though they have potentially changed in the past 10 years resulting from financial austerity measures. The candidate predictors and outcomes were measured using standardized instruments by graduate psychologists who were not directly involved in the care of the participants, which minimized the measurement bias. We used an operationalized, and well-established outcome definition for nonremission, which further minimizes measurement bias. We provided four measures of model performance—discrimination, two measures of calibration, and decision curve analysis (Steyerberg & Vergouwe, 2014). Though the baseline measures were meant to be measured on the first presentation to early intervention service, in practice, there was variation: in NEDEN 32% within three weeks of presentation; in Outlook 21% within three weeks. The model will apply to patients at least three weeks after their presentation to early intervention services.

There are some weaknesses to the study. Ethnicity was not included as a predictor in our model. This is some evidence that treatment resistance may be predicted by ethnicity (Smart et al., 2021). Only around half of the eligible participants consented to participate in the NEDEN study, which may affect the generalizability of our models to the general FEP population. However, those who did not consent were largely similar at baseline to people who did (Birchwood et al., 2014). Outcome data was not available for 34% of the NEDEN cohort and 52% in the Outlook cohort, which could further limit the validity of the results. As a result, while the number of events in the validation cohort was >100 (103), the number of nonevents was <100 (88). This is slightly less than suggested criteria. The method used for imputation of missing predictors using all the available data including outcome data and deletion of imputed outcome data has the advantage of the predictor imputation benefitting from the full data structure, whilst protecting the regression estimates from often problematic outcome imputations (Steyerberg, 2019). This approach has been

subject to criticism (Sullivan et al., 2015), though it is recognized that outcome imputation remains controversial. Further, there were differences in rates of missing data between the development and validation datasets. The outcome was measured only at the six- and 12-month time points. Study subjects may not have met remission criteria for the entire six months in between. The cohorts did not collect biomarkers of illness including inflammatory or neuroimaging data which a previous study in clinical high-risk populations has found to increase prognostic certainty when added to models based on clinical variables (Koutsouleris et al., 2018). Another weakness is that we have not accounted for treatment effects, which can lead to suboptimal model performance, albeit only in the presence of strong treatment effects. We assumed that standardized treatment was provided to all participants as they were drawn from early intervention services.

3.5.2 Comparison with previous studies

Three prediction models for outcome in first-episode psychosis have been reported, though they are yet to be used in clinical practice. One study has examined the prediction of social recovery in FEP participants from a randomised controlled trial (Koutsouleris et al., 2016), while the other two studies have examined prediction for remission and recovery measures in cohort studies (Leighton, Krishnadas, et al., 2019; Leighton, Upthegrove, et al., 2019). The discrimination performance for the remission outcome in our study is higher than that reported for models in two previous studies (c-statistic of 0.63 on external validation (Leighton, Krishnadas, et al., 2019), and 0.70 and 0.61 on internal and external validation respectively (Leighton, Upthegrove, et al., 2019)), which could be explained by the smaller sample sizes used in their development and validation (Leighton, Krishnadas, et al., 2019), and the higher number of predictors used for their model development (Leighton, Upthegrove, et al., 2019). Measures of calibration and clinical usefulness have not been provided by the other two studies, which adds to the novelty and importance of the current study.

3.5.3 Implications for clinicians and policymakers

The early identification of FEP individuals with higher risk prediction of nonremission may allow for changes to their treatment strategies, leading to improved remission rates. Though suggestions for such a strategy to improve remission rates have been made previously, there have been limited attempts towards a targeted approach to identify FEP individuals at high risk of nonremission.

Health services globally has introduced measures to improve access to services and to ensure that FEP individuals receive evidence-based care. A validated prediction model closely aligns with the policy agenda of early identification of FEP individuals with a high risk of nonremission so that their care can be optimized, and resources targeted according to need.

3.5.4 Future research

Prospective validation in additional cohorts from plausibly related settings is required to establish the utility of our model in clinical settings. This will help to compare the model predictions versus clinical intuition and address the issue of treatment effect. Future research also needs to address what biomarkers, such as neuroimaging and immune markers, add to the performance of the model. The model should be validated in a range of clinical settings for its use in services outside England, and in settings that do not have early intervention psychosis services, which may show a need for local updating to improve the accuracy of predictions for specific settings.

Chapter 4 Prediction modelling in first episode psychosis: an assessment of biological disease markers and machine learning classifiers

4.1 Overview of this chapter

This chapter presents an internal validation analysis of the potential for biological disease markers and machine learning classifiers in prediction modelling in first episode psychosis. A manuscript is in preparation.

4.2 Introduction

Psychosis is a major mental illness characterised by hallucinations, delusions and thought disorder. On average, more than 40% fail to achieve symptomatic remission (Lally et al., 2017). However, this average is often a poor guide for individual patients. At present, clinicians struggle to predict outcome in individuals and consequently an empirical approach to care is taken with all patients receiving the same treatment (National Institute for Health and Care Excellence (NICE), 2014). In contrast, the goal of clinical prediction model research is to enable the targeting of treatments according to individual patient risk. Expensive or higher risk treatment may be reserved for those at higher risk. This is the basis of stratified medicine (Hingorani et al., 2013; Steyerberg et al., 2013). We have recently published an external validation study of a binary logistic regression first episode psychosis (FEP) nonremission risk prediction model fit by maximum likelihood estimation (MLE) built using solely clinical variables (Leighton et al., 2021).

The majority of clinical prediction modelling studies for binary outcomes are based on logistic regression. In our recent systematic review of prediction models in FEP, only two of the 13 studies included used machine learning to develop prediction models while the rest used logistic regression (either MLE or regularised MLE) (Lee et al., 2022). However, across the field there is a growing interest in machine learning methods which promise to better capture nonlinearity and model complex interactions in medical data (van der Ploeg et al., 2016). Yet, such flexible machine learning models especially have been demonstrated to show particular issues with calibration in spite of good

discrimination performance (van der Ploeg et al., 2016; van Hoorde et al., 2015).

Across the field of psychiatry, there is a drive to find novel biomarkers of disease state or trait (Carvalho et al., 2020). However, the potential of such biological variables to improve clinical prediction model performance in FEP has not been adequately explored (Lee et al., 2022). Compelling evidence points to an association between inflammation and psychosis. Recent meta-analysis has highlighted the potential of inflammatory ratios, especially the monocyte/lymphocyte ratio (MLR) and neutrophil/lymphocyte ratio (NLR), as a biomarker in psychosis (Mazza et al., 2020). Inflammation is in turn linked to increased glutamate expression in the brain via the upregulation of the enzyme indoleamine 2, 3-dioxygenase (Schwarcz et al., 2012). Meta-analysis of magnetic resonance spectroscopy (MRS) glutamate measures has shown potential as a putative biomarker in psychotic disorders including schizophrenia (Merritt et al., 2016).

Our study has two aims: 1) to compare the discrimination and calibration performance of a binary MLE logistic regression FEP nonremission risk prediction model built using solely clinical variables to models built using putative peripheral inflammatory and MRS glutamate disease biomarkers in addition to clinical variables; 2) to compare the performance of FEP nonremission risk prediction models derived by logistic regression fit using MLE and elastic net to models built using machine learning methods including naïve Bayes, random forest, linear and radial support vector machines (SVMs).

4.3 Methods

4.3.1 Data source and study population

In this post-hoc analysis, we used data from the Lilly F1D-MC-HGDH trial. Lilly has not contributed to or approved, and is not in any way responsible for, the contents of this publication. The F1D-MC-HGDH trial was a double-blind, multicentre, randomised controlled trial of Olanzapine versus Haloperidol treatment in 263 participants meeting diagnostic criteria for a FEP (including schizophrenia, schizophreniform disorder, and schizoaffective disorder) as

defined by the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). The trial was carried out in 14 international centres between March 1997 and July 2001.

Participants were enrolled in the study if they met the following inclusion and exclusion criteria as outlined previously (Lieberman et al., 2005): age 16 to 40 years; onset of psychosis before age 35; DSM-IV diagnosis of schizophrenia, schizophreniform, or schizoaffective disorder; no more than 16 weeks previous antipsychotic treatment or clozapine treatment; no current substance dependence (except caffeine and nicotine) by DSM-IV within one month before study entry; no current serious suicidal risk; female participants not pregnant or breastfeeding; premorbid intelligence quotient of 70 or higher; no requirement for concurrent anticonvulsants, benzodiazepines (except for agitation and control of extrapyramidal symptoms), antidepressants, psychostimulants, or other antipsychotic medications at study entry; and no contraindication for neuroimaging. Each participant (or their authorised legal representative) had to understand the nature of the study and sign an informed consent document. Each site's institutional review board approved the study.

4.3.2 Outcome measure

Our outcome measure was symptom nonremission, defined as failing to meet the Remission in Schizophrenia Working Group criteria using the Positive And Negative Syndrome Scale (PANSS), at three months. The Remission in Schizophrenia Working Group defined remission as scores of less than or equal to three in PANSS items P1 Delusions, P2 Conceptual Disorganization, P3 Hallucinatory Behavior, N1 Blunted Affect, N4 Apathetic Social Withdrawal, N6 Lack of Spontaneity and G9 Unusual Thought Content (Andreasen et al., 2005).

4.3.3 Candidate predictors

Clinical variables were chosen based on the literature and consensus between five psychiatrists working in the field of early intervention in psychosis, as outlined previously (Leighton et al., 2021). Putative peripheral inflammatory and MRS glutamate candidate variables were selected based on a review of the literature and their availability in the Lilly dataset. The peripheral inflammatory

biomarkers considered were peripheral blood MLR and NLR. Combined glutamate and glutamine signal (GLX) as a ratio to creatine (Cr) was measured at three MRS regions of interests (ROIs) – in the frontal lobe, the hippocampus and the basal ganglia in the left hemisphere. Ordinal variables were treated as continuous and due to sample size we did not attempt to model non-linearities in continuous variables.

4.3.4 Magnetic resonance spectroscopy

MRS scans were obtained at 1.5 Tesla. 1H-MRS single-voxel spectroscopy was used to assess proton metabolites in three ROI – the frontal lobe, hippocampus and basal ganglia in the left hemisphere. The combined glutamate and glutamine (GLX) signal was measured as a ratio relative to creatine (Cr). Single-voxel spectra were acquired using a Point RESolved Spectroscopic (PRESS) pulse sequence as part of the Probe P spectroscopy package. Each spectroscopic voxel was prescribed from three-dimensional high resolution axial images (Time to Echo (TE)/Repetition Time (TR) = 35-45 / 175 ms ; 256 x 128 matrix, field-of-view (FOV) = 22 cm ; 15mm slice thickness; one number of excitations (NEX)). The 1H-MRS sequence for the frontal cortex, basal ganglia and hippocampus were as follows: TR 1500ms, TE 45 ms (range 35 ms - 45 ms), voxel dimension 15 (superior-inferior) x 20 (right-left) x 20 (anterior-posterior) mm, 2048 points, 2500Hz (range 2000 Hz – 2500 Hz), NEX (256 minimum, 512 maximum).

4.3.5 Sample size

The original Lilly study was powered to detect a change in using the Positive and Negative Syndrome Scale (PANSS) scores. Our study is a post-hoc analysis. According to the Prediction model Risk Of Bias Assessment Tool (PROBAST) guidance, for prediction model development the number of events per variable (EPV) should be at least ten to minimise overfitting (Wolff et al., 2019). With 86 events, we considered models with seven clinical predictors with or without one additional biological variable. The seven clinical variables selected were those with the highest absolute standardised regression coefficients from our earlier study that were also present in the Lilly dataset – in descending order PANSS P3, Premorbid Adjustment Scale highest level of functioning, PANSS P2, PANSS N4, Global Assessment of Functioning (GAF) Symptoms, sex and PANSS G6 (Leighton

et al., 2021). Due to predictor availability in the Lilly dataset, we substituted GAF symptom score for the clinical global impression severity scale.

4.3.6 Missing data

For prediction modelling missing data were multiply imputed (m = 5) by chained equations using all predictors, auxiliary variables, and outcomes based on the assumption that data was missing at random. Imputed outcome data were then deleted. It is proposed that this strategy leads to more efficient estimates than an ordinary multiple imputation strategy while also protecting the estimates from problematic imputations in the outcome variable (von Hippel, 2007).

4.3.7 Statistical analysis

All statistical analyses were performed using R, Comprehensive R Archive Network (CRAN) version 4.1.0,(R Core Team, 2021) and code is available in Appendix 4.

For baseline demographic and clinical comparisons, independent t-tests were used for normally distributed data, with Welch's correction if the assumption of homogeneity of variances was not met. For data that was not normally distributed, Wilcoxon's rank sum test was used. For categorical variables, Pearson's chi-squared test was used. PANSS measures were mapped to van der Gaag's five-factor model for baseline comparisons (van der Gaag et al., 2006).

Prediction model performance was tested at internal validation by n (10) fold repeated (50) cross-validation. A model was built on nine tenths of the data and performance was estimated on the left out tenth. This was repeated 500 times. Where models required tuning of hyperparameters, cross validation was nested with the hyperparameter tuning performed on an inner n (10) fold repeated (five) cross-validation. This process was repeated five times for each multiple imputation. Estimates and statistical tests were pooled using Rubin's rules (Rubin, 1987).

The distribution of the probabilities at internal validation across all cross-validation folds was inspected using histograms on the first imputed dataset.

Discrimination was quantified by the concordance (c) statistic. For binary outcomes the c-statistic is equal to the area under the receiver operating characteristic (ROC) curve which plots the sensitivity versus 1-specificity for consecutive probability thresholds of the predicted risk. The c-statistic can be interpreted as the probability that a randomly selected participant with the outcome will be ranked higher than a randomly selected participant without the outcome (Steyerberg & Vergouwe, 2014; van Calster et al., 2019).

Calibration was assessed based on the logistic calibration framework first proposed by Cox in 1958 (Cox, 1958). Herein, calibration was assessed by regressing the observed binary outcomes (Y) on the log odds of the predictions (the linear predictor (LP)) with a logistic model: $\text{logit}(Y) = a + b_{LP} \times LP$. The coefficient b_{LP} is known as the calibration slope while the intercept a is interpreted as the calibration-in-the-large when the slope is fixed at unity ($a|b_{LP}=1$). The calibration slope b_{LP} is ideally 1 when a model is well calibrated. If $b_{LP} < 1$, this indicates that the prediction model overfits and its risk estimates are too extreme (high risks are overestimated and low risks are underestimated). If $b_{LP} > 1$, the prediction model tends to underfit the data and the opposite pattern is observed (i.e. its predictions are too modest). The calibration-in-the-large ($a|b_{LP}=1$) is ideally 0 indicating mean calibration whereby the observed event rate in a dataset equals the average of the predicted risks. Predicted risks are underestimated on average if $a|b_{LP}=1 > 0$ and overestimated if $a|b_{LP}=1 < 0$ (Huang et al., 2020; van Calster et al., 2019; van Calster et al., 2016). At internal validation, (by either cross-validation or bootstrap) the focus is usually on assessment of the calibration slope. This is because if model fitting is by standard statistical estimation methods such as maximum likelihood, the calibration-in-the-large is guaranteed to be ideal (van Calster et al., 2019; van Calster et al., 2016). However, this does not hold for some complex model fitting procedures like neural networks (Karhade et al., 2018). As such, we report both the calibration-in-the-large and the calibration slope for our internal validation study.

For our first analysis, we compared the performance of a MLE logistic regression prediction model built using solely clinical variables to models built with the addition of a candidate biological variable, as detailed above.

For our second analysis, we compared the performance of a prediction model built with seven clinical variables using to logistic regression fit by MLE and elastic net to a number of popular machine learning approaches using the 'caret' package (Kuhn, 2008). Elastic net logistic regression was fit by regularised MLE tuned over a grid of alpha and lambda hyperparameters using the 'glmnet' package (Friedman et al., 2010). Alpha controls the balance between ridge and lasso penalty and lambda controls the amount of penalty. Naïve Bayes is a probabilistic classifier based on applying Bayes' theorem with the "naïve" assumption of conditional independence between the predictor variables. Using the 'naivebayes' package we tuned whether numeric predictors were handled by assuming they follow Gaussian distributions or whether kernel density estimation was used to estimate their class-conditional distributions (Majka, 2020). Random forest is an ensemble classifier that builds and combines multiple decision trees. Using the 'party' package, we implemented random forest in a conditional inference framework, tuning how many variables are available to select while splitting a tree at each node (Hothorn et al., 2006). SVMs perform classification tasks by constructing hyperplanes which separate data into classes by maximising a margin which is the distance from the hyperplane to the nearest training point. We tuned the cost hyperparameter which determines how hard or soft the margin is. The choice of kernel determines the type of hyperplane. We used a linear kernel using the 'e1071' package (Meyer et al., 2021) and a radial kernel using the 'kernlab' package (Karatzoglou et al., 2004).

One-way analysis of variance (ANOVA) was used to compare the performance metrics of the models in each of the above analyses. The calibration-in-the-large and calibration slope have been shown to follow a normal distribution but to ensure normality the c-statistic was logit transformed before analysis as recommended by Snell *et al* (Snell et al., 2018). Tukey's post-hoc tests were applied. If data was heterogeneous Welch's ANOVA was used and Games-Howell post-hoc tests were applied which adjusts p-values using Tukey's method. In the absence of the ability to pool post-hoc tests using Rubin's rules within the R programming environment, the Median P Rule was applied across multiple imputations (Eekhout et al., 2017).

4.4 Results

4.4.1 Baseline comparisons

Of the total 263 participants, 168 had PANSS data at 3 months. As shown in Table 4-1, the remitters did not differ from non-remitters in age (W = 3164; p = 0.3), sex ($x^2(1) = 0.568$; $p = 0.451$) and smoking status ($x^2(1) = 1.278$; $p = 0.258$). However, they did differ in baseline PANSS scores. Those who remitted had lower PANSS positive ($t(166) = 2.4$; $p = 0.02$) and PANSS negative scores (W = 4677; $p = 0.0003$). There was no difference in baseline characteristics between participants randomised to haloperidol versus olanzapine therapy (Table 4-2).

Table 4-1 Baseline characteristics for remitters versus non-remitters

Baseline Characteristic	Non-remission at 3 months (N=86)	Remission at 3 months (N=82)	Statistic	*p*-value	Effect Size (95% CI)
Age (Years) Median (IQR)	22.8 (20.1, 26.9)	23.2 (21.1, 26.2)	W = 3164	0.3	r = 0.089 (0.005, 0.240)
Sex (Female) N (%)	13 (15.1)	16 (19.5)	$x^2(1)$ = 0.568	0.451	*OR* = 0.736 (0.301, 1.771)
Non-Smoker N (%)	41 (47.7)	32 (39.0)	$x^2(1)$ = 1.278	0.258	*OR* = 1.421 (0.737, 2.753)
PANSS Positive Mean (SD)	25.1 (4.20)	23.4 (4.49)	$t(166)$ = 2.4	0.02	d = 0.378 (0.060, 0.680)
PANSS Negative Median (IQR)	21.5 (16.0, 26.0)	17.5 (10.2, 22.0)	W = 4677	0.0003	r = 0.282 (0.140, 0.420)
PANSS Disorganisation Median (IQR)	30.0 (24.0, 34.0)	28.0 (22.0, 32.8)	W = 3948	0.2	r = 0.103 (0.005, 0.250)
PANSS Excitement Median (IQR)	17.5 (14.0, 20.8)	16.0 (14.0, 20.0)	W = 3852	0.3	r = 0.080 (0.003, 0.230)
PANSS Emotional Distress Median (IQR)	22.5 (20.0, 25.0)	22.0 (18.0, 25.0)	W = 3870	0.3	r = 0.085 (0.003, 0.230)

IQR – interquartile range; SD – standard deviation; PANSS – Positive and Negative Syndrome Scale; *W* – Wilcoxon rank sum test; χ^2 – Pearson's chi-squared test; *t* – independent t-test; 95% CI – 95% confidence intervals; *r* – Pearson's correlation coefficient; *OR* – odds ratio; *d* – Cohen's d; mean and SD are depicted where the data was normal; median and IQR depicted where the data was not normally distributed.

Table 4-2 Baseline characteristics for haloperidol versus olanzapine groups

Baseline Characteristic	Haloperidol Therapy (n=132)	Olanzapine Therapy (n=131)	Statistic	*p*-value	Effect Size (95% CI)
Age (Years) Median (IQR)	22.8 (20.3, 26.8)	22.7 (20.3, 25.9)	W = 9071	0.5	r = 0.042 (0.002, 0.170)
Sex (Female) N (%)	21 (15.9)	27 (20.6)	$x^2(1)$ = 0.974	0.324	*OR* = 0.730 (0.367, 1.433)
Non-Smoker N (%)	54 (40.9)	63 (48.1)	$x^2(1)$ = 1.373	0.241	*OR* = 0.748 (0.456, 1.253)
PANSS Positive Median (IQR) Missing N (%)	24.0 (21.0, 28.0) 0 (0)	23.5 (20.0, 27.0) 1 (0.8)	W = 9418	0.2	r = 0.085 (0.006, 0.210)
PANSS Negative Mean (SD) Missing N (%)	20.3 (8.32) 0 (0)	19.2 (6.85) 1 (0.8)	$t(252)$ = 1.1	0.3	d = 0.140 (-0.110, 0.380)
PANSS Disorganisation Median (IQR) Missing N (%)	29.0 (22.0, 33.0) 0 (0)	28.0 (23.0, 33.0) 1 (0.8)	W = 8396	0.8	r = 0.019 (0.002, 0.150)
PANSS Excitement Median (IQR) Missing N (%)	18.0 (13.8, 21.0) 0 (0)	17.0 (14.0, 20.0) 1 (0.8)	W = 9230	0.3	r = 0.066 (0.005, 0.180)
PANSS Emotional Distress Median (IQR) Missing N (%)	22.0 (19.0, 26.0) 0 (0)	22.0 (19.0, 25.8) 1 (0.8)	W = 8886	0.6	r = 0.031 (0.003, 0.160)

IQR – interquartile range; SD – standard deviation; PANSS – Positive and Negative Syndrome Scale; *W* – Wilcoxon rank sum test; χ^2 – Pearson's chi-squared test; *t* – independent t-test; 95% CI – 95% confidence intervals; *r* – Pearson's correlation coefficient; *OR* – odds ratio; *d* – Cohen's d; mean and SD are depicted where the data was normal; median and IQR depicted where the data was not normally distributed.

4.4.2 Analysis 1 – clinical variables vs clinical + biological variable models

Results are outlined in Table 4-3 and Figure 4-1. There was no significant effect of model type on the c-statistic ($F(5, 14.6) = 0.184$; $p = 0.9640$), calibration-in-the-large ($F(5, 176512.09) = 0.021$; $p = 0.9998$), and calibration slope ($F(5, 114.8) = 0.206$; $p = 0.9594$). Discrimination was between 0.66 and 0.67 for all models. Calibration-in-the-large was close to optimal ($a|b_{LP}=1 = 0$) for all models. The mean calibration slope was closest to optimal ($b_{LP} = 1$) for the original model with varying degrees of overfitting for the other models with the addition of a biological variable albeit this difference was not significant. The distributions of the predicted probabilities were broadly similar across all models each without obvious systemic issues with predicted probabilities (Figure 4-2**)**.

Table 4-3 Performance metrics for clinical +/- biological variable models

	Original 7 Clinical Variable Model	Original Model + Basal Ganglia GLX/Cr	Original Model + Frontal GLX/Cr	Original Model + Hippocampus GLX/Cr	Original Model + MLR	Original Model + NLR	Statistic	*p*-value
c-statistic (95% CI)	0.6703 (0.6559, 0.6843)	0.6666 (0.6492, 0.6835)	0.6760 (0.6551, 0.6963)	0.6656 (0.6420, 0.6884)	0.6621 (0.6461, 0.6776)	0.6651 (0.6507, 0.6792)	$F(5, 14.6) = 0.184$	0.9640
CITL (95% CI)	0.0061 (-0.0164, 0.0287)	0.0030 (-0.0200, 0.0259)	0.0041 (-0.0202, 0.0283)	0.0100 (-0.0180, 0.0381)	0.0067 (-0.0164, 0.0297)	0.0073 (-0.0157, 0.0303)	$F(5, 176512.09) = 0.021$	0.9998
Calibration Slope (95% CI)	0.9745 (0.8908, 1.058)	0.9513 (0.8568, 1.046)	0.9559 (0.8624, 1.049)	0.9274 (0.8200, 1.035)	0.9100 (0.8292, 0.9908)	0.9284 (0.8491, 1.008)	$F(5, 114.8) = 0.206$	0.9594

Performance metrics for each model with estimates combined using Rubin's rules across five multiple imputations. The c-statistic was logit transformed before estimates were combined and statistical testing. 95% CI – 95% confidence intervals; *F* – one-way independent ANOVA; CITL – calibration-in-the-large; GLX/Cr – combined glutamate and glutamine signal measured as a ratio relative to creatine; MLR – monocyte/lymphocyte ratio; NLR – neutrophil/lymphocyte ratio.

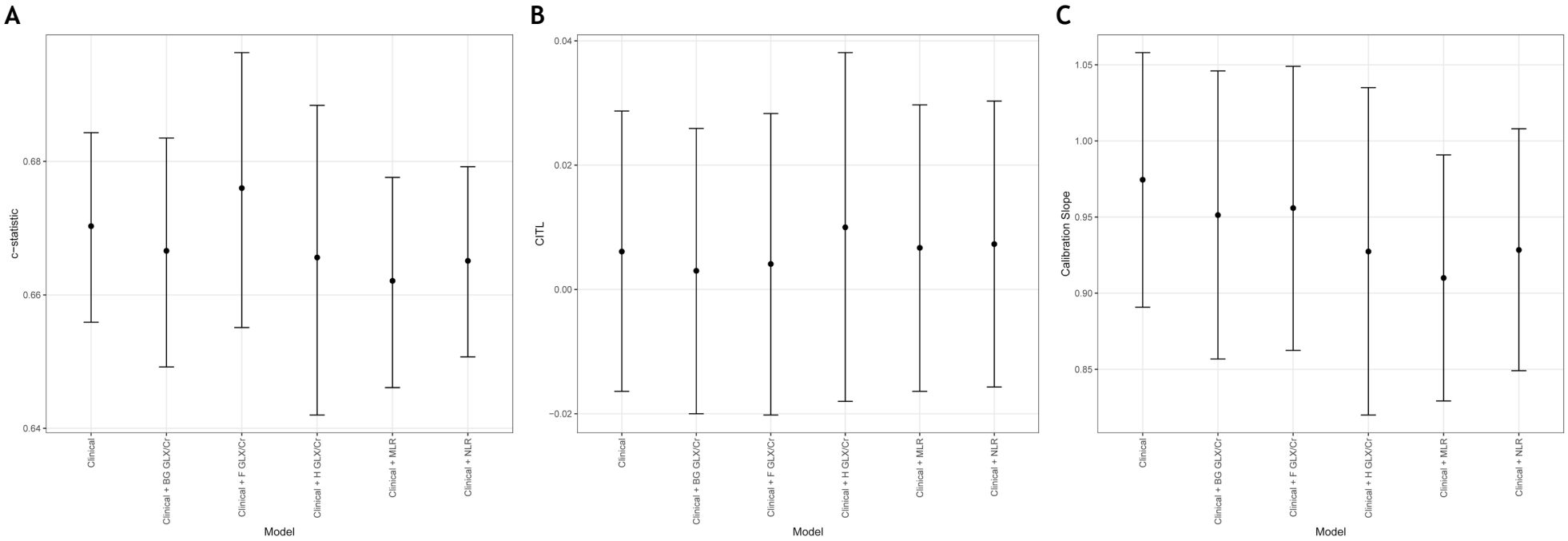


Figure 4-1 Performance metrics for clinical +/- biological variable models. A – c-statistic; B – Calibration-in-the-large; C – Calibration slope. BG – basal ganglia; F – frontal cortex; H – hippocampus; GLX/Cr – combined glutamate and glutamine signal measured as a ratio relative to creatine; MLR – monocyte/lymphocyte ratio; NLR – neutrophil/lymphocyte ratio.

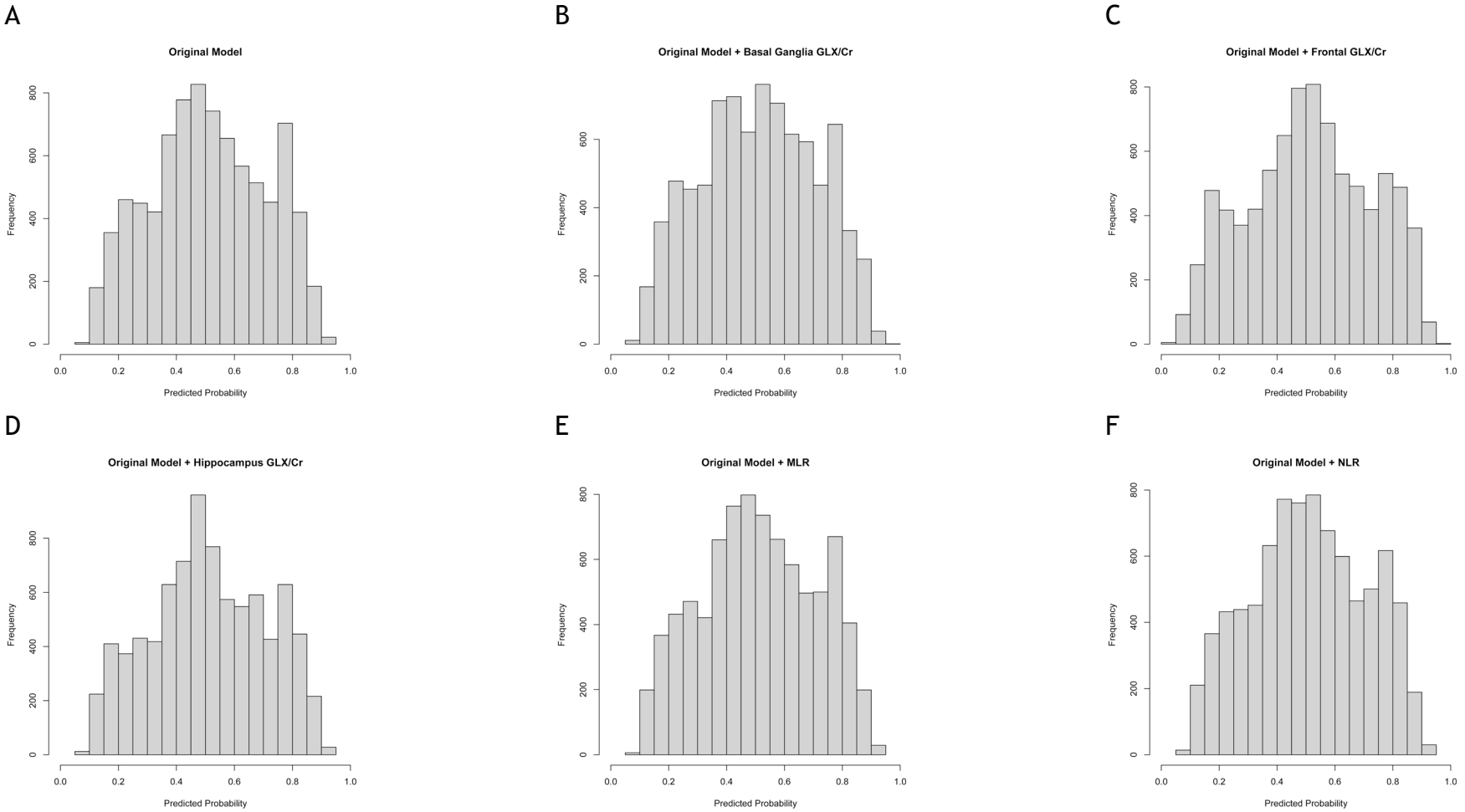


Figure 4-2 Distribution of probabilities at internal validation for clinical +/- biological variable models. A – original seven clinical variable model; B – original model + basal ganglia GLX/Cr; C – original model + frontal GLX/Cr; D – original model + hippocampus GLX/Cr; E – original model + MLR; F – original model + NLR. GLX/Cr – combined glutamate and glutamine signal measured as a ratio relative to creatine; MLR – monocyte/lymphocyte ratio; NLR – neutrophil/lymphocyte ratio.

4.4.3 Analysis 2 – MLE & elastic net logistic regression vs machine learning models

Results are outlined in Table 4-4 and Figure 4-3. There was a significant effect of model type on the c-statistic ($F(5, 398.66) = 7.225$; $p = <0.0001$). Post-hoc Tukey's tests showed that compared to MLE logistic regression but not elastic net logistic regression, Naïve Bayes had significantly improved discrimination. In addition, linear and radial SVM had significantly worse discrimination than elastic net logistic regression.

The was a significant effect of model type on calibration-in-the-large ($F(5, 5.87) = 56.356$; $p = <0.0001$). Post-hoc Games-Howell tests, showed that the calibration-in-the-large was significantly worse for Naïve Bayes and radial SVM compared to all other models.

There was no significant effect of model type on calibration slope ($F(5, 218.1) = 1.139$; $p = 0.3405$). However, MLE logistic regression was the only model with a mean slope close to optimal ($b_{LP} = 1$) and with a low variance across cross validation folds. The variance of the calibration slope was much larger for elastic net logistic regression, random forest and linear SVM compared to the three other techniques.

Table 4-4 Performance metrics for MLE & elastic net logistic regression versus machine learning models.

	MLE LR	Elastic Net LR	Naïve Bayes	Random Forest	Linear SVM	Radial SVM	Statistic	*p*-value
c-statistic (95% CI)	0.6703 (0.6559, 0.6843)	0.6846 (0.6697, 0.6992)	0.6972 (0.6834, 0.7106)	0.6746 (0.6624, 0.6866)	0.6600 (0.6460, 0.6737)	0.6542 (0.6415, 0.6667)	$F(5, 398.66) = 7.225$	<0.0001
CITL (95% CI)	0.0061 (-0.0164, 0.0287)	-0.0002 (-0.0135, 0.0131)	0.4548 (0.3900, 0.5196)	0.0035 (-0.0087, 0.0156)	0.0074 (-0.0062, 0.0211)	0.0608 (0.0418, 0.0799)	$F(5, 5.87) = 56.356$	<0.0001
Calibration Slope (95% CI)	0.9745 (0.8908, 1.058)	3.994 (0.6304, 7.359)	0.7009 (0.6317, 0.7702)	5.203 (-5.720, 16.12)	10.89 (-21.72, 43.50)	1.737 (1.522, 1.953)	$F(5, 218.1) = 1.139$	0.3405

Performance metrics for each model with estimates combined using Rubin's rules across five multiple imputations. The c-statistic was logit transformed before estimates were combined and statistical testing. 95% CI – 95% confidence intervals; CITL – calibration-in-the-large; *F* – one-way independent ANOVA; LR – logistic regression; MLE – maximum likelihood estimation; SVM – support vector machine.

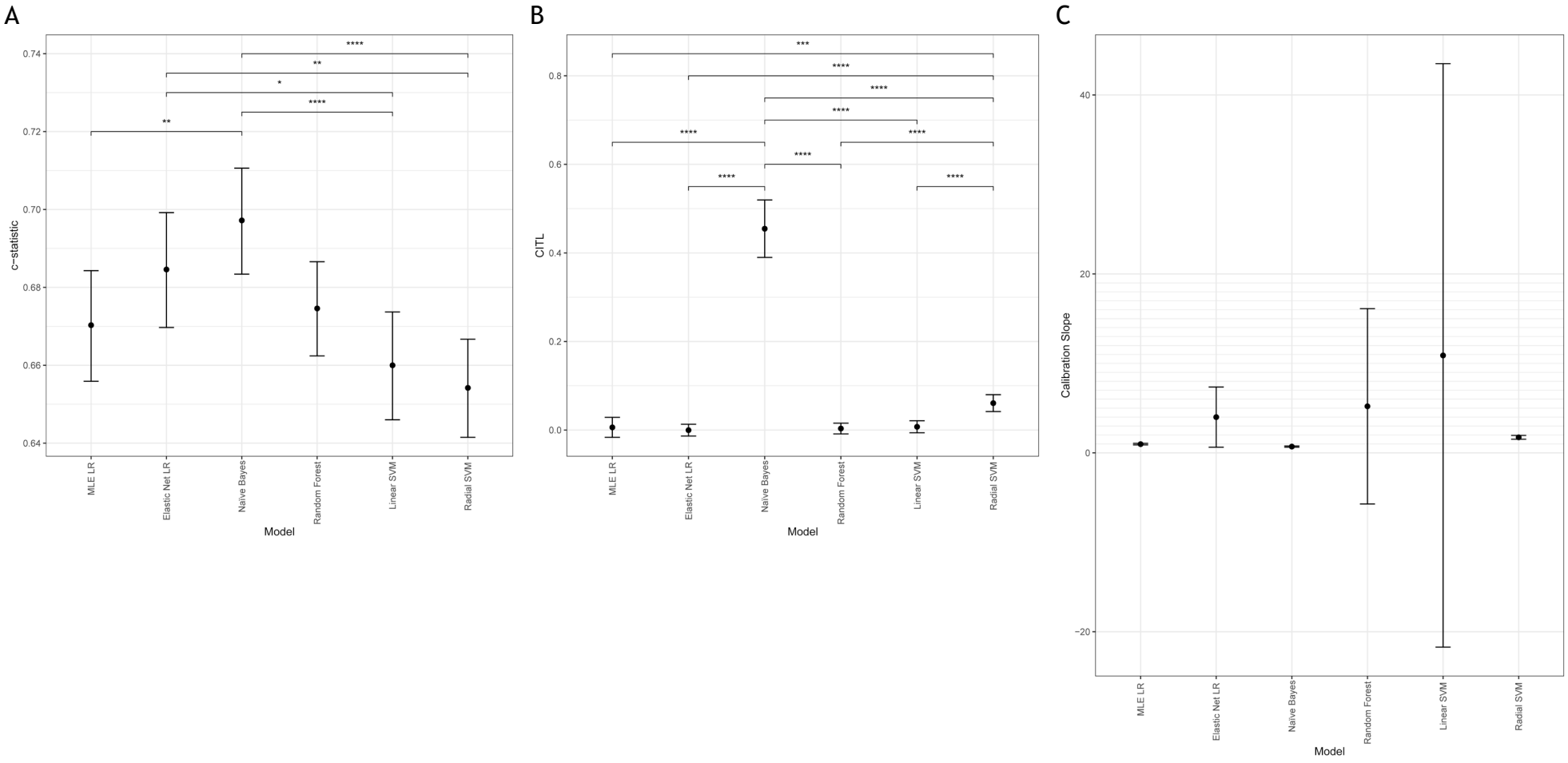


Figure 4-3 Performance metrics for MLE & elastic net logistic regression versus machine learning models. A – c-statistic; B – Calibration-in-the-large; C – Calibration slope. LR – logistic regression; MLE – maximum likelihood estimation; SVM – support vector machine.

The distributions of the predicted probabilities across all the cross-validation folds showed more modest predictions for the elastic net logistic regression, random forest, linear and radial SVM than MLE logistic regression while Naïve Bayes showed more extreme predictions. This is reflected by the mean calibration slopes which compared to MLE logistic regression indicated varying degrees of underfitting (b_{LP} = >1) on average for all other methods except Naïve Bayes which showed overfitting (b_{LP} = <1) on average. Naïve Bayes predicted probabilities also tended to be underestimated on average reflective of its mean calibration-in-the-large much greater than zero (Figure 4-4).

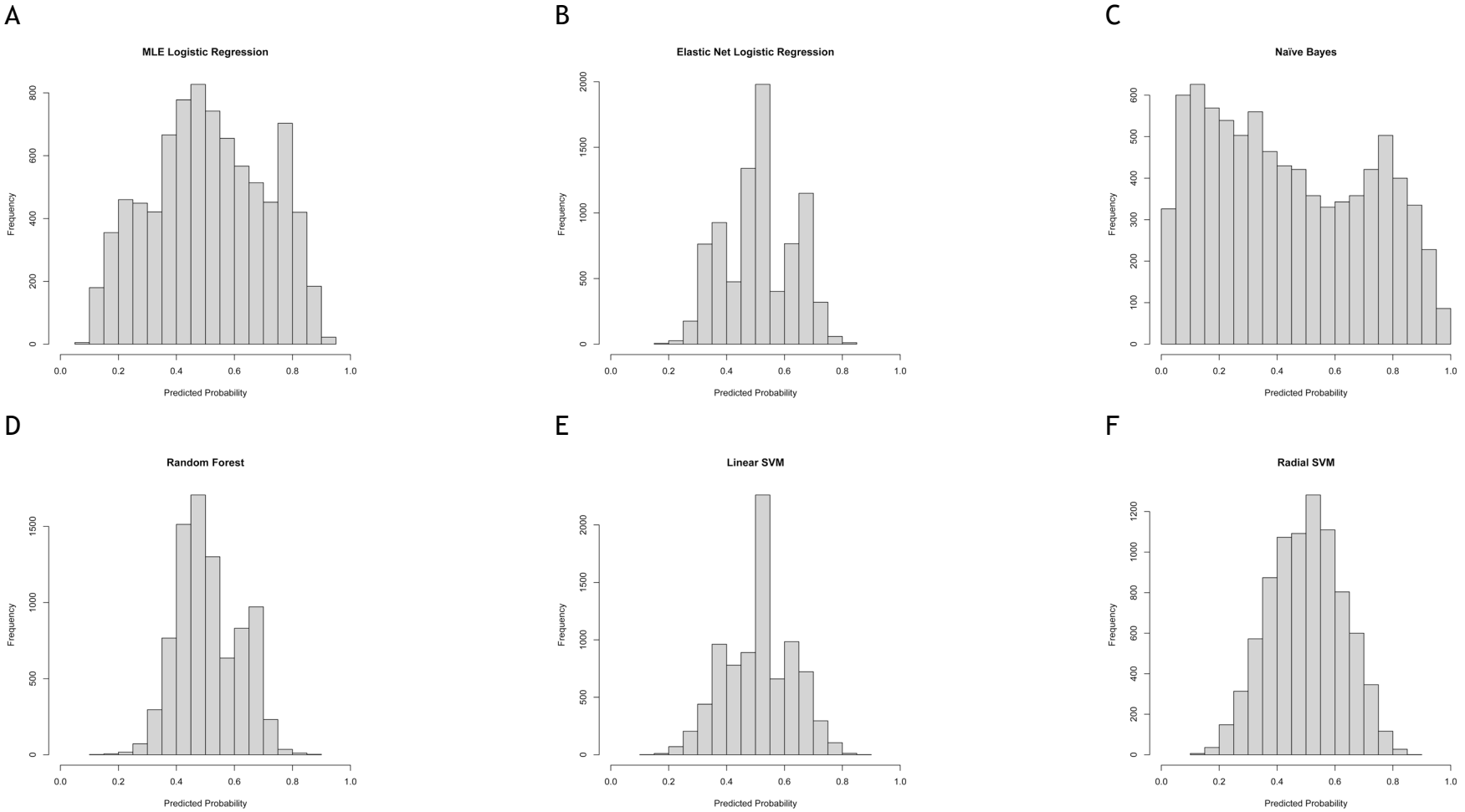


Figure 4-4 Distribution of probabilities at internal validation for MLE & elastic net logistic regression versus machine learning models. A – MLE logistic regression; B – Elastic net logistic regression; C – Naïve Bayes; D – Random forest; E – Linear SVM; F – Radial SVM. MLE – maximum likelihood estimation; SVM – support vector machine.

4.5 Discussion

This is the first study to directly compare machine learning methods to logistic regression in a first episode psychosis population. Further, we are only aware of one other first episode psychosis study that considered biological variables in addition to clinical and demographic predictors to predict outcome in a psychosis population. In that study of 523 patients, the biological variables (genetic) did not add value (de Nijs et al., 2021).

Our results from analysis 1, suggest that that models with an additional biological variable were no better than models with clinical variables alone in terms of discrimination or calibration. In analysis 2, we demonstrate that the Naïve Bayes machine learning model was better than MLE but not elastic net logistic regression in terms of discrimination. However, for all models except MLE logistic regression there were problems with calibration. Naïve Bayes had a mean calibration-in-the-large of greater than zero. This was reflected in the distribution of its predictions which tended to underestimate risk. This would lead to undertreatment if the model was deployed to a clinical population. Radial SVM also had a calibration-in-the-large greater than zero albeit not so extreme as Naïve Bayes. Of all the models, the only calibration slope with a mean close to the ideal and low variance across cross-validation folds was MLE logistic regression.

Most clinical prediction modelling studies only focus on discrimination as a measure of model performance. Discrimination is how well a prediction model can distinguish between those with an outcome and those without. A model that discriminates well should give higher risk estimates to patients with the outcome than those without (Steyerberg & Vergouwe, 2014; van Calster et al., 2019). Calibration is an important aspect of model performance that is often overlooked. It is the agreement between the observed and predicted risks. Specifically, a model is well calibrated if the event rate is X% among patients with a predicted risk of X%. For example, if we predict a 10% risk that a patient will relapse from a disease, the observed proportion should be 10 relapses per 100 patients with such a prediction (Steyerberg & Vergouwe, 2014; van Calster et al., 2016). Calibration is a measure of how reliable a model's predictions are. This is especially important in a clinical context because treatment decisions are

often based on whether a patient's predicted risk meets or exceeds a specific threshold. For example, the National Institute for Health and Care Excellence (NICE) recommend the use of the QRISK algorithm for quantifying cardiovascular risk. A medication (a statin) is recommended for patients whose risk exceeds 10% (Hippisley-Cox et al., 2008; National Institute for Health and Care Excellence (NICE), 2016a). It is vital that the predicted risk estimates of QRISK are reliable – that the observed proportion of patients with the outcome at 10% is indeed 10 in 100. Otherwise, there could result in over or undertreatment, each with its associated harm. Despite its clear importance for understanding the utility of a prediction, calibration is frequently underreported. In our recent systematic review of prediction models in FEP, we found that while 54% of studies reported discrimination only 38% reported any measure of calibration (Lee et al., 2022).

Limited previous research has examined the calibration of machine learning models in comparison to logistic regression in psychiatric populations. Perlis compared calibration in models to predict treatment resistance in depression and showed that MLE logistic regression models were better calibrated than machine learning approaches (naïve Bayes, random forest and SVM). The author hypothesised that this is reflective of the fact that machine learning approaches train models to maximise discrimination regardless of calibration (Perlis, 2013). Lindheim et al compared MLE logistic regression to three machine learning methods (naïve Bayes, classification and regression trees and random forest) in an internal validation study to predict the likelihood of a bipolar diagnosis at screening. Logistic regression was shown to have superior discrimination and calibration at internal validation (Lindhiem et al., 2020). More generally, a 2019 systematic review of 71 studies comparing the performance of clinical prediction models built using logistic regression (either MLE or regularised MLE) to those using machine learning did not find a performance benefit of machine learning in terms of discrimination when excluding studies at high risk of bias. Of the 71 studies, calibration was only reported in 21% and just four studies reported the calibration slope and or intercept. Only three studies in the review came from the field of psychiatry, none of which assessed calibration (Christodoulou et al., 2019). In our systematic review of 13 FEP prediction modelling studies, only two used machine learning techniques while 11 used logistic regression (either MLE or regularised MLE). We were unable to make any comparisons between the

performance of machine learning models and logistic regression as the studies using machine learning did not report any measure of discrimination or calibration performance (Lee et al., 2022).

4.5.1 Strengths and limitations

Our study has several strengths. The present analysis is the first time the data from the original Lilly trial has been used to develop prediction models. It is a unique dataset with longitudinal follow up in FEP and measurement of both clinical and biological variables. An earlier study used the data to develop a simple prediction rule based on PANSS symptom score change at two weeks to predict 12-week symptom response but did not develop a prediction model. Further, this study did not report any measures of discrimination or calibration (Stauffer et al., 2011). In terms of our methodology, we employed a robust 10 fold 50 times repeated internal validation procedure which was nested when tuning hyperparameters to guard against over fitting. In addition, we used multiple imputation. Compared to single imputation or complete case analysis, multiple imputation results in less biased parameter estimates and narrows the uncertainty around missing values by generating several imputations. Finally, we considered both discrimination and calibration performance when assessing our prediction models.

Our study has several limitations. The assessment of logistic calibration typically requires a sample size of 100 events and 100 non-events (van Calster et al., 2016). We had a slightly smaller sample size with 86 events and 82 non-events with an EPV of 12. In addition, simulation studies have suggested that machine learning techniques including SVM, neural networks and random forest require substantially higher EPVs (often >200) to mitigate overfitting and optimism in model performance (van der Ploeg et al., 2014). Further, good calibration metrics using the logistic calibration framework can still be derived from models which are poorly calibrated if particular probability regions are miscalibrated (Huang et al., 2020). Flexible calibration curves (which plot the predicted probabilities on the x-axis against the actual observed proportions on the y-axis using a smoothing technique) allow assessment of probability regions for miscalibration, but require larger sample sizes of at least 200 events and 200 non-events (van Calster et al., 2016). The original Lilly trial measured a limited

number of inflammatory biomarkers assessed and MRS was only measured at 1.5 Tesla. Many biomarkers with a greater evidence base for disturbance in psychosis were not available for our current analysis. Moreover, much of the evidence for glutamate dysfunction in psychosis comes from more modern MRS studies at higher field strengths (Merritt et al., 2016). The exact coordinates of the MRS voxels in the three ROIs in the left hemisphere were not detailed when the Lilly trial data was provided for analysis. Further, variation in scanner across sites may impact the precision of the results. Finally, mechanisms to recalibrate poorly calibrated models exist including logistic recalibration where the intercept is updated and existing coefficients reweighted, Platt scaling, isotonic regression and Bayesian Binning into Quantiles but this was out with the scope of our current analysis (Huang et al., 2020).

4.5.2 Conclusions

In our prediction modelling analysis, we show that the addition of a biological variable does not improve the performance of a logistic regression model built using clinical variables in this dataset. Further, we demonstrate that machine learning or elastic net logistic regression did not result in improved global performance compared to MLE logistic regression.

Chapter 5 Delirium and the risk of developing dementia: a cohort study of 12949 patients

5.1 Overview of this chapter

The central theme of my PhD thesis is prognosis. Prognosis is the determination of risk of future health outcomes in people with a given health condition. Prognostic research is of considerable importance. Globally, there are more people living with health conditions than ever before. Prognostic research seeks to improve the outcomes of people living with health conditions. However, there is a disparity between the potential and actual impact of prognostic research. In response, initiatives like the PROGnosis RESearch Strategy (PROGRESS) have established standards for higher quality prognostic research. PROGRESS centres on four themes: fundamental prognosis research, which investigates the course of health conditions in the context of their current care, prognostic factor research, which looks at specific factors associated with prognosis, prognostic model research, which is concerned with the development, validation and impact of models incorporating multiple prognostic factors, and, stratified medicine research, which focuses on the use of prognostic information to tailor treatments to individuals according to their predicted risk (Hemingway et al., 2013; Hingorani et al., 2013; Riley et al., 2013; Steyerberg et al., 2013).

In recent years, prognosis research has benefitted from the exponential growth in the availability of routinely collected health data. Its use is a key recommendation from PROGRESS. Routinely collected data is data collected without a specific a priori research question. Sources of routinely collect data include disease registries and electronic healthcare records (Benchimol et al., 2015). It is extremely valuable for prognosis research because it allows pragmatic cost-effective research to be conducted in an entirely naturalistic clinical setting, with much larger numbers of participants. In addition, routinely collected data enables the researcher to circumvent any issues with selection bias.

A principal aim of my PhD fellowship was to be the establishment of the Electronic Measures in Psychosis – Assessing Trajectory and Health-Outcomes (EMPATH) platform for the collection of outcome data for first episode psychosis

patients treated withing Esteem NHS Greater Glasgow & Clyde (GG&C) early intervention in psychosis service. EMPATH aims operationalise the routine collection of standard outcome measures within the service. Data collected within EMPATH will be securely deposited within NHS GG&C's Safe Haven which allows research use of linked unconsented routinely collected datasets. I plan to use this source of routinely collected data to undertake prognosis research into patients with a first episode of psychosis. Unfortunately, there have been significant delays in developing and deploying the EMPATH platform as a consequence of the global coronavirus pandemic such that deployment was postponed until 2023. Routinely collected data will be available for prognosis research from 2024 which is after the end of my PhD fellowship.

In order to establish the feasibility of using routinely collected data from NHS GG&C for prognosis research, in advance of accessing routinely collected data from EMPATH, I looked at another clinical field with unanswered prognosis questions: delirium and the risk of subsequent dementia. This work falls under the first PROGRESS theme, fundamental prognosis research, and second PROGRESS theme, prognostic factor research. The data for this study came from the West of Scotland Safe Haven (Hemingway et al., 2013).

5.2 Introduction

Delirium and dementia are two of the most common causes of cognitive impairment in the elderly population, but their interrelationship is poorly understood (Fong et al., 2015). Dementia is characterised by an irreversible progressive global cognitive decline. It is associated with huge financial and wider societal costs. In the UK, the annual cost of dementia is £35 billion, two-thirds of which is borne by people with dementia and their families (Alzheimer's Society, 2020). Delirium is characterised by an acute and fluctuating disturbance in attention and awareness with associated disturbance in cognition (e.g., memory deficit, disorientation, language, visuospatial ability or perception), which cannot be explained by another neurocognitive disorder and does not occur in the context of a severely reduced level of arousal, such as coma. It is a serious and life-threatening neuropsychiatric syndrome, which is a direct physiological consequence of another medical condition, substance intoxication or withdrawal, toxins or multiple aetiologies (Slooter et al., 2020). Delirium is

very common in the elderly and present in up to 50% of patients over the age of 65 admitted to hospital (Inouye et al., 2014). Delirium is a clinical diagnosis, which is often under-recognised and frequently overlooked. This has led to a number of high-profile campaigns to increase the awareness and recognition of delirium across the UK and the wider world (Khachaturian et al., 2020).

Dementia is the primary risk factor for delirium and delirium is a major risk factor for subsequent dementia (Fong et al., 2015; Jackson et al., 2017). It is not yet clear if delirium is a simply a marker of brain vulnerability, whether the impact of delirium on dementia is derived from its precipitating cause or whether delirium itself leads to permanent neuronal damage. Delirium is preventable in 30%–40% of cases and is, therefore, an important modifiable risk factor for dementia (Inouye et al., 2014).

Several clinical studies provide evidence to support the relationship between delirium and dementia. A 2010 meta-analysis of two studies (n=241) found that delirium was associated with an increased risk of dementia (*RR* 5.7, 95% CI 1.3 to 24.0) (Witlox et al., 2010). A 2021 meta-analysis of six studies (n=901) showed that delirium was associated with increased odds of developing new dementia compared with patients without delirium (*OR* 11.9, 95% CI 7.29 to 19.6) (Pereira et al., 2021). The relationship has also been explored in a small population-based cohort study of 553 individuals aged 85+, which found an increased risk of incident dementia following episode of delirium (*OR* 8.7, 95% CI 2.1 to 35) (Davis et al., 2012).

However, to date, the field lacks large studies with long-term follow-up of delirium in subjects initially free of dementia to clearly establish outcomes (Fong et al., 2015; Inouye et al., 2014).

Our study has two objectives:

1. To estimate the cumulative incidence of dementia among those who experience an episode of delirium but who have not yet been diagnosed with dementia prior to that episode.

2. To model the effect of age at delirium diagnosis, sex and socioeconomic deprivation on the rate of occurrence of dementia among those still at risk (i.e., the cause-specific hazard of dementia).

5.3 Methods

We adhere to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and the Reporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statements (Benchimol et al., 2015; Vandenbroucke et al., 2007).

We undertook a retrospective cohort study of patients over the age of 65 who had been diagnosed with an index episode of delirium but who had not been diagnosed with dementia prior to their index episode of delirium. Patients from the National Health Service (NHS) Greater Glasgow & Clyde (GG&C) health board were included. Patients with a diagnosis of delirium made before 1 May 2020 were included back as far as the records allowed. The earliest delirium diagnosis was 21 April 1996. Patients were followed from their first episode of delirium up until 1 October 2020 when the data were collected. The primary outcome event of interest was diagnosis of dementia. A competing event, death before dementia diagnosis, was observed. Patients who had not experienced either event before the end of the follow-up period were coded as censored. Patients who experienced their outcome event on the same day as their index delirium diagnosis were considered to have survived 0.5 days.

West of Scotland Safe Haven at NHS GG&C created the study population from the database population. The diagnoses of delirium and dementia were clinical diagnoses based on the International Classification of Diseases 10th Revision (ICD-10) made by the treating clinician (see Appendix 5). Diagnoses could have been made in accident and emergency (A&E), as an inpatient or outpatient or on death. Age at delirium diagnosis, sex and Scottish Index of Multiple Deprivation (SIMD) 2009 quintile (lowest equals most deprived) were included as covariates. SIMD 2009 was based on their most recent postal address. All subjects had information about covariates—there were no missing data. The total number of relevant delirium patients in the NHS GG&C Safe Haven database determined the sample size.

As outlined above, competing risks are present as a participant is at risk of two mutually exclusive events. Using the Kaplan-Meier estimate of the survival function to estimate the incidence function in the presence of competing risks generally results in upward biases in the estimation of the incidence function. Instead, we used the cumulative incidence function (CIF), which allows for the estimation of the incidence of the occurrence of an event (dementia) while taking competing risk (death without a dementia diagnosis) into account. The CIF for the *k*th cause is defined as: $CIF_k(t) = Pr(T \leq t, D = k)$, where D denotes the type of event that occurred, and T denotes the time from baseline time until the occurrence of the event. The function $CIF_k(t)$ denotes the probability of experiencing the *k*th event before time t and before the occurrence of a different type of event (Austin et al., 2016).

We also modelled the effect of covariates (age at incident delirium, sex and deprivation quintile) on the cause-specific hazard function. The cause-specific hazard function is the instantaneous rate of occurrence of the primary event (dementia) in subjects who have not yet experienced either event (dementia or death without dementia). The exponentiated regression coefficient from the cause-specific hazard model represents the amount of relative change in the cause-specific hazard function associated with a 1-unit change in the covariate. The cause-specific hazard model is well suited to studying the aetiology of a disease (Lau et al., 2009). We fit the cause-specific hazard model by estimating a Cox proportional hazards model and treating subjects who experience a competing event as being censored at the time of occurrence of the competing event. Post-model assumption testing included testing the proportional hazard's assumption via Schoenfeld residuals, using the difference in beta values (DFBETAS) to check for influential observations and assessing the functional form of covariates via Martingale residuals. Age had a non-linear functional form, so the final Cox model was refitted using a penalised cubic spline term for age. The results of our post-model assumption testing are available in Appendix 6.

All analyses were performed using R, CRAN V.4.0.0 (R Core Team, 2020) (with the 'survival' (Therneau, 2020; Therneau & Grambsch, 2000), 'cmprsk' (Gray, 2020), and 'survminer' (Kassambara et al., 2021) packages) and code is available in Appendix 7.

5.3.1 Ethics approval

The West of Scotland Safe Haven has ethical approval (17/WS/0237) to create a research database using routinely collected, un-consented patient data. Delegated research ethics approval was granted for linkage to NHS patient data by the Local Privacy and Advisory Committee at NHS GG&C under approval (GSH/18/AM/004).

5.4 Results

12949 patients with a relevant index episode of delirium followed up for an average of 741 days (minimum=0.5 days, maximum=8855 days) were included in the study. 3530 (27%) of these patients had a subsequent diagnosis of dementia and 5788 (45%) died without a diagnosis of dementia, leaving 3631 (28%) who were coded as censored by the study end date. The diagnosis of dementia was made on death in 643 (18%) of patient who were diagnosed with dementia. This information is summarised in Figure 5-1.

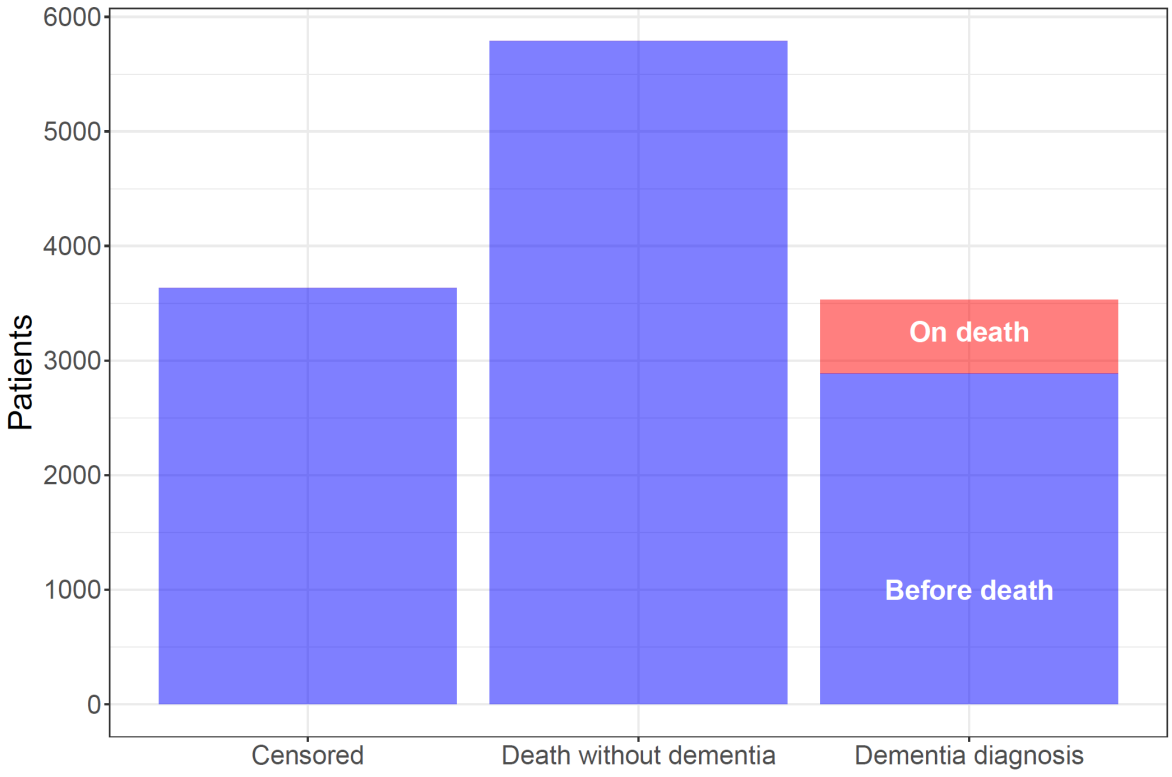


Figure 5-1 The outcomes for patients with an index episode of delirium follow-up for an average of 741 days (minimum = 0.5 days, maximum = 8855 days).

The diagnosis of new index episodes of delirium increased in frequency over time with some seasonal variation as per Figure 5-2.

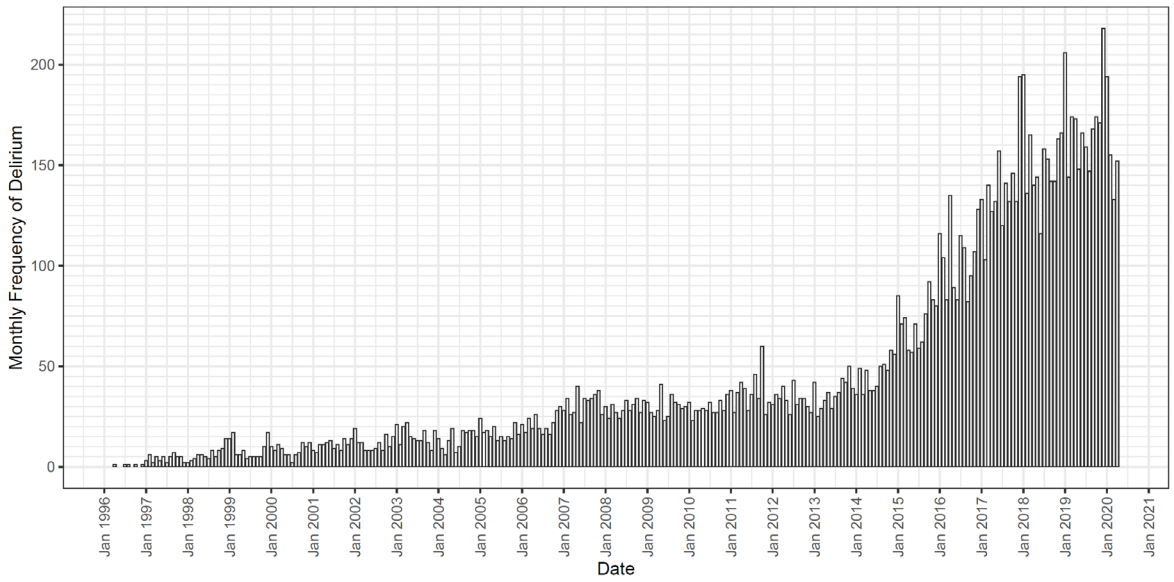


Figure 5-2 The monthly frequency of new index delirium diagnoses in patients who had not been diagnosed with dementia prior to this episode of delirium.

Descriptive statistics for the patients in the study are reported in Table 5-1.

Table 5-1 Descriptive statistics for all patients included in the study

Variable	Total Sample (n = 12949)	Dementia Diagnosis (n= 3530)	Death without a dementia diagnosis (n = 5788)
Age at index episode of delirium Mean (SD)	82.3 (7.8)	83.2 (7.0)	82.7 (8.1)
Male Sex No. (%)	5036 (39%)	1262 (36%)	2458 (42%)
SIMD 2009 Quintile No. (%)	1st – 4976 (38%) 2nd – 2341 (18%) 3rd – 1986 (15%) 4th – 1673 (13%) 5th – 1973 (15%)	1st – 1359 (38%) 2nd – 595 (17%) 3rd – 620 (18%) 4th – 515 (15%) 5th – 441 (12%)	1st – 2247 (39%) 2nd – 1050 (18%) 3rd – 885 (15%) 4th – 672 (12%) 5th – 934 (16%)

SIMD – Scottish Index of Multiple Deprivation.

The estimated cumulative incidences of dementia and for the competing risk of death without a dementia diagnosis are presented in Figure 5-3. The estimated cumulative incidence of dementia, accounting for the competing risk of death without a dementia diagnosis, was 9.0% (95% CI 8.5% to 9.5%) by six months, 13.6% (95% CI 13.0% to 14.2%) by a year, 31.0% (95% CI 30.1% to 31.9%) by five years, 35.5% (95% CI 34.5% to 36.5%) by 10 years, and 36.3% (95% CI 35.2% to

37.3%) by 20 years. The estimated cumulative incidence of the competing risk of death without a dementia diagnosis was 20.0% (95% CI 19.3% to 20.7%) by six months, 27.1% (95% CI 26.3% to 27.9%) by a year, 49.2% (95% CI 48.2% to 50.2%) by five years, 55.3% (95% CI 54.3% to 56.4%) by 10 years and 57.4% (95% CI 56.2% to 58.5%) by 20 years.

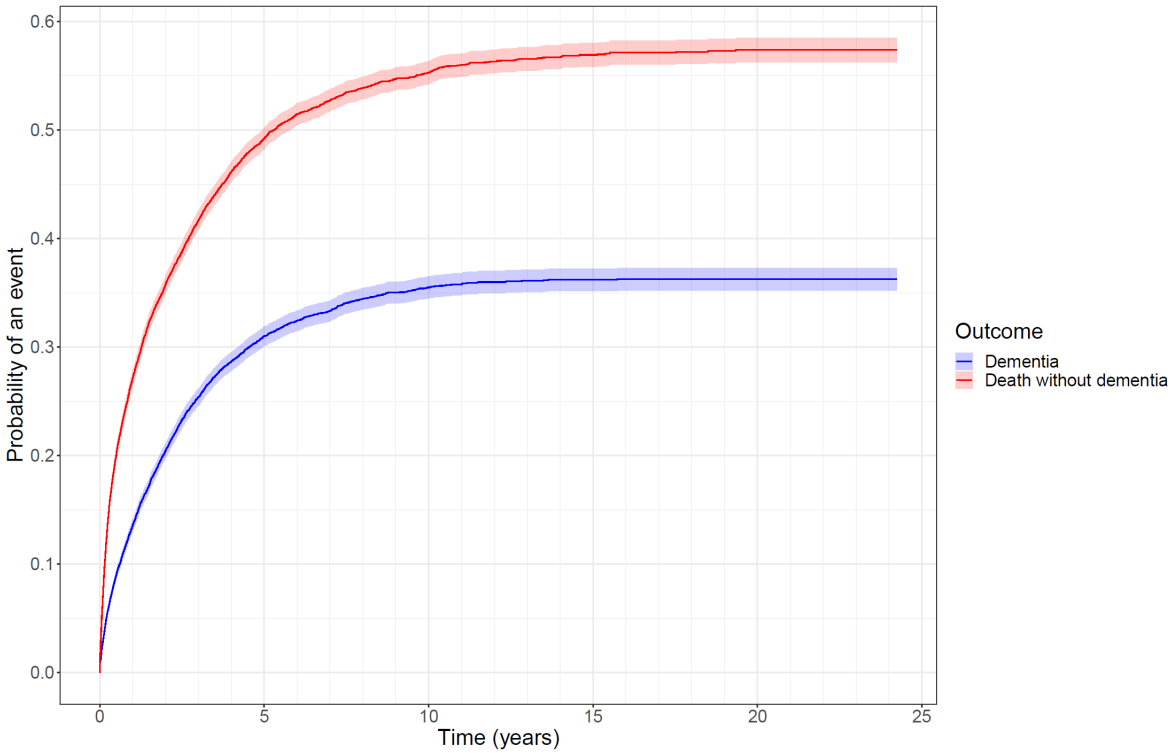


Figure 5-3 Cumulative incidence functions for dementia (blue) and for death without dementia (red) in patients with an index episode of delirium by time in years with 95% CIs.

The multivariable adjusted cause-specific hazard ratios for sex and SIMD 2009 deprivation quintile are illustrated in Figure 5-4.

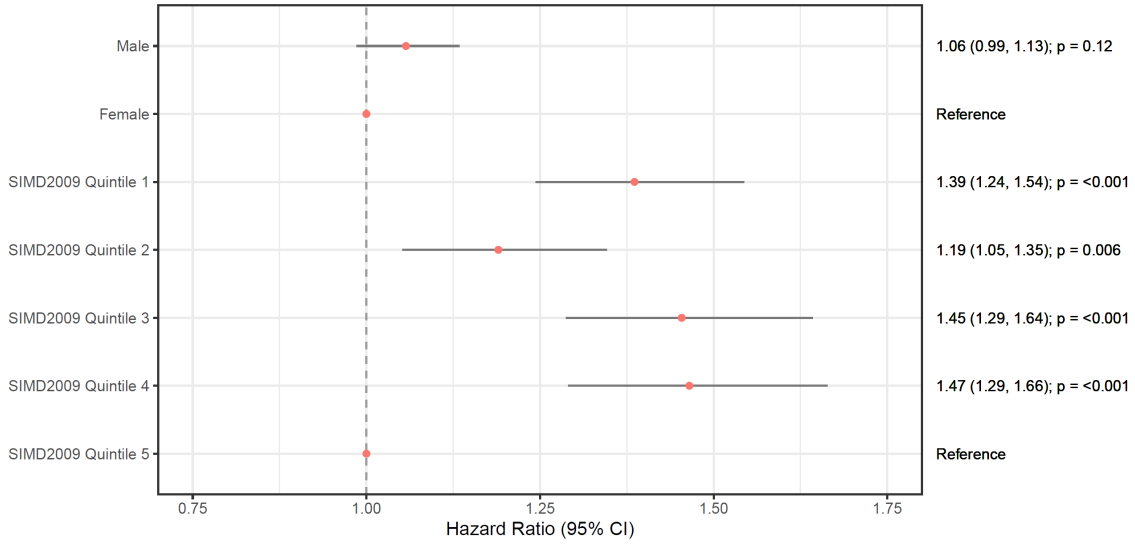


Figure 5-4 Multivariable adjusted cause-specific hazard ratios for dementia diagnosis in patients with an index episode of delirium. The cause-specific hazard ratios of the four most deprived SIMD 2009 quintiles are greater than the least deprived quintile (reference). There does not appear to be a relationship between sex and cause-specific hazard of dementia in patients with an index episode of delirium. SIMD – Scottish Index of Multiple Deprivation.

The multivariable-adjusted cause-specific hazard ratios for age at delirium diagnosis is illustrated in Figure 5-5.

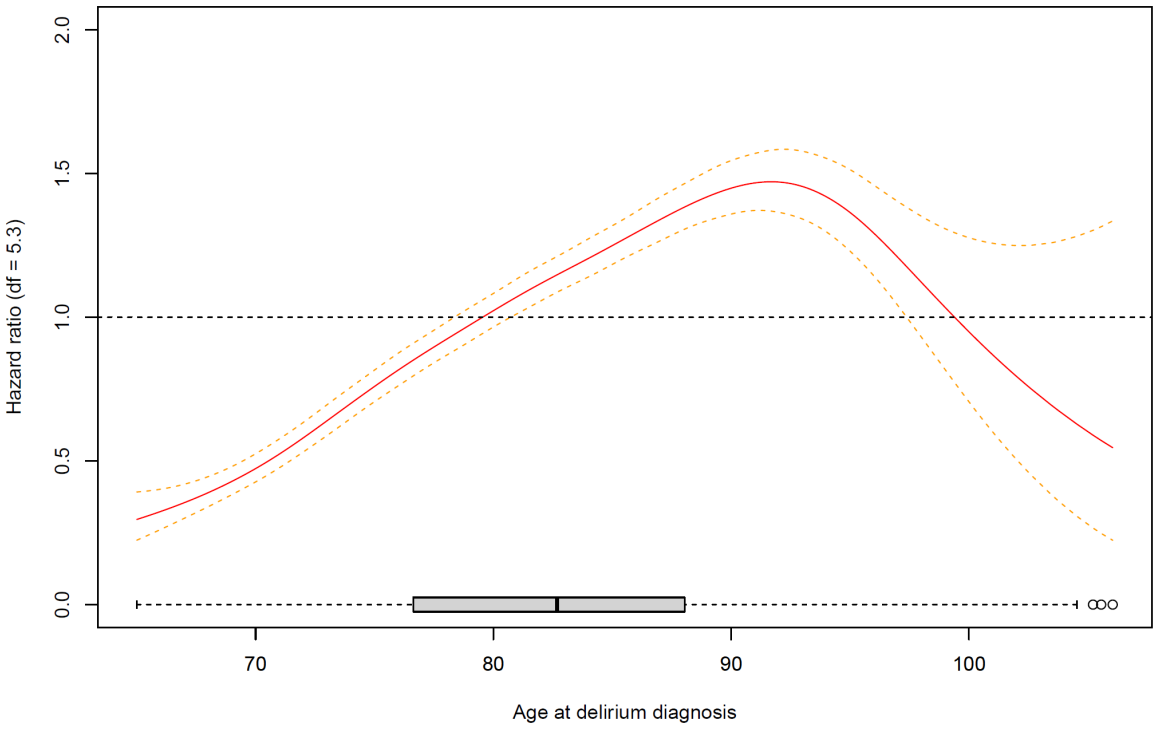


Figure 5-5 Association of age at delirium diagnosis with cause-specific hazard of dementia in Cox model with penalised spline after multivariable adjustment with 95% confidence intervals (reference 79.5 years; p≤0.001). The cause-specific hazard of dementia increases with age of delirium diagnosis from age 65 until around age 90, when it plateaus then decreases. df – degrees of freedom.

5.5 Discussion

To the best of our knowledge, this study represents the largest cohort (n=12949) followed up for the longest period of time (up to 8855 days; mean 741 days) within the published literature examining the new diagnosis of dementia following an episode of delirium. The results show that a first episode of delirium after the age of 65 is associated with a substantial risk of subsequently developing dementia (31% by 5 years). This is in line with data from smaller previously published studies (Davis et al., 2012; Fong et al., 2015; Witlox et al., 2010). Our data also show that delirium is associated with substantial mortality, in addition to the risk of dementia. This underlines the seriousness of delirium and the importance of prompt diagnosis and treatment of underlying cause. Our research supports the concept of delirium as both an indicator of physiological frailty as well as a possible precipitating and accelerating factor in cognitive and physical decline. Within NHS GG&C, there has been a trend of increases in diagnosis of delirium over time. This may indicate that recent high-profile delirium recognition campaigns are having the desired impact including the Think Delirium campaign, which was introduced in NHS GG&C in 2015 (Healthcare Improvement Scotland, 2014; Khachaturian et al., 2020). Findings from the Cox-regression analysis show that the multivariable-adjusted cause-specific hazard of dementia among those diagnosed with delirium increases with higher levels of deprivation and also with advancing age, plateauing and decreasing in extremes of age. However, there does not appear to be a relationship with sex.

The most frequent causes of delirium involve significant systemic inflammation. Inflammation is well recognised as a major precipitant of delirium (Cunningham, 2011). There exists an extensive network of mechanisms that allow neuroimmune communication (Chavan et al., 2017). In recent years, the effects of inflammatory insult on central nervous structure and function have become increasingly well characterised (Cunningham, 2011). Dementia is a disorder which, except in rare single-gene inherited syndromes, has a complex aetiology involving multiple contributory interacting factors. These include ageing, obesity, diabetes, hypertension and smoking—the common strand to these risk factors is the systemic preponderance of inflammatory molecules (Yaffe et al., 2004). Inflammation is thought to have a central mechanistic role in the

pathogenesis of both Alzheimer's dementia (Kinney et al., 2018) and vascular dementia (Iadecola, 2013), the two most common subtypes. While acute inflammation is protective to the brain under most circumstances, prolonged or excess release of proinflammatory molecules within the vulnerable or aged brain may activate various downstream cellular cascades relevant to the emergence of dementia (Hoeijmakers et al., 2016).

These phenomena may be relevant in the context of our findings that support the link between hospitalisation with delirium and subsequent dementia diagnosis. It remains a matter of discussion whether delirium is purely a marker of susceptibility to developing dementia, or unmasks/accelerates unrecognised dementia, or indeed, whether delirium may have direct neurotoxic effects that can be causal in the pathogenesis of dementia (Fong et al., 2015). Evidence from the Vantaa 85+ population-based study may provide evidence to support the latter hypothesis. Neuropathological correlates of dementia such as neurofibrillary tau, ß-amyloid plaque burden, vascular lesions, Lewy-body pathology and ApoE4 allele status were not found to be positively associated with subjects who developed dementia following delirium, while in contrast, a strong association existed in those that developed dementia without a delirium history (Davis et al., 2012). Although the Vantaa study was not powered to be conclusive, it may suggest that, in some cases, dementia following delirium represents a different aetiological pathway to the development of dementia, rather than being purely a vulnerability marker/accelerant of pre-existing disease.

In our multivariable analysis, the cause-specific hazard of dementia increases with age of delirium diagnosis from age 65 until around age 90, when it plateaus then decreases. This is consistent with existing evidence in the general population demonstrating a doubling of both the prevalence and instance of dementia every five to six years until the age of 90 (Qiu & Fratiglioni, 2018). Evidence for trends in dementia diagnosis among the oldest old is limited by sample size. However, two large population-based cohort studies found the increases in the incidence of dementia plateau or even decline beyond age 90. It is suggested that among the oldest old, risk factors for dementia may not be related to the ageing process itself but with age-associated risk factors such as

hypertension, hyperlipidaemia and heart disease (Hall et al., 2005; Jia et al., 2020; Miech et al., 2002).

We found that living in an area of deprivation is associated with an increased cause-specific hazard of developing dementia following an incident episode of delirium after adjusting for age at delirium diagnosis and sex. This supports an earlier finding that the hazard of dementia is increased among those living in areas with higher levels of deprivation in an English population cohort study of 6220 adults over the age of 65 (Cadar et al., 2018). Unfortunately, we did not have information available to adjust for personal indicators of socioeconomic status like personal wealth, educational attainment or occupation, so we are not able to clearly determine whether individual factors were driving this area deprivation effect. However, previous research has shown that living in an area of higher deprivation is associated with poorer health outcomes even after adjusting for personal wealth, education and employment (Stafford & Marmot, 2003).

Our study has a number of strengths including the large sample size and long length of follow-up. Furthermore, by virtue of being registry based, our study is pragmatic and the setting is entirely naturalistic. We have properly accounted for the impact of competing risks by using the CIF rather than Kaplan-Meier estimator and we have modelled the effect of covariates on the cause-specific hazard of dementia in those who experience an episode of delirium. We adhere to gold-standard STROBE reporting guidelines.

There were several limitations. The cohort largely consisted of patients diagnosed within secondary care. Only diagnoses made at death were included from primary care. This introduced a selection bias for more severe cases of delirium requiring assessment at A&E, on admission to hospital or on death. Equally, it is possible that our cases could have had earlier incidences of delirium, perhaps within primary care, which were milder and not coded and indexed to our data set. Similarly, the majority of dementia diagnoses were made within secondary care. If patients moved out of NHS GG&C after their index delirium diagnosis but before their outcome occurred, their outcome would not be known except if it was made at death. As such, it is possible that the proportion of patients who developed dementia was underestimated due to

attrition bias (patients were censored when they should not have been). Furthermore, in those whose dementia diagnosis was made on death, it is possible that this dementia diagnosis was made in primary care at an earlier time point patient and, thus, dementia survival was overestimated. In addition, our cohort is drawn from all medical records over a specific timeframe rather than being set up as a prospective cohort study. We rely on clinicians accurately and reliably coding the diagnosis of delirium at the point of clinical care being administered rather than trained research assistants. While we believe the system of diagnostic coding to be robust within NHS GG&C, it is likely that some cases of dementia or delirium may be missed or inaccurately diagnosed or coded. For example, there is a clear trend of increasing diagnosis of delirium over time within NHS GG&C. This is unlikely to represent a true increase in the underlying rates of delirium but rather represent an increase in the recognition and coding of delirium, perhaps driven by a number of high-profile delirium recognition campaigns, leading to a general increase in awareness of the condition (Khachaturian et al., 2020). Moreover, our multivariable model lacks several important covariates like medical comorbidities, lifestyle factors like diet and smoking or genetics which have been clearly identified as important risk factors for dementia (National Institute for Health and Care Excellence (NICE), 2021a). Finally, when we designed our study, we set it up as a cohort study of patients with an incident episode of delirium to determine the risk of dementia, not as a case–control study, with patient with delirium and matched controls without delirium. As such, we were unable to determine the net effect of delirium itself on dementia diagnosis. Future work should consider a case–control design to answer this important question.

In conclusion, our study reinforces the link between delirium and future dementia within a unique and well-powered data set. It has key clinical implications. We have shown that delirium in over 65s carries a 31% risk of developing dementia and an even greater risk of death in the five years postdiagnosis. This highlights the importance of recognising delirium and preventing it where possible. Future research is required to determine whether the recognition and early treatment of delirium could reduce the risk of subsequent dementia or death. Moreover, at present, there is no consensus about follow-up and monitoring of cognitive function after an episode of

delirium in the elderly. Our findings seem to support closer follow-up of delirium and proactive screening for dementia, but this has implications for service provision, particularly as the population ages. Indeed, it may be that those who experience an episode of delirium represent an ‘at risk’ group who could be candidates for future novel targeted therapies for dementia prevention and early-stage treatment. Finally, important questions about the pathophysiology of delirium remain to be answered. It is unclear whether delirium is a marker or an accelerant of irreversible cognitive decline. The field lacks strong data on the mechanistic relationship between delirium and dementia and indeed the cellular/molecular landscape in delirium and dementia. This is best generated through a combination of neuroimaging approaches, quality animal research and human biomarker studies (Inouye et al., 2014).

Chapter 6 Discussion

6.1 Thesis overview

6.1.1 Thesis primary aim

The primary aim of this thesis was to conduct prognostic model research in first episode psychosis. From this broad aim, I sought to answer four questions:

1. Is prediction of individual patient outcome possible in first episode psychosis using clinical variables?

2. Does prediction model performance remain robust at external validation?

3. Does prediction model performance improve with the addition of biologically relevant disease markers as predictors?

4. Does prediction model performance improve with the application of advanced machine learning classifiers compared to logistic regression?

Chapter 2 explored these four questions in the context of a systematic review of prognostic prediction models developed for predicting poor outcome in first episode psychosis. This systematic review provided evidence that prediction of individual patient outcome is possible in first episode psychosis. Thirteen eligible studies were identified reporting 31 prognostic prediction models. However, just four studies reported external validation with discrimination performance ranging from a c-statistic of 0.556 to 0.876. Calibration and clinical utility were reported in two of those studies, both of which were acceptable. The majority of the 13 studies developed models using logistic regression with only two studies considering machine learning techniques, both employing support vector machines. The majority of the 13 studies used solely clinical variables with only one study utilising biological variables including genetic and environmental factors. No comparison was possible between the performance of logistic regression and machine learning studies because neither machine learning study reported any measures of discrimination, calibration or clinical utility. Similarly, no comparison was made between the study employing biologically relevant disease markers and the others employing solely clinical variables because the

biomarker study did not report discrimination, calibration or measures of clinical utility. Indeed, across the majority of studies considered in the systematic review, methodological limitations were a common theme with only two studies found to have a low risk of bias according to the PROBAST tool. PROBAST defines risk of bias to occur when "shortcomings in study design, conduct or analysis lead to systematically distorted estimates of a model's predictive performance" (Wolff et al., 2019). The common methodological limitations identified included a lack of appropriate validation (e.g. only apparent validation considered), issues with handling of missing data (e.g. by complete case analysis only) and the lack of reporting of calibration, discrimination and measures of clinical utility. Altogether, this review showed that the potential for prediction of individual patient outcomes in first episode psychosis has not yet been fully realised.

Chapter 3 sought to address questions one and two while employing methodological best practice and avoiding the pitfalls identified in Chapter 2 which had led to high risk of bias in the majority of studies. Specifically, a prognostic prediction model was developed and validated using logistic regression and clinical variables which was able to predict robustly individual patient outcome (symptomatic nonremission) in first episode psychosis. The model was developed in 673 patients with first episode psychosis recruited between 2005 and 2010 from 14 early intervention services in NHS England and externally validated in 191 patients recruited between 2006 and 2009 from a further 11 early intervention services in the NHS England. The model showed fair discrimination with a c-statistic of 0.73 (0.64–0.81) and calibration with an intercept of -0.014 (-0.34, 0.31) and slope 0.85 (0.42, 1.27). The model also demonstrated clinical utility across the range of probability thresholds chosen by clinicians. Once this first episode psychosis nonremission prediction model has been prospectively validated, it could facilitate the early identification of patients at high risk of nonremission and prioritise the timely delivery of effective phase-specific treatments like clozapine for treatment resistance.

Chapter 4 addressed questions three and four by internal validation in a cohort of 168 first episode psychosis patients from the Lilly F1D-MC-HGDH trial - a double-blind, multicentre, randomised controlled trial of Olanzapine versus Haloperidol treatment. To date the majority of first episode psychosis prediction

models have been developed using logistic regression and include only clinical variables. The potential for biological disease markers to improve model performance in FEP has not been adequately explored. Further, there is a growing interest in applying machine learning methods which promise to capture better nonlinearity and model complex interactions in medical data. The discrimination (c-statistic) and calibration (calibration-in-the-large and calibration slope) performance of a logistic regression first episode non-remission risk prediction model built using solely clinical variables was compared to models built with the addition of peripheral inflammatory or magnetic resonance spectroscopy glutamate biomarkers. The performance of prediction models derived by maximum likelihood estimation and elastic net logistic regression was then compared to models built using machine learning including naïve Bayes, random forest, linear and radial support vector machines. When comparing a logistic regression model with clinical variables to models with the addition of a biological variable, there was no significant effect of model type on the c-statistic ($F(5, 14.6) = 0.184$; $p = 0.9640$), calibration-in-the-large ($F(5, 176512.09) = 0.021$; $p = 0.9998$), or calibration slope ($F(5, 114.8) = 0.206$; $p = 0.9594$). Comparing logistic regression models to machine learning models showed a significant effect of model type on the c-statistic ($F(5, 398.66) = 7.225$; $p = <0.0001$) and calibration-in-the-large ($F(5, 5.87) = 56.356$; $p = <0.0001$) but not calibration slope ($F(5, 218.1) = 1.139$; $p = 0.3405$). In post-hoc tests, Naïve Bayes showed superior discrimination performance compared to maximum likelihood estimation but not elastic net logistic regression. All models except maximum likelihood estimation logistic regression demonstrated problems with calibration. Taken together these results show that, in this dataset, for the prediction of first episode psychosis the addition of a biological variable does not improve the performance of a logistic regression model built using clinical variables. Further, machine learning or elastic net logistic regression did not result in improved global performance compared to maximum likelihood estimation logistic regression.

6.1.2 Thesis secondary aim

The secondary aim for this thesis was to answer a final question:

5. Can routinely collected electronic healthcare record data could be used for prognostic research in the National Health Service in Greater Glasgow and Clyde?

The coronavirus pandemic delayed collection of routine data in first episode psychosis in NHS GG&C such that it was not possible to use these data for my thesis. In Chapter 5 I took the opportunity to examine this question in a more common psychiatric presentation, delirium, in the hope that information from this would inform future prospective studies in first episode psychosis. Delirium is an important risk factor for subsequent dementia. However, the field lacks large studies with long-term follow-up of delirium in subjects initially free of dementia to clearly establish clinical trajectories. I undertook a retrospective cohort study of all patients over the age of 65 diagnosed with an episode of delirium who were initially dementia free at onset of delirium within NHS GG&C between 1996 and 2020 using the Safe Haven database (NHS Greater Glasgow & Clyde, 2023). The cumulative incidence of dementia was estimated accounting for the competing risk of death without a dementia diagnosis. The effects of age at delirium diagnosis, sex and socioeconomic deprivation on the cause-specific hazard of dementia were modelled via Cox regression. 12949 patients with an incident episode of delirium were included and followed up for an average of 741 days. The estimated cumulative incidence of dementia was 31% by 5 years. The estimated cumulative incidence of the competing risk of death without dementia was 49.2% by 5 years. The cause-specific hazard of dementia was increased with higher levels of deprivation and also with advancing age from 65, plateauing and decreasing from age 90. There did not appear to be a relationship with sex. This study reinforced the link between delirium and future dementia in a large cohort of patients. It highlights the importance of early recognition of delirium and prevention where possible. Finally, this study demonstrated the feasibility of using routinely collected electronic health record data from NHS GG&C via the Safe Haven database for prognostic research. This will inform future prospective prognostic modelling studies into first episode psychosis using Safe Haven.

6.2 Strengths and limitations

6.2.1 Methodological considerations

The strengths and limitations have been reviewed in each chapter but the salient points bear repeating. A key strength of this thesis is the inclusion of the first systematic review of prediction modelling studies in first episode psychosis. This review presented a thorough critique of the prediction model literature with particular attention given to their study characteristics, methodologies and model performance metrics. The systematic review found evidence for the prediction of individual patient outcome in first episode psychosis, with four of the thirteen studies including external validation. However, the review was limited by the poor reporting of discrimination and calibration measures across the included studies such that it was often difficult to make comparisons between prediction models. Many studies presented solely classification metrics such as accuracy. The problem with solely reporting classification metrics is that they vary both across models and across different probability thresholds for the same model. This had a detrimental effect on the review's ability to properly address my third and fourth research questions. Specifically, no comparison between the single study employing biologically relevant disease markers as predictors and those using solely clinical solely clinical variables was possible nor was a comparison between the studies employing machine learning methods and those using regression techniques. Another limitation was the fact that many of the studies and models considered different outcomes in first episode psychosis (six different outcomes considered across the 13 included studies). Direct comparisons between the performance of models built to predict different outcomes is not possible. Additionally, every study developed new prediction models rather than attempting to validate or update an existing model. If researchers only create new prediction models, this discards historical data and previous research efforts (Jenkins et al., 2021). Perhaps due to the focus on only building new models, no study took an existing prediction model close to clinical practice – there were no assessments of a model's clinical impact and no model in the systematic review was deployed clinically.

Building on this systematic review, an important strength of chapters 3 and 4 was the use of gold standard methodological practices improving on

methodological limitations identified in chapter 2. Specific methodological improvements in chapters 3 and 4 include the handling of missing data. Unlike the majority of studies identified in chapter 2, missing data in chapters 3 and 4 was handled by multiple imputation using all available data including auxiliary variables. This increases the statistical power and reduces bias. A further methodological improvement common to both chapters 3 and 4 was the use of expert knowledge for variable selection with consideration given to using an appropriate number of events per predictor variable for the sample size. A methodological strength of chapter 3 was the external validation of the prediction model in a similar population derived from an independent study. As outlined above this answered a key question for my thesis – that prediction model performance would remain robust at external validation. However, this thesis would have been further strengthened if questions three and four, the assessment of biological variables and the consideration of machine learning methods, had also been tested at external validation in chapter 4 rather than only at internal validation. Another weakness of chapter 4 was the relatively modest sample size, but this was mitigated by performing robust internal validation by ten-fold cross-validation repeated 50 times.

6.2.2 Cohort selection

An additional strength of chapter 3 was the use of representative samples of first episode psychosis patients drawn from early intervention services in England for both the development and validation cohorts. This improved generalisability of the findings to first episode psychosis patients in early intervention services. However, there were limitations. While the development and validation cohorts were drawn from naturalistic studies, there was still potential for selection bias to be introduced as participants who did not present to services, who declined consent or who dropped out were excluded from the analysis. Reassuringly, for the development cohort in chapter 3, there was little evidence of differences between those who consented and those who declined consent, but no such data was available for the validation cohort (Birchwood et al., 2014). In contrast, the population for chapter 4 was derived from a randomised controlled trial. This randomised controlled trial population differed to participants in chapter 3. Namely, the participants were not drawn from early intervention services. Further, the majority were recruited from the

United States which has a very different healthcare model to the NHS as well as a different sociodemographic and ethnic composition to the UK. Moreover, stringent trial conditions were not representative of routine clinical practice. This impacted the generalisability of the findings.

There were further limitations resulting from the patient populations used for chapters 3 and 4. The participants in chapters 3 were recruited and followed more than a decade ago and, in the case of chapter 4, more than two decades ago. First episode psychosis healthcare and the wider geopolitical climate have changed considerably in recent years such that the relevance of prediction models developed to current clinical practice using this historical data may be limited. Specifically, calibration drift has been identified as a key concern when attempting to deploy prediction models in ever changing clinical environments where differences arise over time between the population on which a model was developed and the population on which it is intended to be applied (Davis et al., 2020). As these prediction models were developed on historical data, calibration issues may be present from the outset if the models are to be applied to new patients. This may necessitate updating the models using new patient data before considering trialling their application to clinical practice. Indeed, one of the key aims of my planned PhD fellowship was to operationalise the collection of routine clinical information for patients with first episode psychosis in NHS GG&C. The goal was to prospectively validate my existing prediction models on new patients as they entered the service using real world clinical data derived from an NHS early intervention service. Unfortunately, implementation of data collection was severely delayed as a consequence of the global coronavirus pandemic such that this was not possible during my PhD timeline. As mitigation, I sought to answer my first four thesis aims using existing first episode psychosis datasets, rather than using new data.

Additional limitations resulting from the dataset used for chapter 4 of my thesis included the limited number of biologically relevant disease markers available to test as predictors. It is possible that the usefulness of these biological variables for prediction was curtailed by the lack of a mechanistic link to the pathophysiology of psychosis thereby limiting their effect size. Many biomarkers with a greater evidence base for disturbance in psychosis were not available, including proinflammatory cytokines and chemokines, which numerous

systematic reviews and meta-analyses have identified as potential biomarkers in psychosis (Cakici et al., 2020; Goldsmith et al., 2016; Miller et al., 2011; Upthegrove et al., 2014). Further, much of the evidence for glutamate dysfunction in psychosis comes from more modern magnetic resonance spectroscopy studies conducted at higher field strengths than 1.5 Telsa (Merritt et al., 2016; Merritt et al., 2021).

6.2.3 Routinely collected data

Chapter 5 had a number of strengths including the large sample size and long length of follow-up. Further, by virtue of being registry based using routinely collected clinical data from NHS GG&C, the study was entirely naturalistic and pragmatic. Insights drawn from naturalistic studies are often more generalisable to real-world practice as they do not have the same strict inclusion/exclusion criteria and highly controlled study settings as randomised controlled trials (Cook & Thigpen, 2019). However, there are disadvantages of using registry data. Whereas the prognostic factors and outcomes in chapters 3 and 4 were collected using standardised instruments by trained researchers, for chapter 5, the integrity of the data used relies on clinicians accurately and reliably coding the delirium and dementia diagnoses. Further, unlike a planned prospective study, the covariates included in the model were only those available in the dataset. As such, chapter 5 lacked several important covariates such as medical comorbidities, lifestyle factors (e.g. diet and smoking) and genetics which have been identified as important risk factors for dementia (National Institute for Health and Care Excellence (NICE), 2021a). Further, patients included in the study were primarily drawn from those treated within secondary care. This would likely have resulted in a selection bias for more severe delirium and dementia diagnoses. Finally, continuity of my thesis would have been improved if the routinely collected data for first episode psychosis had been available rather than changing the focus to delirium and dementia for chapter 5.

6.3 Future directions

A key aim of my original PhD fellowship was to be the establishment of the Electronic Measures in Psychosis – Assessing Trajectory and Health-Outcomes (EMPATH) platform for the collection of outcome data for first episode psychosis

patients treated within the Esteem NHS GG&C early intervention in psychosis service. There were significant delays in developing and deploying the EMPATH platform as a consequence of the global coronavirus pandemic such that deployment was postponed until 2023. Routinely collected data should be available for prognosis research from 2024. Building on the lessons learned from Chapter 5, I plan to use routinely collected clinical data gathered by EMPATH for future prognostic research into first episode psychosis. For example, data-linkage with EMPATH could allow the testing of additional routinely collected inflammatory biological disease markers (from patients' baseline clinical bloods) as predictor variables in larger sample sizes.

Secondly, the prognostic prediction models developed in this thesis are developed to use static baseline information to predict a fixed future end-point. As such, they are only relevant to patients at entry to early intervention services. The same model should not be used to make subsequent predictions in the same patients followed up over time. Indeed, the vast majority of clinical prediction models developed and all those recommended by NICE and SIGN for clinical practice use predictors at baseline to predict a fixed future end point. However, this does not reflect reality – real-world clinical problems are not stationary. Rather, predictions should be able to be updated at later timepoints in a patient's journey in response to the natural evolution of their illness and any changes in management. Future work should focus on dynamic prediction models as a solution. Dynamic prediction models are updated over time as more information is collected. Past predictions can be combined with new information at future time points and longitudinal risk factors can be incorporated as time-dependent predictors. Dynamic prediction has demonstrated greater accuracy compared to static models (Bone et al., 2021). Dynamic modelling solutions include iterative logistic regression, joint modelling and Bayesian methods for continuous updating (Bone et al., 2021; Jenkins et al., 2018; Yuen et al., 2018; Yuen et al., 2020). Yet, despite clear benefits, currently dynamic modelling is rarely applied to clinical problems (Jenkins et al., 2018).

Finally, the ultimate goal of prognostic model research is to enable the targeting of treatments according to individual patient risk. Expensive or higher risk treatment may be reserved for those at higher risk. This the basis of stratified medicine (Hingorani et al., 2013; Steyerberg et al., 2013). The aim for the

model is not to replace clinicians but to help inform a shared approach to decision making (Steyerberg et al., 2013). Successful examples of prognostic models integrated into clinical practice include the PREDICT tool which is recommended to guide adjuvant therapy in individuals with invasive breast cancer (National Institute for Health and Care Excellence (NICE), 2018; Wishart et al., 2010), and, the QRISK tool which is recommended to guide lipid lowering treatment based on an individual's cardiovascular risk (Hippisley-Cox et al., 2017; National Institute for Health and Care Excellence (NICE), 2016a). Recent systematic review has shown that no model within the field of psychiatry has reached this point (Meehan et al., 2022).

Chapter 3 of this thesis described the development and external validation of a first episode psychosis nonremission risk prediction model. Development and external validation are the first and second key steps required to take a prognostic prediction model to clinical practice, as outlined in the PROGRESS framework (Steyerberg et al., 2013). However, good model performance, even on external validation, will not translate to clinical utility if clinicians are unable to safely apply a model into clinical practice. Moreover, premature implementation of a model into clinical practice may even be harmful if it leads to people who might otherwise benefit from treatment being denied access (Hingorani et al., 2013). The third and final step required to take a model to clinical practice is assessment of a prediction model's clinical impact.

The gold standard for a clinical impact study is a comparative design ideally via a pragmatic cluster-based randomised controlled trial, whereby, one healthcare group is randomised to carry out usual care while the other group is given access to individualised predictions from a prognostic model to guide treatment (Moons et al., 2009). This helps to establish, by the accurate identification of individuals who will have poor outcomes, whether we can meaningful intervene to improve their prognosis.

However, prior to a full-scale clinical impact study, I plan to conduct a feasibility study. A feasibility study is a recommended first step to establish how best to integrate a prognostic model into clinical practice and to identify what its measurable impacts may be, including on health outcomes and the cost-effectiveness of care (de Hond et al., 2022). The study will help determine the

practicality and acceptability of delivering Chapter 3's first episode psychosis nonremission risk prediction model in a clinical setting, and identify any barriers to implementation or any unintended negative effects. The model was developed using retrospective data. The planned feasibility study will also allow prospective validation of the model in a new cohort of first episode psychosis patients in a real-world clinical setting.

To facilitate this feasibility study, I have had Chapter 3's first episode psychosis nonremission risk prediction model developed into a password protected user-friendly web-app (Figure 6-1). The web-app securely and pseudo-anonymously saves the entered patient predictor data and their calculated outcome probability, linking them to the clinician user of the web-app.

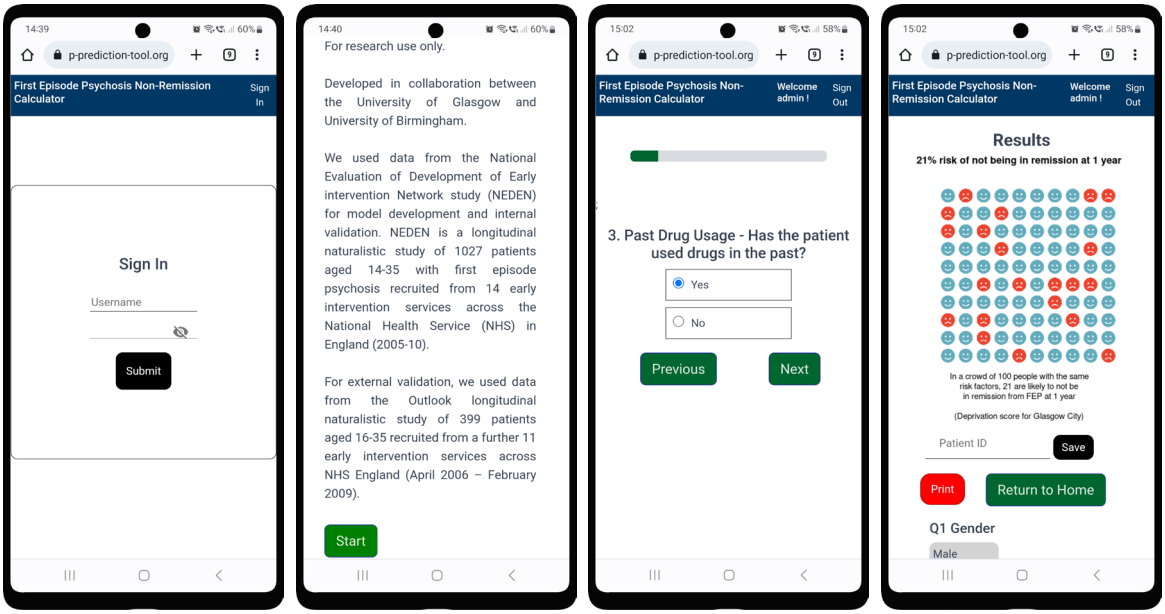


Figure 6-1 First episode psychosis nonremission risk prediction model web-app screenshots on mobile for planned feasibility study.

In order for a psychiatry model to reach the point of deployment to clinical practice, like the QRISK tool in cardiovascular disease, its clinical impact must be established. As part of this process, a clinical intervention based on the model needs to be recommended and a probability threshold above which it should be applied. The fundamental purpose of clinical prediction models is to improve patient outcomes. The implicit assumption of the model is that providing enough forewarning should allow the patient to modify and improve their outcome. Given this based on a model's predictions, interventions should be recommended (Lenert et al., 2019). In addition, clinicians are required to

decide a probability threshold above which they would offer this intervene. The probability threshold is chosen based on the benefits and harms of the proposed intervention in order to optimise the balance between false positives and false negatives. For example, in the case of a model predicting cancer, the clinician would choose a lower probability threshold (more false positives, less false negatives) to offer a non-invasive screening test and a higher probability threshold (less false positives, more false negatives) to suggest an invasive and potentially harmful biopsy (Vickers & Elkin, 2006). In weighing the benefits and harms of the proposed intervention it is essential to consider the views of the patient. However, most models never get to the point of clinical use and so neither the proposed intervention nor the probability threshold above which the clinician would apply it are specified. QRISK is a rare example of a model with a specified intervention (statin treatment) and a proposed probability threshold for this intervention based on its harms and benefits (greater than or equal to 10% risk of developing a heart attack or stroke over the next 10 years) (National Institute for Health and Care Excellence (NICE), 2016a).

In chapter 3, an intervention was proposed for the first episode psychosis nonremission risk prediction model, namely, “enhanced monitoring” over routine care leading to early identification and intervention for treatment resistance, substance misuse, or nonconcordance. Based on discussion with NHS early-intervention specialists (eight NHS Consultant Psychiatrists), a probability threshold of 40%-60% for “enhanced monitoring” was suggested. Future work is required with clinicians, patients and policy-makers to reach a consensus decision about an intervention that is achievable and realistic together with an appropriate probability threshold to apply it. This will form an integral part of a planned clinical impact study. Thereafter, the first episode psychosis nonremission risk prediction model can be applied to clinical practice. This will allow clinicians to provide treatment to patients who need it most, facilitating timely intervention, and enable efficient and effective clinical care.

6.4 Conclusions

This thesis provided an overview of the prediction modelling field in first episode psychosis highlighting the importance of methodological rigour. It outlined the development and external validation of a first episode psychosis nonremission

risk prediction model in two large naturalistic cohorts of patients. The model could allow clinicians to intervene earlier to change trajectories and improve prognosis in first episode psychosis but first requires prospective validation and its clinical impact established in a future trial. There was an exploration of the potential for biological disease markers and machine learning classifiers to augment model performance in first episode psychosis. The potential was not borne out in this analysis but further external validation studies in larger sample sizes with additional biomarkers are necessary. Finally, it has proven feasible to use routinely collected clinical data for prognostic research in NHS GG&C in delirium. With the establishment of the EMPATH platform, I look forward to future prognostic work in first episode psychosis as routinely collected clinical data is deposited. Altogether, this thesis made several contributions to the growing body of clinical prognostic research in first episode psychosis and delirium. In particular, considerable progress has been made towards the deployment of a useable and informative clinical prediction model which will improve care for people with first episode psychosis.

Appendix 1 Search strategy for Chapter 2

PsycINFO Search:

Psychosis Terms:

1.Acute Psychosis/ or Psychosis/
2.first episode psychosis.m_titl.
3.psychosis.m_titl.

4. 1 or 2 or 3

Outcomes Terms:

5.Treatment Outcomes/ or Health Outcomes/ or Psychotherapeutic Outcomes/ or Psychosocial Outcomes/ or Symptom Remission/
6."recovery (disorders)"/ or relapse prevention/
7.treatment resistant disorders/
8."quality of life"/ or "health related quality of life"/ or "quality of work life"/
9.vocational rehabilitation/
10.relapse prevention.m_titl.
11.(outcome* or remission or recovery).m_titl.
12."treatment resis*.".m_titl.
13.quality of life.m_titl.
14.social recovery.m_titl.
15.vocational recovery.m_titl.

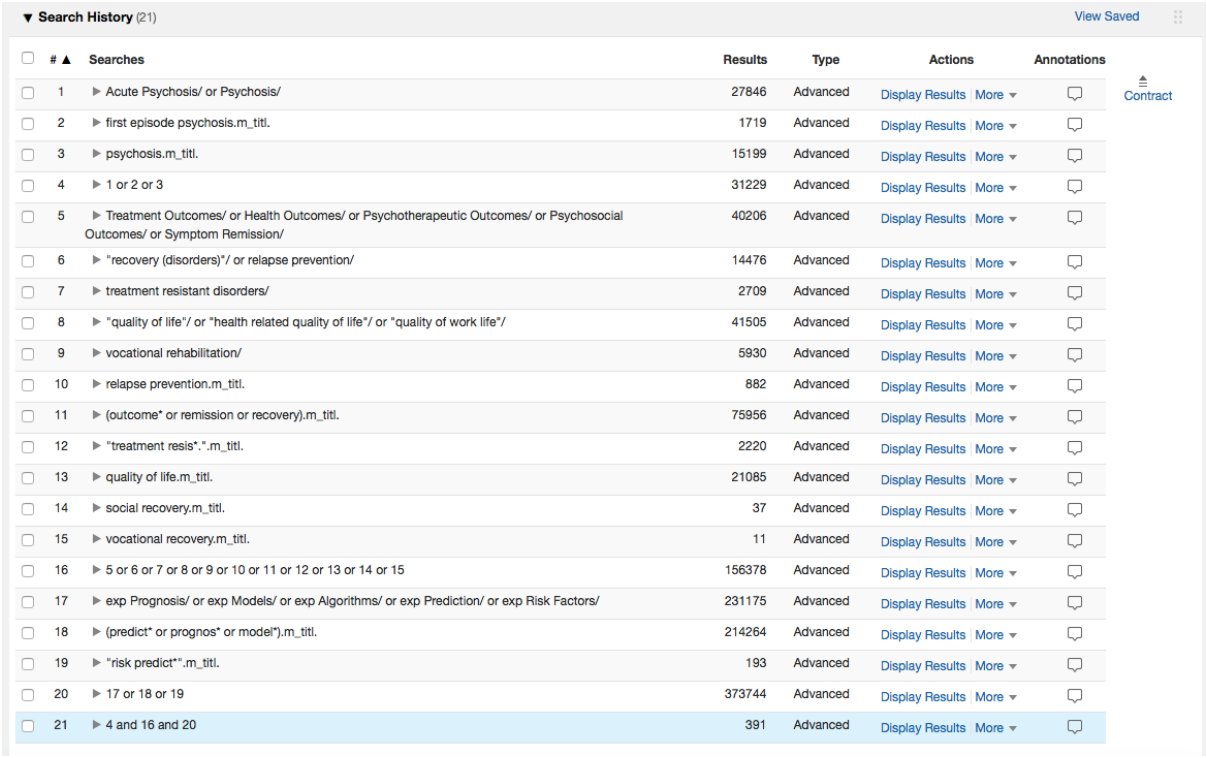
16. 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15

Prediction Terms:

17.exp Prognosis/ or exp Models/ or exp Algorithms/ or exp Prediction/ or exp Risk Factors/
18.(predict* or prognos* or model*).m_titl.
19."risk predict*".m_titl.

20. 17 or 18 or 19

21. 4 and 16 and 20



▼ Search History (21) View Saved

# ▲	Searches	Results	Type	Actions	Annotations
1	▶ Acute Psychosis/ or Psychosis/	27846	Advanced	Display Results | More	
2	▶ first episode psychosis.m_titl.	1719	Advanced	Display Results | More	
3	▶ psychosis.m_titl.	15199	Advanced	Display Results | More	
4	▶ 1 or 2 or 3	31229	Advanced	Display Results | More	
5	▶ Treatment Outcomes/ or Health Outcomes/ or Psychotherapeutic Outcomes/ or Psychosocial Outcomes/ or Symptom Remission/	40206	Advanced	Display Results | More	
6	▶ "recovery (disorders)"/ or relapse prevention/	14476	Advanced	Display Results | More	
7	▶ treatment resistant disorders/	2709	Advanced	Display Results | More	
8	▶ "quality of life"/ or "health related quality of life"/ or "quality of work life"/	41505	Advanced	Display Results | More	
9	▶ vocational rehabilitation/	5930	Advanced	Display Results | More	
10	▶ relapse prevention.m_titl.	882	Advanced	Display Results | More	
11	▶ (outcome* or remission or recovery).m_titl.	75956	Advanced	Display Results | More	
12	▶ "treatment resis*.".m_titl.	2220	Advanced	Display Results | More	
13	▶ quality of life.m_titl.	21085	Advanced	Display Results | More	
14	▶ social recovery.m_titl.	37	Advanced	Display Results | More	
15	▶ vocational recovery.m_titl.	11	Advanced	Display Results | More	
16	▶ 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15	156378	Advanced	Display Results | More	
17	▶ exp Prognosis/ or exp Models/ or exp Algorithms/ or exp Prediction/ or exp Risk Factors/	231175	Advanced	Display Results | More	
18	▶ (predict* or prognos* or model*).m_titl.	214264	Advanced	Display Results | More	
19	▶ "risk predict*".m_titl.	193	Advanced	Display Results | More	
20	▶ 17 or 18 or 19	373744	Advanced	Display Results | More	
21	▶ 4 and 16 and 20	391	Advanced	Display Results | More	

Contract

<u>EMBASE Search</u>:

Psychosis Terms:

1.*acute psychosis/ or *psychosis/
2.psychosis.m_titl.

3. 1 or 2

Outcomes Terms:

4.*treatment outcome/
5.*outcomes research/
6.*remission/
7.*“quality of life”/
8.*relapse/
9.*vocational rehabilitation/
10.relapse prevention.m_titl.
11.(outcome* or remission or recovery).m_titl.
12.“treatment resis*”.m_titl.
13.quality of life.m_titl.
14.social recovery.m_titl.
15.vocational recovery.m_titl.

16. 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15

Prediction Terms:

17.*prognosis/

18.*computer model/ or *psychological model/ or *anatomic model/ or *individual based population model/ or *mathematical model/ or *statistical model/
19.*algorithm/
20.*algorithm/ or *classification algorithm/ or *coding algorithm/
21.*prediction/
22.*computer prediction/ or *"prediction and forecasting"/
23.*risk factor/
24.(predict* or prognos* or model*)m_titl.
25."risk predict*".m_titl.

26. 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25

27. 3 and 16 and 26

# ▲	Searches	Results	Type	Actions	Annotations
1	*acute psychosis/ or *psychosis/	41983	Advanced	Display Results | More	
2	psychosis.m_titl.	24129	Advanced	Display Results | More	
3	1 or 2	46952	Advanced	Display Results | More	
4	*treatment outcome/	22783	Advanced	Display Results | More	
5	*outcomes research/	7973	Advanced	Display Results | More	
6	*remission/	15354	Advanced	Display Results | More	
7	*"quality of life"/	101021	Advanced	Display Results | More	
8	*relapse/	15467	Advanced	Display Results | More	
9	*vocational rehabilitation/	4438	Advanced	Display Results | More	
10	relapse prevention.m_titl.	960	Advanced	Display Results | More	
11	(outcome* or remission or recovery).m_titl.	591661	Advanced	Display Results | More	
12	*treatment resis*".m_titl.	5147	Advanced	Display Results | More	
13	quality of life.m_titl.	99823	Advanced	Display Results | More	
14	social recovery.m_titl.	48	Advanced	Display Results | More	
15	vocational recovery.m_titl.	18	Advanced	Display Results | More	
16	4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15	743595	Advanced	Display Results | More	
17	*prognosis/	44741	Advanced	Display Results | More	
18	*computer model/ or *psychological model/ or *anatomic model/ or *individual based population model/ or *mathematical model/ or *statistical model/	61441	Advanced	Display Results | More	
19	*algorithm/	54580	Advanced	Display Results | More	
20	*algorithm/ or *classification algorithm/ or *coding algorithm/	56488	Advanced	Display Results | More	
21	*prediction/	33843	Advanced	Display Results | More	
22	*computer prediction/ or *"prediction and forecasting"/	938	Advanced	Display Results | More	
23	*risk factor/	74514	Advanced	Display Results | More	
24	(predict* or prognos* or model*).m_titl.	1288283	Advanced	Display Results | More	
25	*risk predict*".m_titl.	4254	Advanced	Display Results | More	
26	17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25	1442386	Advanced	Display Results | More	
27	3 and 16 and 26	383	Advanced	Display Results | More	

Contract

<u>CINAHL Plus Search</u>:

Psychosis Terms:

S1 (MH "Psychotic Disorders")
S2 TI psychosis

S3 S1 or S2

Outcomes Terms:

S4 (MH "Outcomes (Health Care)") OR (MH "Treatment Outcomes") OR (MH "Outcomes Research")
S5 (MH "Recovery")
S6 (MH "Quality of Life") OR (MH "Psychological Well-Being")
S7 (MH "Rehabilitation, Vocational") OR (MH "Rehabilitation, Psychosocial")

S8 TI relapse prevention
S9 TI (outcome* OR remission OR recovery)
S10 TI treatment resis*
S11 TI quality of life
S12 TI social recovery
S13 TI vocational recovery

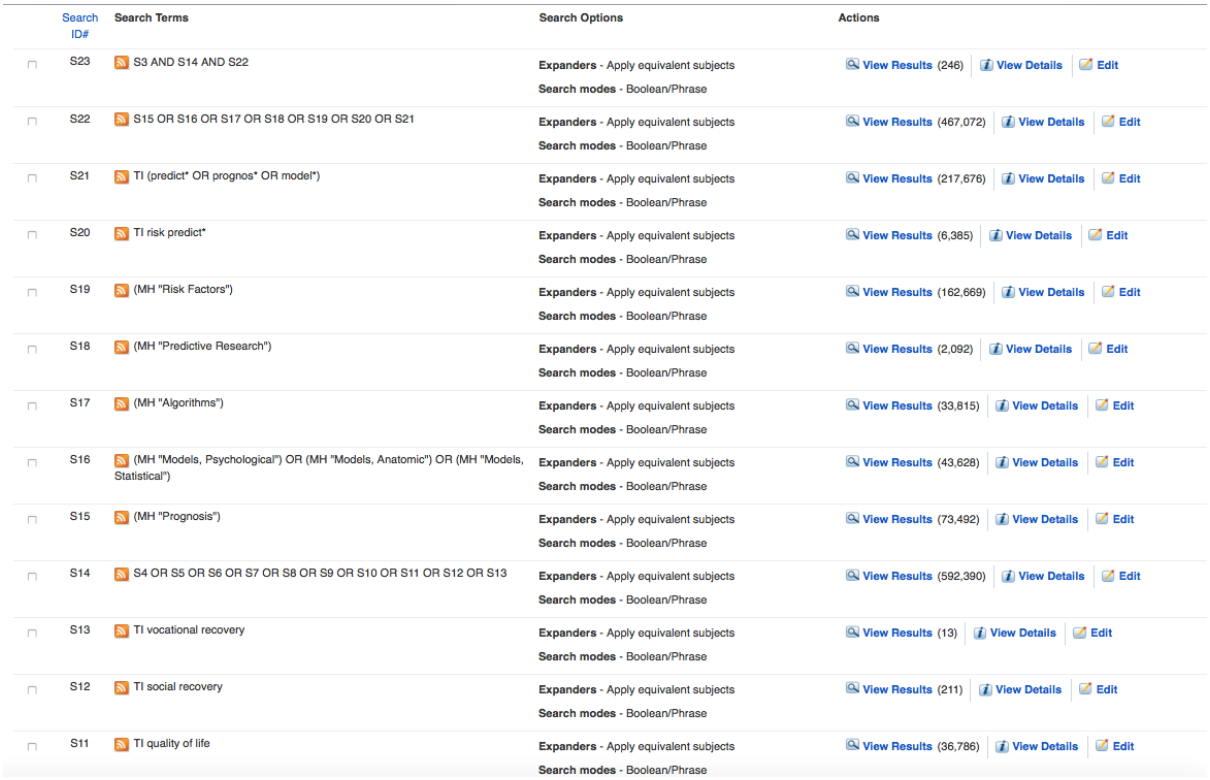
S14 S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S11 OR S12 OR S13

Prediction Terms:

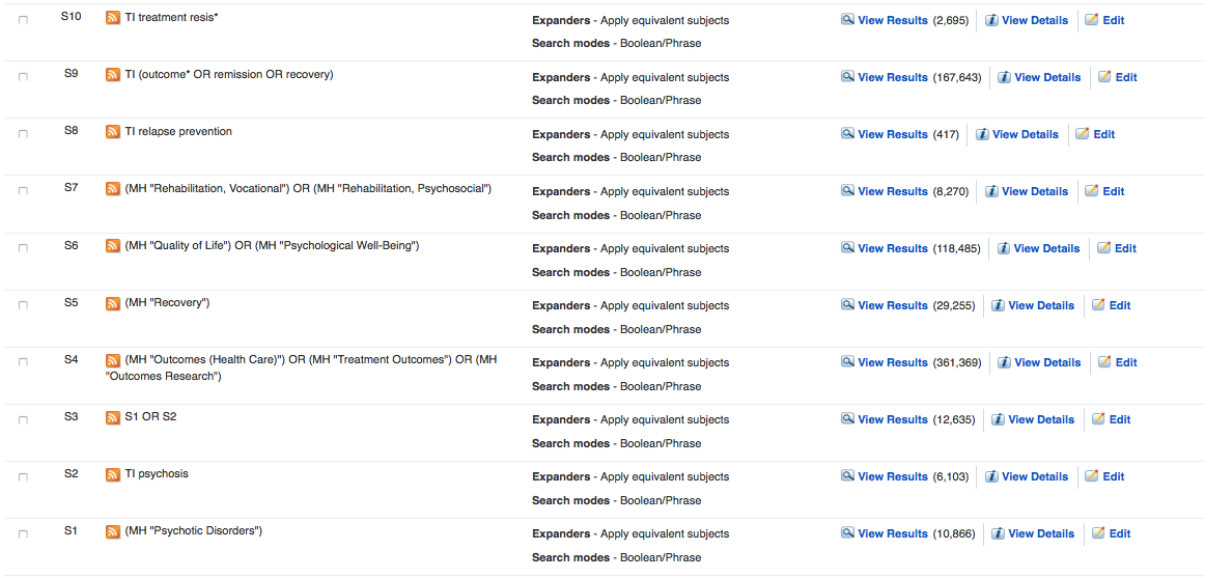
S15 (MH “Prognosis”)
S16 (MH “Models, Psychological”) OR (MH “Models, Anatomic”) OR (MH “Models, Statistical”)
S17 (MH “Algorithms”)
S18 (MH “Predictive Research”)
S19 (MH “Risk Factors”)
S20 TI risk predict*
S21 TI (predict* OR prognos* OR model*)

S22 S15 OR S16 OR S17 OR S18 OR S19 OR S20 OR S21

S23 S3 AND S14 AND S22



Search ID#	Search Terms	Search Options	Actions
S23	S3 AND S14 AND S22	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (246) | View Details | Edit
S22	S15 OR S16 OR S17 OR S18 OR S19 OR S20 OR S21	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (467,072) | View Details | Edit
S21	TI (predict* OR prognos* OR model*)	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (217,676) | View Details | Edit
S20	TI risk predict*	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (6,385) | View Details | Edit
S19	(MH "Risk Factors")	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (162,669) | View Details | Edit
S18	(MH "Predictive Research")	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (2,092) | View Details | Edit
S17	(MH "Algorithms")	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (33,815) | View Details | Edit
S16	(MH "Models, Psychological") OR (MH "Models, Anatomic") OR (MH "Models, Statistical")	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (43,628) | View Details | Edit
S15	(MH "Prognosis")	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (73,492) | View Details | Edit
S14	S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S11 OR S12 OR S13	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (592,390) | View Details | Edit
S13	TI vocational recovery	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (13) | View Details | Edit
S12	TI social recovery	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (211) | View Details | Edit
S11	TI quality of life	**Expanders** - Apply equivalent subjects **Search modes** - Boolean/Phrase	View Results (36,786) | View Details | Edit



S10	TI treatment resis*	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	View Results (2,695)	View Details	Edit
S9	TI (outcome* OR remission OR recovery)	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	View Results (167,643)	View Details	Edit
S8	TI relapse prevention	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	View Results (417)	View Details	Edit
S7	(MH "Rehabilitation, Vocational") OR (MH "Rehabilitation, Psychosocial")	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	View Results (8,270)	View Details	Edit
S6	(MH "Quality of Life") OR (MH "Psychological Well-Being")	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	View Results (118,485)	View Details	Edit
S5	(MH "Recovery")	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	View Results (29,255)	View Details	Edit
S4	(MH "Outcomes (Health Care)") OR (MH "Treatment Outcomes") OR (MH "Outcomes Research")	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	View Results (361,369)	View Details	Edit
S3	S1 OR S2	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	View Results (12,635)	View Details	Edit
S2	TI psychosis	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	View Results (6,103)	View Details	Edit
S1	(MH "Psychotic Disorders")	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase	View Results (10,866)	View Details	Edit

<u>Web of Science - Core Collection Search</u>:

Psychosis Terms:

#1 TS=Psychosis

Outcome Terms:

#2 TI=(outcome* OR recovery OR remission OR “quality of life” OR treatment resis*)

Prediction Terms:

#3 TI=(predict* OR prognos* OR model*)

#4 #3 AND #2 AND #1

# 4	493	#3 AND #2 AND #1 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years*	Edit
# 3	3,154,077	TI=(predict* OR prognos* OR model*) *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years*	Edit
# 2	783,746	TI=(outcome* OR recovery OR remission OR "quality of life" OR treatment resis*) *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years*	Edit
# 1	62,890	TS=Psychosis *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years*	Edit

<u>PubMed Search</u>:

Psychosis Terms

#1 psychosis[Title/Abstract]

Outcome Terms

#2 ((((((((((“outcome assessment, health care”[MeSH Major Topic]) OR “treatment outcome”[MeSH Major Topic]) OR “quality of life”[MeSH Major Topic]) OR “mental health recovery”[MeSH Major Topic]) OR “rehabilitation, vocational”[MeSH Major Topic]) OR relapse prevention[Title]) OR treatment

resis*[Title]) OR outcome*[Title]) OR remission[Title]) OR recovery[Title]) OR “quality of life”[Title]) OR social recovery[Title]) OR vocational recovery[Title]

Prediction Terms

#3 (((((((((“prognosis”[MeSH Major Topic]) OR “forecasting”[MeSH Major Topic]) OR “algorithms”[MeSH Major Topic]) OR “models, psychological”[MeSH Major Topic]) OR “models, statistical”[MeSH Major Topic]) OR “risk factors”[MeSH Major Topic]) OR predict*[Title]) OR prognos*[Title]) OR model*[Title]) OR risk predict*[Title]

#4

((psychosis[Title/Abstract]) AND ((((((((((“outcome assessment, health care”[MeSH Major Topic]) OR “treatment outcome”[MeSH Major Topic]) OR “quality of life”[MeSH Major Topic]) OR “mental health recovery”[MeSH Major Topic]) OR “rehabilitation, vocational”[MeSH Major Topic]) OR relapse prevention[Title]) OR treatment resis*[Title]) OR outcome*[Title]) OR remission[Title]) OR recovery[Title]) OR “quality of life”[Title]) OR social recovery[Title]) OR vocational recovery[Title])) AND ((((((((((“prognosis”[MeSH Major Topic]) OR “forecasting”[MeSH Major Topic]) OR “algorithms”[MeSH Major Topic]) OR “models, psychological”[MeSH Major Topic]) OR “models, statistical”[MeSH Major Topic]) OR “risk factors”[MeSH Major Topic]) OR predict*[Title]) OR prognos*[Title]) OR model*[Title]) OR risk predict*[Title])

History

Download history Clear history

Search	Add to builder	Query	Items found	Time
#4	Add	Search **((psychosis[Title/Abstract]) AND ((((((((((“outcome assessment, health care”[MeSH Major Topic]) OR “treatment outcome”[MeSH Major Topic]) OR “quality of life”[MeSH Major Topic]) OR “mental health recovery”[MeSH Major Topic]) OR “rehabilitation, vocational”[MeSH Major Topic]) OR relapse prevention[Title]) OR treatment resis*[Title]) OR outcome*[Title]) OR remission[Title]) OR recovery[Title]) OR “quality of life”[Title]) OR social recovery[Title]) OR vocational recovery[Title])) AND ((((((((((“prognosis”[MeSH Major Topic]) OR “forecasting”[MeSH Major Topic]) OR “algorithms”[MeSH Major Topic]) OR “models, psychological”[MeSH Major Topic]) OR “models, statistical”[MeSH Major Topic]) OR “risk factors”[MeSH Major Topic]) OR predict*[Title]) OR prognos*[Title]) OR model*[Title]) OR risk predict*[Title])**	328	10:20:08
#3	Add	Search **(((((((((“prognosis”[MeSH Major Topic]) OR “forecasting”[MeSH Major Topic]) OR “algorithms”[MeSH Major Topic]) OR “models, psychological”[MeSH Major Topic]) OR “models, statistical”[MeSH Major Topic]) OR “risk factors”[MeSH Major Topic]) OR predict*[Title]) OR prognos*[Title]) OR model*[Title]) OR risk predict*[Title]**	1165269	10:19:42
#2	Add	Search **((((((((((“outcome assessment, health care”[MeSH Major Topic]) OR “treatment outcome”[MeSH Major Topic]) OR “quality of life”[MeSH Major Topic]) OR “mental health recovery”[MeSH Major Topic]) OR “rehabilitation, vocational”[MeSH Major Topic]) OR relapse prevention[Title]) OR treatment resis*[Title]) OR outcome*[Title]) OR remission[Title]) OR recovery[Title]) OR “quality of life”[Title]) OR social recovery[Title]) OR vocational recovery[Title]**	526390	10:11:26
#1	Add	Search **psychosis[Title/Abstract]**	36962	10:05:33

Google Scholar Search:

Allintitle: psychosis AND (predict OR prognos OR model)

Appendix 2 R code for Chapter 3

```
library(doParallel)
library(mice)
library(readr)
library(caret)
library(CalibrationCurves)
library(pmsampsize)
library(pROC)
library(gtools)
library(dcurves)
library(psfmi)
library(dplyr)

#enable multicore (windows) which roughly halfs time for analysis runs
cl <- makeCluster(detectCores(), type = 'PSOCK')
registerDoParallel(cl)

options(max.print = 1000000)

#don't use scientific notation (revert back with options(scipen=0)
options(scipen = 999)
options(digits = 4)

#sample size with 14 expert chosen predictors
pmsampsize(
  type = "b",
  rsquared = 0.25,
  parameters = 14,
  shrinkage = 0.9,
  prevalence = 0.5
)

#Load study data
#EDEN
eden = read_csv("eden_all.csv")
eden$Study = NULL
eden = csv_to_factor(eden)
#
tempData <- mice(eden,m=10,seed=987)

# 10-fold CV repeated 5 times
control <- trainControl(
  method = "repeatedcv",
  number = 10,
  repeats = 5,
  classProbs=TRUE,
  summaryFunction=twoClassSummary,
  savePredictions = T)

finalModels = list()
crossValModels = list()
for (i in seq(1:tempData$m))
{
  #Get imputed data
  eden_imp = complete(tempData,i)
  #just take the columns we are using except outcome as standardising first
  eden_imp_exp = eden_imp[,c(2,15,18,25,49,53,54,55,63,72,100,111,112,120)]
  #standardise the columns before building model
  preProcValues = preProcess(eden_imp_exp, method = c("center", "scale"))
  eden_imp_exp_stand = predict(preProcValues, eden_imp_exp)
  #Add factor outcome back in
  eden_imp_exp_stand$M12_PANSS_Period_Rem = eden$M12_PANSS_Period_Rem
  #Remove rows with missing outcomes
  eden_imp_exp_stand_MID = eden_imp_exp_stand[complete.cases(eden_imp_exp_stand), ]
  #need to return design matrix to reestimate intercept
  finalModels[[i]] = glm(M12_PANSS_Period_Rem ~ ., data = eden_imp_exp_stand_MID, family = "binomial", x = T)
  crossValModels[[i]] = train(M12_PANSS_Period_Rem ~ ., data = eden_imp_exp_stand_MID, method = "glm", metric =
"ROC", trControl=control, na.action=na.pass)
}

#Export data
#write_csv(finalModels[[1]]$data, "dev_eden_1.csv", na = "")
#change number between 1 to 10 for all imputed datasets

#Export data for standardising
#eden_imp = complete(tempData,1)
```

```
#just take the columns we are using including outcome
#eden_imp_exp = eden_imp[,c(2,15,18,25,49,53,54,55,63,72,100,111,112,120,126)]


#Get AUCs means and SEs across folds for each MI dataset
internalMeanROCValues = list()
internalMeanROCSEs = list()
for (i in seq(1:tempData$m))
{
  #hacky way to get ROCs for train objects - ignore GLM2 just there because can't use resamples for just one train object
  miResamps = resamples(list(GLM=crossValModels[[i]], GLM2 = crossValModels[[1]]))
  internalMeanROCValues[[i]] = mean(miResamps$values[,2])
  internalMeanROCSEs[[i]] = sqrt(var(miResamps$values[,2]))/sqrt(length(miResamps$values[,2]))
}
#Correctly pooled C-statistic and 95% CI using Rubin's Rules with logit transformation
pool_auc(internalMeanROCValues, internalMeanROCSEs, nimp = 10, log_auc = T)

#Alternative way to above calculating own ROCs
internalMeanROCValuesAlt = list()
internalMeanROCSEsAlt = list()
#and to calculate mean calibration ints and slopes per fold across MI
internalMeanCalIntValues = list()
internalMeanCalIntSEs = list()
internalMeanCalSlopeValues = list()
internalMeanCalSlopeSEs = list()

for (i in seq(1:tempData$m))
{
  rocs = crossValModels[[i]]$pred %>%
    group_by(Resample) %>%
    summarise(aucs = as.vector(roc(
      predictor = No,
      response = obs,
      ci = T,
      levels = c("No", "Yes"),
      direction = ">"
    )$auc))
  internalMeanROCValuesAlt[[i]] = mean(rocs$aucs)
  internalMeanROCSEsAlt[[i]] = sqrt(var(rocs$aucs))/sqrt(length(rocs$aucs))

  curves = crossValModels[[i]]$pred %>%
    group_by(Resample) %>%
    summarise(ints = as.vector(
      val.prob.ci.3(p=No, y=as.character(obs)=="No", g=5, pl=T, logistic.cal = T, lty.log=9,
                    col.log="red", lwd.log=1.5, col.ideal="blue", lwd.ideal=0.5, smooth = "rcs")$stats[["Intercept"]]
    ),
              slopes = as.vector(
                val.prob.ci.3(p=No, y=as.character(obs)=="No", g=5, pl=T, logistic.cal = T, lty.log=9,
                              col.log="red", lwd.log=1.5, col.ideal="blue", lwd.ideal=0.5, smooth = "rcs")$stats[["Slope"]]
              ))

  internalMeanCalIntValues[[i]] = mean(curves$ints)
  internalMeanCalIntSEs[[i]] = sqrt(var(curves$ints))/sqrt(length(curves$ints))
  internalMeanCalSlopeValues[[i]] = mean(curves$slopes)
  internalMeanCalSlopeSEs[[i]] = sqrt(var(curves$slopes))/sqrt(length(curves$slopes))
}
#Correctly pooled C-statistic and 95% CI using Rubin's Rules with logit transformation
pool_auc(internalMeanROCValuesAlt, internalMeanROCSEsAlt, nimp = 10, log_auc = T)
#Correctly pooled internal Calibration intercept and SE - should be zero for internal validation
rubin.rules(unlist(internalMeanCalIntValues), unlist(internalMeanCalIntSEs))
pool_auc_2(est_auc = internalMeanCalIntValues, est_se = internalMeanCalIntSEs, nimp = 10, log_auc = F)
#Correctly pooled internal Calibration slope and SE
rubin.rules(unlist(internalMeanCalSlopeValues), unlist(internalMeanCalSlopeSEs))
pool_auc_2(est_auc = internalMeanCalSlopeValues, est_se = internalMeanCalSlopeSEs, nimp = 10, log_auc = F)

#pool AUCs using Rubin's Rules (concatenated predictions)
internalROCs = list()
internalROCsValues = list()
internalROCsSEs = list()
for(i in seq(1:tempData$m))
{
  internalROCs[[i]] = roc(
    predictor = crossValModels[[i]]$pred$No,
    response = crossValModels[[i]]$pred$obs,
    ci = T,
    levels = c("No", "Yes"),
    direction = ">"
  )
  internalROCsValues[[i]] = internalROCs[[i]]$auc
```

```
  internalROCsSEs[[i]] = (internalROCs[[i]]$auc - internalROCs[[i]]$ci[1])/1.96
}
#Correctly pooled C-statistic and 95% CI using Rubin's Rules with logit transformation
pool_auc(internalROCsValues, internalROCsSEs, nimp = 10, log_auc = T)

#Individual permutation tests for each multiple imputation dataset
psPermInternal = list()
set.seed(987)
for(i in seq(1:tempData$m))
{
  #permutation p value
  auc_null = NULL
  for(j in seq (1:10001))
  {
    perm = permute(crossValModels[[i]]$pred$obs)
    auc_null = c(auc_null, roc(predictor = crossValModels[[i]]$pred$No, response = perm, levels=c("No", "Yes"),
direction=">")$auc)
  }
  psPermInternal[[i]] = (1+sum(auc_null >= internalROCsValues[[i]]))/10001
}
psPermInternal

internalCalIntValues = list()
internalCalIntSEs = list()
internalCalSlopeValues = list()
internalCalSlopeSEs = list()
for(i in seq(1:tempData$m))
{
  internalCal = val.prob.ci.3(p=crossValModels[[i]]$pred$No, y=as.character(crossValModels[[i]]$pred$obs)=="No", g=5,
logistic.cal = T, lty.log=9,
                        col.log="red", lwd.log=1.5, col.ideal="blue", lwd.ideal=0.5, smooth = "rcs")

  internalCalIntValues[[i]] = internalCal$stats[["Intercept"]]
  internalCalIntSEs[[i]] = (internalCal$stats[["Intercept"]] - internalCal$cl.interc[1])/1.96
  internalCalSlopeValues[[i]] = internalCal$stats[["Slope"]]
  internalCalSlopeSEs[[i]] = (internalCal$stats[["Slope"]] - internalCal$cl.slope[[1]])/1.96
}

#Correctly pooled internal Calibration intercept and SE - should be zero for internal validation
rubin.rules(unlist(internalCalIntValues), unlist(internalCalIntSEs))
pool_auc_2(est_auc = internalCalIntValues, est_se = internalCalIntSEs, nimp = 10, log_auc = F)
#Correctly pooled internal Calibration slope and SE
rubin.rules(unlist(internalCalSlopeValues), unlist(internalCalSlopeSEs))
pool_auc_2(est_auc = internalCalSlopeValues, est_se = internalCalSlopeSEs, nimp = 10, log_auc = F)


#combined data
internalPreds = NULL
internalOutcomes = NULL
for(i in seq(1:tempData$m))
{
  internalPreds = c(internalPreds,crossValModels[[i]]$pred$No)
  internalOutcomes = c(internalOutcomes, as.character(crossValModels[[i]]$pred$obs))
}

#calibration plot
#increase memory allocated to R
memory.limit(size=56000)
intval_cal = val.prob.ci.2(p=internalPreds, y=internalOutcomes=="No", g=10, logistic.cal = T, lty.log=9,
            col.log="red", lwd.log=1.5, col.ideal="blue", lwd.ideal=0.5, dostats = T, smooth = "rcs")
#shrinkage factor = 0.8381
shrinkage = intval_cal["Slope"]

#Pool results to get predictor estimates based on Rubin's rule (coefficients reversed as using to predict non-remission)
View(-summary(pool(finalModels), conf.int = T, exponentiate = F, conf.level = 0.95)[,c(2,7,8)])
View(exp(-summary(pool(finalModels), conf.int = T, exponentiate = F, conf.level = 0.95)[,c(2,7,8)]))
#With shrinkage applied (N.B. intercept still to be re-estimated)
View(-summary(pool(finalModels), conf.int = T, exponentiate = F, conf.level = 0.95)[,c(2,7,8)]*shrinkage)
View(exp(-summary(pool(finalModels), conf.int = T, exponentiate = F, conf.level = 0.95)[,c(2,7,8)]*shrinkage))

#for each imputed dataset recalculate intercept with pooled shrunk final coefficients
#then average the intercepts to get final intercept
shrunk_coefs = summary(pool(finalModels))[[2]]*shrinkage
models = list()
#alternatively using rms package as per Steyerberg - see commented code
#library(rms)
intercepts = NULL
for(i in seq(1:tempData$m))
{
```

```
  shrunk_LP = finalModels[[i]]$x[,2:15] %*% shrunk_coefs[2:15]
  models[[i]] = glm(finalModels[[i]]$y ~ offset(shrunk_LP), family = "binomial")
  #intercepts = c(intercepts, lrm.fit(y = finalModels[[i]]$y, offset= finalModels[[i]]$x[,2:15] %*%
shrunk_coefs[2:15])$coef[1])
}
#pooled recalculated intercept
recalc_intercept = summary(pool(models))[[2]]
#mean(intercepts)

#Get a GLM model
finalModel = finalModels[[1]]
#replace coefficients with pooled ones shrunk coefs
finalModel$coefficients = shrunk_coefs
#replace intercept with pooled recalculated intercept
finalModel$coefficients[1] = recalc_intercept

############################################################
#Additional code for alternative internal validation via Harrell's bootstrap method
#and heuristic Van Houwelingen's shrinkage factor

finalModelsAlt = list()
vanH = list()

apparentROC = list()
apparentSlopeModel = list()
boot_results = list()

for (i in seq(1:tempData$m))
{
  #Get imputed data
  eden_imp = complete(tempData,i)
  #just take the columns we are using except outcome as standardising first
  eden_imp_exp = eden_imp[,c(2,15,18,25,49,53,54,55,63,72,100,111,112,120)]
  #standardise the columns before building model
  preProcValues = preProcess(eden_imp_exp, method = c("center", "scale"))
  eden_imp_exp_stand = predict(preProcValues, eden_imp_exp)
  #Add factor outcome back in
  eden_imp_exp_stand$M12_PANSS_Period_Rem = eden$M12_PANSS_Period_Rem
  #Remove rows with missing outcomes
  eden_imp_exp_stand_MID = eden_imp_exp_stand[complete.cases(eden_imp_exp_stand), ]
  #need to return design matrix to reestimate intercept
  finalModelsAlt[[i]] = glm(M12_PANSS_Period_Rem ~ ., data = eden_imp_exp_stand_MID, family = "binomial", x = T, y = T)
  #Get apparent auc
  pred_prob = predict(finalModelsAlt[[i]],type="response")
  apparentROC[[i]] = roc(eden_imp_exp_stand_MID$M12_PANSS_Period_Rem~pred_prob,ci=TRUE,levels = c("No",
"Yes"),direction = "<")
  #Get calibration slope models
  pred_LP = predict(finalModelsAlt[[i]],type="link")
  apparentSlopeModel[[i]] = glm(eden_imp_exp_stand_MID$M12_PANSS_Period_Rem ~
pred_LP,family="binomial",x=TRUE,y=TRUE)

  # Calculate heuristic shrinkage

  # Obtain chi2
  null_model <- glm(M12_PANSS_Period_Rem~1, data = eden_imp_exp_stand_MID, family="binomial")
  chisq <- anova(null_model,finalModelsAlt[[i]],test="Chisq")
  chi2 <- chisq$Deviance[2]
  # display degrees of freedom (df)
  df_log <- chisq$Df[2]
  # Combining these elements we can obtain the heuristic
  # shrinkage of Van Houwelingen which is given by (chi2 - df)/chi2
  # where df = 10 for the predictors here
  vanH[[i]] <- (chi2 - df_log)/chi2

  #Do the whole process with bootstrapping to compare with cross validation
  boot_results[[i]] = manual_boot(eden_imp_exp_stand_MID,500)

}

#mean heuristic shrinkage across MI
mean(unlist(vanH))

#bootstrap values
bootROCValues = list()
bootSlopeValues = list()
for (i in seq(1:tempData$m))
{
  bootROCValues[[i]] = apparentROC[[i]]$auc - (mean(boot_results[[i]]$app_c_stat)-mean(boot_results[[i]]$test_c_stat)) #
c-stat
```

```
  bootSlopeValues[[i]] = apparentSlopeModel[[i]]$coef[2]- (mean(boot_results[[i]]$app_c_slope)-
mean(boot_results[[i]]$test_c_slope)) # c-slope
}

#Get the mean bootstrapped values across the multiple imputations
#mean c-statistic
mean(unlist(bootROCValues))
#mean calibration slope
mean(unlist(bootSlopeValues))

############################################################
#Load study data
#standardising outlook test data on itself
#outlook
outlook = read_csv("outlook_all.csv")
outlook$Study = NULL
#remove correlated variables - identified as problem in MICE
outlook$PCT_Average_Rank_2007 = NULL
outlook$PCT_Employment_Scale_2007 = NULL
outlook$PCT_Extent_2007 = NULL
outlook$PCT_Local_Concentration_2007 = NULL
outlook$PCT_Income_Scale_2007 = NULL
outlook = csv_to_factor(outlook)

tempData2 <- mice(outlook,m=10,seed=987)

Results = list()
Obs = list()
DUP = list()
for (i in seq(1:tempData2$m))
{
  #Get imputed data
  outlook_imp = complete(tempData2,i)
  #just take the columns we are using except outcome as standardising first
  outlook_imp_exp = outlook_imp[,c(2,15,18,25,49,53,54,55,63,72,100,111,112,120)]
  #standardise the columns before building model
  preProcValues = preProcess(outlook_imp_exp, method = c("center", "scale"))
  outlook_imp_exp_stand = predict(preProcValues, outlook_imp_exp)
  #Add factor outcome back in
  outlook_imp_exp_stand$M12_PANSS_Period_Rem = outlook$M12_PANSS_Period_Rem
  #Remove rows with missing outcomes
  outlook_imp_exp_stand_MID = outlook_imp_exp_stand[complete.cases(outlook_imp_exp_stand), ]
  Results[[i]] = predict(finalModel, outlook_imp_exp_stand_MID, type = "response", na.action = na.pass)
  #Predicting No
  Results[[i]] = 1-Results[[i]]
  Obs[[i]] =  outlook_imp_exp_stand_MID$M12_PANSS_Period_Rem
  DUP[[i]] = outlook_imp_exp_stand_MID$ADJ_DUP
}

#pool AUCs using Rubin's Rules
externalROCs = list()
externalROCsValues = list()
externalROCsSEs= list()
for(i in seq(1:tempData2$m))
{
  externalROCs[[i]] = roc(
    predictor = Results[[i]],
    response = Obs[[i]],
    ci = T,
    levels = c("No", "Yes"),
    direction = ">"
  )
  externalROCsValues[[i]] = externalROCs[[i]]$auc
  externalROCsSEs[[i]] = (externalROCs[[i]]$auc - externalROCs[[i]]$ci[1])/1.96
}
#Correctly pooled C-statistic and 95% CI using Rubin's Rules with logit transformation
pool_auc(externalROCsValues, externalROCsSEs, nimp = 10, log_auc = T)

#Individual permutation tests for each multiple imputation dataset
psPermExternal = list()
set.seed(987)
for(i in seq(1:tempData2$m))
{
  #permutation p value
  auc_null = NULL
  for(j in seq (1:10001))
  {
    perm = permute(Obs[[i]])
    auc_null = c(auc_null, roc(predictor = Results[[i]], response = perm, levels=c("No", "Yes"), direction=">")$auc)
```

```
  }
  psPermExternal[[i]] = (1+sum(auc_null >= externalROCsValues[[i]]))/10001
}
psPermExternal

externalCalIntValues = list()
externalCalIntSEs = list()
externalCalSlopeValues = list()
externalCalSlopeSEs = list()
for(i in seq(1:tempData2$m))
{
  externalCal = val.prob.ci.3(p=Results[[i]], y=Obs[[i]]=="No", g=5, logistic.cal = T, lty.log=9,
                              col.log="red", lwd.log=1.5, col.ideal="blue", lwd.ideal=0.5)

  externalCalIntValues[[i]] = externalCal$stats[["Intercept"]]
  externalCalIntSEs[[i]] = (externalCal$stats[["Intercept"]] - externalCal$cl.interc[1])/1.96
  externalCalSlopeValues[[i]] = externalCal$stats[["Slope"]]
  externalCalSlopeSEs[[i]] = (externalCal$stats[["Slope"]] - externalCal$cl.slope[[1]])/1.96
}

#Correctly pooled Calibration intercept and SE
rubin.rules(unlist(externalCalIntValues), unlist(externalCalIntSEs))
pool_auc_2(est_auc = externalCalIntValues, est_se = externalCalIntSEs, nimp = 10, log_auc = F)

#Correctly pooled Calibration slope and SE
rubin.rules(unlist(externalCalSlopeValues), unlist(externalCalSlopeSEs))
pool_auc_2(est_auc = externalCalSlopeValues, est_se = externalCalSlopeSEs, nimp = 10, log_auc = F)

#calibration plot - ignore confidence intervals, use above via Rubin's rules instead
#first imputed dataset - don't display stats
pdf("figure2_correct.pdf", width = 7, height = 7)
val.prob.ci.2(p=Results[[1]], y=Obs[[1]]=="No", g=5, logistic.cal = T, lty.log=9,
              col.log="red", lwd.log=1.5, col.ideal="blue", lwd.ideal=0.5, statloc = F)
dev.off()

#combined data
#no confidence intervals for dca so combining data just gives average
externalPreds = NULL
externalOutcomes = NULL
externalDUP = NULL
for(i in seq(1:tempData2$m))
{
  externalPreds = c(externalPreds,Results[[i]])
  externalOutcomes = c(externalOutcomes, as.character(Obs[[i]]))
  externalDUP = c(externalDUP, DUP[[i]])
}

#dca
dca_ext = NULL
dca_ext$M12_PANSS_Period_Rem = as.integer(externalOutcomes=="No")
dca_ext$Model = externalPreds
dca_ext$DUP = externalDUP
#get data across all thresholds
dca_ext_calc = dca(M12_PANSS_Period_Rem ~ Model + DUP, data = as.data.frame(dca_ext), as_probability =  "DUP",
                   thresholds = seq(0.3, 0.77, by = 0.01))
pdf("figure3.pdf", width = 7, height = 7)
dca_ext_calc %>%
  plot(smooth = TRUE)
dev.off()

#Write dca to csv
write_csv(dca_ext_calc$dca,"dca.csv")

############################################################
#standardising test data using EDEN training values

#get preprocessing rules for standardisation using first imputed EDEN dataset
eden_imp2 = complete(tempData,1)
#just take the columns we are using except outcome as standardising first
eden_imp_exp2 = eden_imp2[,c(2,15,18,25,49,53,54,55,63,72,100,111,112,120)]
#standardise the columns before building model
preProcValuesEDEN = preProcess(eden_imp_exp2, method = c("center", "scale"))

Results2 = list()
Obs2 = list()
DUP2 = list()
for (i in seq(1:tempData2$m))
{
  #Get imputed data
```

```
  outlook_imp2 = complete(tempData2,i)
  #just take the columns we are using except outcome as standardising first
  outlook_imp_exp2 = outlook_imp2[,c(2,15,18,25,49,53,54,55,63,72,100,111,112,120)]

  #standardise using preprocessing rules from EDEN first imputed dataset
  outlook_imp_exp_stand2 = predict(preProcValuesEDEN, outlook_imp_exp2)
  #Add factor outcome back in
  outlook_imp_exp_stand2$M12_PANSS_Period_Rem = outlook$M12_PANSS_Period_Rem
  #Remove rows with missing outcomes
  outlook_imp_exp_stand_MID2 = outlook_imp_exp_stand2[complete.cases(outlook_imp_exp_stand2), ]
  Results2[[i]] = predict(finalModel, outlook_imp_exp_stand_MID2, type = "response", na.action = na.pass)
  #Predicting No
  Results2[[i]] = 1-Results2[[i]]
  Obs2[[i]] =  outlook_imp_exp_stand_MID2$M12_PANSS_Period_Rem
  DUP2[[i]] = outlook_imp_exp_stand_MID2$ADJ_DUP
}

#pool AUCs using Rubin's Rules
externalROCs2 = list()
externalROCsValues2 = list()
externalROCsSEs2= list()
for(i in seq(1:tempData2$m))
{
  externalROCs2[[i]] = roc(
    predictor = Results[[i]],
    response = Obs[[i]],
    ci = T,
    levels = c("No", "Yes"),
    direction = ">"
  )
  externalROCsValues2[[i]] = externalROCs2[[i]]$auc
  externalROCsSEs2[[i]] = (externalROCs2[[i]]$auc - externalROCs2[[i]]$ci[1])/1.96
}
#Correctly pooled C-statistic and 95% CI using Rubin's Rules with logit transformation
pool_auc(externalROCsValues2, externalROCsSEs2, nimp = 10, log_auc = T)

#Individual permutation tests for each multiple imputation dataset
psPermExternal2 = list()
set.seed(987)
for(i in seq(1:tempData2$m))
{
  #permutation p value
  auc_null = NULL
  for(j in seq (1:10001))
  {
    perm = permute(Obs2[[i]])
    auc_null = c(auc_null, roc(predictor = Results2[[i]], response = perm, levels=c("No", "Yes"), direction=">")$auc)
  }
  psPermExternal2[[i]] = (1+sum(auc_null >= externalROCsValues2[[i]]))/10001
}
psPermExternal2

externalCalIntValues2 = list()
externalCalIntSEs2 = list()
externalCalSlopeValues2 = list()
externalCalSlopeSEs2 = list()
for(i in seq(1:tempData2$m))
{
  externalCal2 = val.prob.ci.3(p=Results2[[i]], y=Obs2[[i]]=="No", g=5, logistic.cal = T, lty.log=9,
                         col.log="red", lwd.log=1.5, col.ideal="blue", lwd.ideal=0.5)

  externalCalIntValues2[[i]] = externalCal2$stats[["Intercept"]]
  externalCalIntSEs2[[i]] = (externalCal2$stats[["Intercept"]] - externalCal2$cl.interc[1])/1.96
  externalCalSlopeValues2[[i]] = externalCal2$stats[["Slope"]]
  externalCalSlopeSEs2[[i]] = (externalCal2$stats[["Slope"]] - externalCal2$cl.slope[[1]])/1.96
}

#Correctly pooled Calibration intercept and SE
rubin.rules(unlist(externalCalIntValues2), unlist(externalCalIntSEs2))
pool_auc_2(est_auc = externalCalIntValues2, est_se = externalCalIntSEs2, nimp = 10, log_auc = F)

#Correctly pooled Calibration slope and SE
rubin.rules(unlist(externalCalSlopeValues2), unlist(externalCalSlopeSEs2))
pool_auc_2(est_auc = externalCalSlopeValues2, est_se = externalCalSlopeSEs2, nimp = 10, log_auc = F)

#calibration plot - ignore confidence intervals, use above via Rubin's rules instead
#first imputed dataset - don't display stats
pdf("figure2_correct2.pdf", width = 7, height = 7)
val.prob.ci.2(p=Results2[[1]], y=Obs2[[1]]=="No", g=5, logistic.cal = T, lty.log=9,
```

```
          col.log="red", lwd.log=1.5, col.ideal="blue", lwd.ideal=0.5, statloc = F)
dev.off()

#combined data
#no confidence intervals for dca so combining data just gives average
externalPreds2 = NULL
externalOutcomes2 = NULL
externalDUP2 = NULL
for(i in seq(1:tempData2$m))
{
  externalPreds2 = c(externalPreds2,Results2[[i]])
  externalOutcomes2 = c(externalOutcomes2, as.character(Obs2[[i]]))
  externalDUP2 = c(externalDUP2, DUP2[[i]])
}

#dca
dca_ext2 = NULL
dca_ext2$M12_PANSS_Period_Rem = as.integer(externalOutcomes2=="No")
dca_ext2$Model = externalPreds2
dca_ext2$DUP = externalDUP2
#get data across all thresholds
dca_ext_calc2 = dca(M12_PANSS_Period_Rem ~ Model + DUP, data = as.data.frame(dca_ext2), as_probability =  "DUP",
                thresholds = seq(0.3, 0.77, by = 0.01))
pdf("figure3_2.pdf", width = 7, height = 7)
dca_ext_calc2 %>%
  plot(smooth = TRUE)
dev.off()

#Write dca2 to csv
write_csv(dca_ext_calc2$dca,"dca2.csv")

#dca first imputed dataset
dca_ext2_1 = NULL
dca_ext2_1$M12_PANSS_Period_Rem = as.integer(Obs2[[1]]=="No")
dca_ext2_1$Model = Results2[[1]]
dca_ext2_1$DUP = DUP2[[1]]

dca_ext_calc2_1 = dca(M12_PANSS_Period_Rem ~ Model + DUP, data = as.data.frame(dca_ext2_1), as_probability =
"DUP",
                thresholds = seq(0.3, 0.77, by = 0.01))

pdf("figure3_2_1.pdf", width = 7, height = 7)
dca_ext_calc2_1 %>%
  plot(smooth = TRUE)
dev.off()

#Write dca2 to csv
write_csv(dca_ext_calc2_1$dca,"dca2_1.csv")

#############################################################################
#Demographic comparisons
#############################################################################
#summary statistics
#Age
eden_final_demographics = eden[complete.cases(eden[ , "M12_PANSS_Period_Rem"]),]
outlook_final_demographics = outlook[complete.cases(outlook[ , "M12_PANSS_Period_Rem"]),]

summary(eden$Age_Entry)
mean(eden$Age_Entry, na.rm = T)
sd(eden$Age_Entry, na.rm = T)

summary(eden_final_demographics$Age_Entry)
mean(eden_final_demographics$Age_Entry, na.rm = T)
sd(eden_final_demographics$Age_Entry, na.rm = T)

summary(outlook$Age_Entry)
mean(outlook$Age_Entry, na.rm = T)
sd(outlook$Age_Entry, na.rm = T)

summary(outlook_final_demographics$Age_Entry)
mean(outlook_final_demographics$Age_Entry, na.rm = T)
sd(outlook_final_demographics$Age_Entry, na.rm = T)

summary(aov(y~group, data = data.frame(group=factor(rep(1:4,
c(1027,673,399,191))),y=c(eden$Age_Entry,eden_final_demographics$Age_Entry,
                                                    outlook$Age_Entry,outlook_final_demographics$Age_Entry))))

#Sex
summary(eden$Sex)
```

```
summary(eden_final_demographics$Sex)
summary(outlook$Sex)
summary(outlook_final_demographics$Sex)

chisq.test(as.table(rbind(c(318,709),c(210,463),c(153,246),c(73,118))), correct = F)

#EET
summary(eden$BL_EET)
summary(eden_final_demographics$BL_EET)
summary(outlook$BL_EET)
summary(outlook_final_demographics$BL_EET)

chisq.test(as.table(rbind(c(589,284),c(383,190),c(225,174),c(106,85))), correct = F)

#Qualification
summary(as.factor(eden$Qualification_Level_Ordinal))
summary(as.factor(eden_final_demographics$Qualification_Level_Ordinal))
summary(as.factor(outlook$Qualification_Level_Ordinal))
summary(as.factor(outlook_final_demographics$Qualification_Level_Ordinal))

chisq.test(as.table(rbind(c(245,399,262,98),c(156,255,173,74),c(89,130,92,69),c(40,67,46,35))), correct = F)

#DUP
summary(eden$ADJ_DUP)
mean(eden$ADJ_DUP, na.rm = T)
sd(eden$ADJ_DUP, na.rm = T)

summary(eden_final_demographics$ADJ_DUP)
mean(eden_final_demographics$ADJ_DUP, na.rm = T)
sd(eden_final_demographics$ADJ_DUP, na.rm = T)

summary(outlook$ADJ_DUP)
mean(outlook$ADJ_DUP, na.rm = T)
sd(outlook$ADJ_DUP, na.rm = T)

summary(outlook_final_demographics$ADJ_DUP)
mean(outlook_final_demographics$ADJ_DUP, na.rm = T)
sd(outlook_final_demographics$ADJ_DUP, na.rm = T)

summary(aov(y~group, data = data.frame(group=factor(rep(1:4, c(1027,673,399,191))),y=c(eden$ADJ_DUP,
                                                    eden_final_demographics$ADJ_DUP,
                                                    outlook$ADJ_DUP,
                                                    outlook_final_demographics$ADJ_DUP))))

#Deprivation
summary(eden$PCT_Average_Scrore_2007)
mean(eden$PCT_Average_Scrore_2007, na.rm = T)
sd(eden$PCT_Average_Scrore_2007, na.rm = T)

summary(eden_final_demographics$PCT_Average_Scrore_2007)
mean(eden_final_demographics$PCT_Average_Scrore_2007, na.rm = T)
sd(eden_final_demographics$PCT_Average_Scrore_2007, na.rm = T)

summary(outlook$PCT_Average_Scrore_2007)
mean(outlook$PCT_Average_Scrore_2007, na.rm = T)
sd(outlook$PCT_Average_Scrore_2007, na.rm = T)

summary(outlook_final_demographics$PCT_Average_Scrore_2007)
mean(outlook_final_demographics$PCT_Average_Scrore_2007, na.rm = T)
sd(outlook_final_demographics$PCT_Average_Scrore_2007, na.rm = T)

summary(aov(y~group, data = data.frame(group=factor(rep(1:4,
c(1027,673,399,191))),y=c(eden$PCT_Average_Scrore_2007,
                                                    eden_final_demographics$PCT_Average_Scrore_2007,
                                                    outlook$PCT_Average_Scrore_2007,
                                                    outlook_final_demographics$PCT_Average_Scrore_2007))))

#PANSS Totals
eden$BL_PANSS_Total = rowSums(eden[,53:82])
outlook$BL_PANSS_Total = rowSums(outlook[,53:82])
eden_final_demographics$BL_PANSS_Total = rowSums(eden_final_demographics[,53:82])
outlook_final_demographics$BL_PANSS_Total = rowSums(outlook_final_demographics[,53:82])

summary(eden$BL_PANSS_Total)
mean(eden$BL_PANSS_Total, na.rm = T)
sd(eden$BL_PANSS_Total, na.rm = T)

summary(eden_final_demographics$BL_PANSS_Total)
mean(eden_final_demographics$BL_PANSS_Total, na.rm = T)
```

```
sd(eden_final_demographics$BL_PANSS_Total, na.rm = T)

summary(outlook$BL_PANSS_Total)
mean(outlook$BL_PANSS_Total, na.rm = T)
sd(outlook$BL_PANSS_Total, na.rm = T)

summary(outlook_final_demographics$BL_PANSS_Total)
mean(outlook_final_demographics$BL_PANSS_Total, na.rm = T)
sd(outlook_final_demographics$BL_PANSS_Total, na.rm = T)

summary(aov(y~group, data = data.frame(group=factor(rep(1:4, c(1027,673,399,191))),y=c(eden$BL_PANSS_Total,
                                                                                       eden_final_demographics$BL_PANSS_Total,
                                                                                       outlook$BL_PANSS_Total,
                                                                                       outlook_final_demographics$BL_PANSS_Total))))

#My custom functions
##############################################################################################

#change char cols to factor
csv_to_factor <- function(imported_csv)
{
  cols_char_csv = colnames(imported_csv[, sapply(imported_csv, class) == 'character'])
  for (i in seq(1:length(cols_char_csv)))
  {
    imported_csv[[cols_char_csv[i]]] = as.factor(imported_csv[[cols_char_csv[i]]])
  }
  return(imported_csv)
}

#Calibration Curves function also returning confidence intervals around intercept, slope and c-statistic
#Only confirmed to work with default options
val.prob.ci.3 <- function(p, y, logit, group, weights = rep(1, length(y)), normwt = F, pl = T,
                          smooth = c("loess","rcs",F), CL.smooth="fill",CL.BT=F,lty.smooth=1,col.smooth="black",lwd.smooth=1,
                          nr.knots=5,logistic.cal = F,lty.log=1,col.log="black",lwd.log=1, xlab = "Predicted probability", ylab =
                            "Observed proportion", xlim = c(-0.02, 1),ylim = c(-0.15,1), m, g, cuts, emax.lim = c(0, 1),
                          legendloc =  c(0.50 , 0.27), statloc = c(0,.85),dostats=T,cl.level=0.95,method.ci="pepe",roundstats=2,
                          riskdist = "predicted", cex=0.75,cex.leg = 0.75, connect.group =
                            F, connect.smooth = T, g.group = 4, evaluate = 100, nmin = 0, d0lab="0", d1lab="1", cex.d01=0.7,
                          dist.label=0.04, line.bins=-.05, dist.label2=.03, cutoff, las=1, length.seg=1,
                          y.intersp=1,lty.ideal=1,col.ideal="red",lwd.ideal=1,...)
{
  if(smooth[1]==F){smooth <- "F"}
  smooth <- match.arg(smooth)
  if(!missing(p))
    if(any(!(p>=0 | p<=1))){stop("Probabilities can not be > 1 or < 0.")}
  if(missing(p))
    p <- 1/(1 + exp( - logit))
  else logit <- log(p/(1 - p))
  if(!all(y%in%0:1)){stop("The vector with the binary outcome can only contain the values 0 and 1.")}
  if(length(p) != length(y))
    stop("lengths of p or logit and y do not agree")
  names(p) <- names(y) <- names(logit) <- NULL
  if(!missing(group)) {
    if(length(group) == 1 && is.logical(group) && group)
      group <- rep("", length(y))
    if(!is.factor(group))
      group <- if(is.logical(group) || is.character(group))
        as.factor(group) else cut2(group, g =
                                   g.group)
    names(group) <- NULL
    nma <- !(is.na(p + y + weights) | is.na(group))
    ng <- length(levels(group))
  }
  else {
    nma <- !is.na(p + y + weights)
    ng <- 0
  }
  logit <- logit[nma]
  y <- y[nma]
  p <- p[nma]
  if(ng > 0) {
    group <- group[nma]
    weights <- weights[nma]
    return(val.probg(p, y, group, evaluate, weights, normwt, nmin)
    )
  }

  # Sort vector with probabilities
  y     <- y[order(p)]
```

```
logit <- logit[order(p)]
p     <- p[order(p)]


if(length(p)>5000 & smooth=="loess"){warning("Number of observations > 5000, RCS is recommended.",immediate. = T)}
if(length(p)>1000 & CL.BT==T){warning("Number of observations is > 1000, this could take a while...",immediate. = T)}


if(length(unique(p)) == 1) {
  #22Sep94
  P <- mean(y)
  lntc <- log(P/(1 - P))
  n <- length(y)
  D <- -1/n
  L01 <- -2 * sum(y * logit - log(1 + exp(logit)), na.rm = T)
  L.cal <- -2 * sum(y * lntc - log(1 + exp(lntc)), na.rm = T)
  U.chisq <- L01 - L.cal
  U.p <- 1 - pchisq(U.chisq, 1)
  U <- (U.chisq - 1)/n
  Q <- D - U

  stats <- c(0, 0.5, 0, D, 0, 1, U, U.chisq, U.p, Q, mean((y - p[
    1])^2), lntc, 0, rep(abs(p[1] - P), 2))
  names(stats) <- c("Dxy", "C (ROC)", "R2", "D", "D:Chi-sq",
                    "D:p", "U", "U:Chi-sq", "U:p", "Q", "Brier",
                    "Intercept", "Slope", "Emax", "Eavg", "ECI")
  return(stats)
}
i <- !is.infinite(logit)
nm <- sum(!i)
if(nm > 0)
  warning(paste(nm, "observations deleted from logistic calibration due to probs. of 0 or 1"))
i.2 <- i
f.or <- lrm(y[i]~logit[i])
f <- lrm.fit(logit[i], y[i])
cl.slope <- confint(f,level=cl.level)[2,]
f2 <-      lrm.fit(offset=logit[i], y=y[i])
if(f2$fail){
  warning("The lrm function did not converge when computing the calibration intercept!",immediate.=T)
  f2 <- list()
  f2$coef <- NA
  cl.interc <- rep(NA,2)
}else{
  cl.interc <- confint(f2,level=cl.level)
}
stats <- f$stats
cl.auc <- CalibrationCurves:::ci.auc(y,p,cl.level,method.ci)

n <- stats["Obs"]
predprob <- seq(emax.lim[1], emax.lim[2], by = 0.0005)
lt <- f$coef[1] + f$coef[2] * log(predprob/(1 - predprob))
calp <- 1/(1 + exp( - lt))
emax <- max(abs(predprob - calp))
if (pl) {
  plot(0.5, 0.5, xlim = xlim, ylim = ylim, type = "n", xlab = xlab,
       ylab = ylab, las=las,...)
  clip(0,1,0,1)
  abline(0, 1, lty = lty.ideal,col=col.ideal,lwd=lwd.ideal)
  do.call("clip", as.list(par()$usr))


  lt <- lty.ideal
  lw.d <- lwd.ideal
  all.col <- col.ideal
  leg <- "Ideal"
  marks <- -1
  if (logistic.cal) {
    lt <- c(lt, lty.log)
    lw.d <- c(lw.d,lwd.log)
    all.col <- c(all.col,col.log)
    leg <- c(leg, "Logistic calibration")
    marks <- c(marks, -1)
  }
  if(smooth!="F"){all.col <- c(all.col,col.smooth)}
  if (smooth=="loess") {
    #Sm <- lowess(p,y,iter=0)
    Sm <- loess(y~p,degree=2)
    Sm <- data.frame(Sm$x,Sm$fitted); Sm.01 <- Sm
```

```
    if (connect.smooth==T & CL.smooth!="fill") {
      clip(0,1,0,1)
      lines(Sm, lty = lty.smooth,lwd=lwd.smooth,col=col.smooth)
      do.call("clip", as.list(par()$usr))
      lt <- c(lt, lty.smooth)
      lw.d <- c(lw.d,lwd.smooth)
      marks <- c(marks, -1)
    }else if(connect.smooth==F & CL.smooth!="fill"){
      clip(0,1,0,1)
      points(Sm,col=col.smooth)
      do.call("clip", as.list(par()$usr))
      lt <- c(lt, 0)
      lw.d <- c(lw.d,1)
      marks <- c(marks, 1)
    }
    if(CL.smooth==T | CL.smooth=="fill"){
      to.pred <- seq(min(p),max(p),length=200)
      if(CL.BT==T){
        cat("Bootstrap samples are being generated.\n\n\n")

        replicate(2000,CalibrationCurves:::BT.samples(y,p,to.pred)) -> res.BT
        apply(res.BT,1,quantile,c(0.025,0.975)) -> CL.BT
        colnames(CL.BT) <- to.pred

        if(CL.smooth=="fill"){
          clip(0,1,0,1)
          polygon(x = c(to.pred, rev(to.pred)), y = c(CL.BT[2,],
                                                      rev(CL.BT[1,])),
                  col = rgb(177, 177, 177, 177, maxColorValue = 255), border = NA)
          if (connect.smooth==T) {
            lines(Sm, lty = lty.smooth,lwd=lwd.smooth,col=col.smooth)
            lt <- c(lt, lty.smooth)
            lw.d <- c(lw.d,lwd.smooth)
            marks <- c(marks, -1)
          }else if(connect.smooth==F){
            points(Sm,col=col.smooth)
            lt <- c(lt, 0)
            lw.d <- c(lw.d,1)
            marks <- c(marks, 1)
          }
          do.call("clip", as.list(par()$usr))
          leg <- c(leg, "Flexible calibration (Loess)")
        }else{

          clip(0,1,0,1)

lines(to.pred,CL.BT[1,],lty=2,lwd=1,col=col.smooth);clip(0,1,0,1);lines(to.pred,CL.BT[2,],lty=2,lwd=1,col=col.smooth)
          do.call("clip", as.list(par()$usr))
          leg <- c(leg,"Flexible calibration (Loess)","CL flexible")
          lt <- c(lt,2)
          lw.d <- c(lw.d,1)
          all.col <- c(all.col,col.smooth)
          marks <- c(marks,-1)
        }

      }else{
        Sm.0 <- loess(y~p,degree=2)
        predict(Sm.0,type="fitted",se=T) -> cl.loess
        clip(0,1,0,1)
        if(CL.smooth=="fill"){
          polygon(x = c(Sm.0$x, rev(Sm.0$x)), y = c(cl.loess$fit+cl.loess$se.fit*1.96,
                                                    rev(cl.loess$fit-cl.loess$se.fit*1.96)),
                  col = rgb(177, 177, 177, 177, maxColorValue = 255), border = NA)
          if (connect.smooth==T) {
            lines(Sm, lty = lty.smooth,lwd=lwd.smooth,col=col.smooth)
            lt <- c(lt, lty.smooth)
            lw.d <- c(lw.d,lwd.smooth)
            marks <- c(marks, -1)
          }else if(connect.smooth==F){
            points(Sm,col=col.smooth)
            lt <- c(lt, 0)
            lw.d <- c(lw.d,1)
            marks <- c(marks, 1)
          }
          do.call("clip", as.list(par()$usr))
          leg <- c(leg, "Flexible calibration (Loess)")
        }else{
          lines(Sm.0$x,cl.loess$fit+cl.loess$se.fit*1.96,lty=2,lwd=1,col=col.smooth)
          lines(Sm.0$x,cl.loess$fit-cl.loess$se.fit*1.96,lty=2,lwd=1,col=col.smooth)
```

```
      do.call("clip", as.list(par()$usr))
      leg <- c(leg,"Flexible calibration (Loess)","CL flexible")
      lt <- c(lt,2)
      lw.d <- c(lw.d,1)
      all.col <- c(all.col,col.smooth)
      marks <- c(marks,-1)
    }

  }

 }else{
   leg <- c(leg, "Flexible calibration (Loess)")}
 cal.smooth <- approx(Sm.01, xout = p)$y
 eavg <- mean(abs(p - cal.smooth))
 ECI <- mean((p-cal.smooth)^2)*100
}
if(smooth=="rcs"){
 par(lwd=lwd.smooth,bty="n",col=col.smooth)
 if(!is.numeric(nr.knots)){stop("Nr.knots must be numeric.")}
 if(nr.knots==5){
   tryCatch(CalibrationCurves:::rcspline.plot(p,y,model="logistic",nk=5,show="prob", statloc = "none"
                    ,add=T,showknots=F,xrange=c(min(na.omit(p)),max(na.omit(p))),lty=lty.smooth),error=function(e){
                     warning("The number of knots led to estimation problems, nk will be set to 4.",immediate. = T)
                     tryCatch(CalibrationCurves:::rcspline.plot(p,y,model="logistic",nk=4,show="prob", statloc = "none"
                                      ,add=T,showknots=F,xrange=c(min(na.omit(p)),max(na.omit(p))),lty=lty.smooth)
                              ,error=function(e){
                                warning("Nk 4 also led to estimation problems, nk will be set to 3.",immediate.=T)
                                CalibrationCurves:::rcspline.plot(p,y,model="logistic",nk=3,show="prob", statloc = "none"
                                         ,add=T,showknots=F,xrange=c(min(na.omit(p)),max(na.omit(p)))
                                         ,lty=lty.smooth)
                              })
                    })
 }else if(nr.knots==4){
   tryCatch(CalibrationCurves:::rcspline.plot(p,y,model="logistic",nk=4,show="prob", statloc = "none"
                    ,add=T,showknots=F,xrange=c(min(na.omit(p)),max(na.omit(p))),lty=lty.smooth),error=function(e){
                     warning("The number of knots led to estimation problems, nk will be set to 3.",immediate.=T)
                     CalibrationCurves:::rcspline.plot(p,y,model="logistic",nk=3,show="prob", statloc = "none"
                              ,add=T,showknots=F,xrange=c(min(na.omit(p)),max(na.omit(p))),lty=lty.smooth)
                    })
 }else if(nr.knots==3){
   tryCatch(CalibrationCurves:::rcspline.plot(p,y,model="logistic",nk=3,show="prob", statloc = "none"
                    ,add=T,showknots=F,xrange=c(min(na.omit(p)),max(na.omit(p))),lty=lty.smooth),
          error=function(e){
            stop("Nk = 3 led to estimation problems.")
          })
 }else{stop(paste("Number of knots = ",nr.knots,sep="", ", only 5 >= nk >=3 is allowed."))}

 par(lwd=1,bty="o",col="black")
 leg <- c(leg,"Flexible calibration (RCS)","CL flexible")
 lt <- c(lt,lty.smooth,2)
 lw.d <- c(lw.d,rep(lwd.smooth,2))
 all.col <- c(all.col,col.smooth)
 marks <- c(marks,-1,-1)
}
if(!missing(m) | !missing(g) | !missing(cuts)) {
 if(!missing(m))
   q <- cut2(p, m = m, levels.mean = T, digits = 7)
 else if(!missing(g))
   q <- cut2(p, g = g, levels.mean = T, digits = 7)
 else if(!missing(cuts))
   q <- cut2(p, cuts = cuts, levels.mean = T, digits = 7)
 means <- as.single(levels(q))
 prop <- tapply(y, q, function(x)mean(x, na.rm = T))
 points(means, prop, pch = 2, cex=1)
 #18.11.02: CI triangles
 ng     <-tapply(y, q, length)
 og     <-tapply(y, q, sum)
 ob     <-og/ng
 se.ob <-sqrt(ob*(1-ob)/ng)
 g                  <- length(as.single(levels(q)))

 for (i in 1:g) lines(c(means[i], means[i]), c(prop[i],min(1,prop[i]+1.96*se.ob[i])), type="l")
 for (i in 1:g) lines(c(means[i], means[i]), c(prop[i],max(0,prop[i]-1.96*se.ob[i])), type="l")

 if(connect.group) {
   lines(means, prop)
   lt <- c(lt, 1)
   lw.d <- c(lw.d,1)
 }
```

```
    else {
      lt <- c(lt, 0)
      lw.d <- c(lw.d, 0)
    }
    leg <- c(leg, "Grouped observations")
    all.col <- c(all.col, col.smooth)
    marks <- c(marks, 2)
  }
}
lr <- stats["Model L.R."]
p.lr <- stats["P"]
D <- (lr - 1)/n
L01 <- -2 * sum(y * logit - logb(1 + exp(logit)), na.rm = TRUE)
U.chisq <- L01 - f$deviance[2]
p.U <- 1 - pchisq(U.chisq, 2)
U <- (U.chisq - 2)/n
Q <- D - U
Dxy <- stats["Dxy"]
C <- stats["C"]
R2 <- stats["R2"]
B <- sum((p - y)^2)/n
# ES 15dec08 add Brier scaled
Bmax  <- mean(y) * (1-mean(y))^2 + (1-mean(y)) * mean(y)^2
Bscaled <- 1 - B/Bmax
stats <- c(Dxy, C, R2, D, lr, p.lr, U, U.chisq, p.U, Q, B,
           f2$coef[1], f$coef[2], emax, Bscaled)
names(stats) <- c("Dxy", "C (ROC)", "R2", "D", "D:Chi-sq",
                  "D:p", "U", "U:Chi-sq", "U:p", "Q", "Brier", "Intercept",
                  "Slope", "Emax", "Brier scaled")
if(smooth=="loess")
  stats <- c(stats, c(Eavg = eavg),c(ECI = ECI))

# Cut off definition
if(!missing(cutoff)) {
  arrows(x0=cutoff,y0=.1,x1=cutoff,y1=-0.025,length=.15)
}
if(pl) {
  if(min(p)>plogis(-7) | max(p)<plogis(7)){

    lrm(y[i.2]~qlogis(p[i.2]))-> lrm.fit.1
    if(logistic.cal)  lines(p[i.2],plogis(lrm.fit.1$linear.predictors),lwd=lwd.log,lty=lty.log,col=col.log)

  }else{logit <- seq(-7, 7, length = 200)
  prob <- 1/(1 + exp( - logit))
  pred.prob <- f$coef[1] + f$coef[2] * logit
  pred.prob <- 1/(1 + exp( - pred.prob))
  if(logistic.cal) lines(prob, pred.prob, lty=lty.log,lwd=lwd.log,col=col.log)
  }
  #         pc <- rep(" ", length(lt))
  #         pc[lt==0] <- "."
  lp <- legendloc
  if (!is.logical(lp)) {
    if (!is.list(lp))
      lp <- list(x = lp[1], y = lp[2])
    legend(lp, leg, lty = lt, pch = marks, cex = cex.leg, bty = "n",lwd=lw.d,
           col=all.col,y.intersp = y.intersp)
  }
  if(!is.logical(statloc)) {
    if(dostats[1]==T){
      stats.2 <- paste('Calibration\n',
                       '...intercept: '
                       , sprintf(paste("%.",roundstats,"f",sep=""), stats["Intercept"]), " (",
                       sprintf(paste("%.",roundstats,"f",sep=""),cl.interc[1])," to ",
                       sprintf(paste("%.",roundstats,"f",sep=""),cl.interc[2]),")",'\n',
                       '...slope: '
                       , sprintf(paste("%.",roundstats,"f",sep=""), stats["Slope"]), " (",
                       sprintf(paste("%.",roundstats,"f",sep=""),cl.slope[1])," to ",
                       sprintf(paste("%.",roundstats,"f",sep=""),cl.slope[2]),")",'\n',
                       'Discrimination\n',
                       '...c-statistic: '
                       , sprintf(paste("%.",roundstats,"f",sep=""), stats["C (ROC)"]), " (",
                       sprintf(paste("%.",roundstats,"f",sep=""),cl.auc[2])," to ",
                       sprintf(paste("%.",roundstats,"f",sep=""),cl.auc[3]),")"
                       , sep = '')
      text(statloc[1], statloc[2],stats.2,pos=4,cex=cex)

    }else{
      dostats <- dostats
      leg <- format(names(stats)[dostats])           #constant length
```

```
      leg <- paste(leg, ":", format(stats[dostats], digits=roundstats), sep =
                   "")
      if(!is.list(statloc))
        statloc <- list(x = statloc[1], y = statloc[2])
      text(statloc, paste(format(names(stats[dostats])),
                          collapse = "\n"), adj = 0, cex = cex)
      text(statloc$x + (xlim[2]-xlim[1])/3 , statloc$y, paste(
        format(round(stats[dostats], digits=roundstats)), collapse =
          "\n"), adj = 1, cex = cex)
    }
   }
   if(is.character(riskdist)) {
    if(riskdist == "calibrated") {
     x <- f$coef[1] + f$coef[2] * log(p/(1 - p))
     x <- 1/(1 + exp( - x))
     x[p == 0] <- 0
     x[p == 1] <- 1
    }
    else x <- p
    bins <- seq(0, min(1,max(xlim)), length = 101)
    x <- x[x >= 0 & x <= 1]
    #08.04.01,yvon: distribution of predicted prob according to outcome
    f0      <-table(cut(x[y==0],bins))
    f1      <-table(cut(x[y==1],bins))
    j0      <-f0 > 0
    j1      <-f1 > 0
    bins0 <-(bins[-101])[j0]
    bins1 <-(bins[-101])[j1]
    f0      <-f0[j0]
    f1      <-f1[j1]
    maxf <-max(f0,f1)
    f0      <-(0.1*f0)/maxf
    f1      <-(0.1*f1)/maxf

    segments(bins1,line.bins,bins1,length.seg*f1+line.bins)
    segments(bins0,line.bins,bins0,length.seg*-f0+line.bins)
    lines(c(min(bins0,bins1)-0.01,max(bins0,bins1)+0.01),c(line.bins,line.bins))
    text(max(bins0,bins1)+dist.label,line.bins+dist.label2,d1lab,cex=cex.d01)
    text(max(bins0,bins1)+dist.label,line.bins-dist.label2,d0lab,cex=cex.d01)

   }
  }
  if(dostats==T){
   cat(paste("\n\n A ",cl.level*100,
             "% confidence interval is given for the calibration intercept, calibration slope and c-statistic. \n\n",
             sep=""))}

  stats_ci <- list("stats" = stats, "cl.interc" = cl.interc, "cl.slope" = cl.slope, "cl.auc" = cl.auc)
  return(stats_ci)
}

#function to calculate rubin's rules
#https://bookdown.org/mwheymans/bookmi/rubins-rules.html
rubin.rules <- function(means, SEs)
{
  n = length(SEs)
  rubin_mean = mean(means)
  variance_within = (sum(SEs^2))/n
  variance_between = (sum((means-rubin_mean)^2))/(n-1)
  variance_total = variance_within + variance_between + variance_between/n
  rubin_se = sqrt(variance_total)

  rubins = list("rubin_mean" = rubin_mean, "rubin_se" = rubin_se)
  return(rubins)
}

#amended pool_auc() function https://rdrr.io/cran/psfmi/src/R/pool_auc.R
#comment out code that binds result between 0 and 1 so can use to pool other performance metrics
#like citl and slope
pool_auc_2 <- function(est_auc,
                       est_se,
                       nimp = 5,
                       log_auc=TRUE)
{

  RR_se <- function(est, se, nimp){
    m <- nimp
    w_auc <-
      mean(se^2) # within variance
```

```
  b_auc <-
    var(est) # between variance
  tv_auc <-
    w_auc + (1 + (1/m)) * b_auc # total variance
  se_total <-
    sqrt(tv_auc)
  r <- (1 + 1 / m) * (b_auc / w_auc)
  v <- (m - 1) * (1 + (1/r))^2
  t <- qt(0.975, v)
  res <- c(se_total, t)
  return(res)
}

est_auc <-
  unlist(est_auc)
est_auc_se <-
  unlist(est_se)
if(length(est_auc) != nimp)
  stop("Include c-statistic value for each imputed dataset")

if(log_auc){
  est_auc_log <-
    log(est_auc/(1-est_auc))
  est_auc_se_log <-
    est_auc_se / (est_auc * (1-est_auc))

  se_total <-
    RR_se(est_auc_log, est_auc_se_log, nimp=nimp) # pooled se

  # Backtransform
  inv.auc <- exp(mean(est_auc_log)) /
    (1 + exp(mean(est_auc_log)))
  inv.auc.u <- exp(mean(est_auc_log) + (se_total[2]*se_total[1])) /
    (1 + exp(mean(est_auc_log) + (se_total[2]*se_total[1])))
  inv.auc.l <- exp(mean(est_auc_log) - (se_total[2]*se_total[1])) /
    (1 + exp(mean(est_auc_log) - (se_total[2]*se_total[1])))
  auc_res <- round(matrix(c(inv.auc.l, inv.auc, inv.auc.u),
                          1, 3, byrow = T), 4)
  dimnames(auc_res) <- list(c("C-statistic (logit)"),
                            c("95% Low", "C-statistic", "95% Up"))
} else {
  mean_auc <-
    mean(est_auc)
  se_total <-
    RR_se(est_auc, est_auc_se, nimp=nimp)
  auc_u <-
    mean(est_auc) + (se_total[2]*se_total[1])
  #if(auc_u > 1) auc_u <- 1.00
  auc_l <- mean(est_auc) - (se_total[2]*se_total[1])
  #if(auc_l < 0) auc_l <- 0.00
  auc_res <-
    round(matrix(c(auc_l, mean_auc, auc_u),
                 1, 3, byrow = T), 4)
  dimnames(auc_res) <-
    list(c("C-statistic"), c("95% Low", "C-statistic", "95% Up"))
}
return(auc_res)
}

## To get bootstrapped answers for c-statistic & c-slope (CITL 0 for internal validation)
manual_boot <- function(data,samples){
 results <- matrix(nrow = samples,ncol = 6)
 set.seed(987)
 for (i in 1:samples) {
   samp_index <- sample(1:nrow(data), nrow(data), rep=TRUE) # create a sampling index vector

   bs_samp <- data[samp_index,] # index the original dataset using the sampling vector to give the bs sample
   model <- glm(M12_PANSS_Period_Rem ~ .,family=binomial, data=bs_samp)  # Fit model to the bootstrap sample
   pr_bs <- predict(model,type="response") # predict probabilities from the bootstrap model in the bs sample
   lp_bs <- predict(model) # predict lp from the bootstrap model in the bs sample

   pr_test <- predict(model,type="response",newdata = data) # predict probabilities from the bootstrap model in the
original sample
   lp_test <- predict(model, newdata = data) # predict lp from the bootstrap model in the original sample

   # calculate the apparent performance of the bootstrap model in the bs sample
   app_cstat_model <- roc(M12_PANSS_Period_Rem~pr_bs, data=bs_samp, levels = c("No", "Yes"), direction = "<")
   results[i,1] <- app_cstat_model$auc
   app_cslope_model <- glm(M12_PANSS_Period_Rem ~ lp_bs,family=binomial(link='logit'), data=bs_samp)
```

```
    results[i,2] <- summary(app_cslope_model)$coefficients[2,1]

    # calculate the test performance of the bootstrap model in the original sample
    test_cstat_model <- roc(M12_PANSS_Period_Rem~pr_test, data=data, levels = c("No", "Yes"), direction = "<")
    results[i,3] <- test_cstat_model$auc
    test_cslope_model <- glm(M12_PANSS_Period_Rem ~ lp_test,family=binomial, data=data)
    results[i,4] <- summary(test_cslope_model)$coefficients[2,1]

    print(i)
  }
  results2 <- as.data.frame(results)
  colnames(results2) <- c("app_c_stat","app_c_slope","test_c_stat","test_c_slope")
  return(results2)
}
```

Appendix 3 Note on internal validation method for Chapter 3

The internal validation was performed by applying k fold repeated cross validation. To calculate performance metrics (principally the c-statistic) from cross-validation two methods have been described in the literature. In the first method, the performance metric is calculated separately in each left out fold, repeated across k folds, then averaged. In the second method, the probabilities are calculated in each left out fold, repeated across k folds, concatenated into a long vector and a single performance metric calculated. This debate is addressed by Zou *et al* (Zou et al., 2012) in their text (here ROC is receiver operating characteristic curve and AUC is area under the curve, ROC AUC is equivalent to the c-statistic for binary outcomes):

> "the machine learning community often uses other strategies to calculate the cross-validated AUC. For example, Bradley pointed out that some averaged AUCs from ROC curves correspond to each partition and others aggregated the outputs of all folds first, producing one ROC and calculating its AUC"

In my original analysis for Chapter 3, the second of the two cross validation methods I described above was used which resulted in the c-statistic of 0.74 (0.72, 0.76) and calibration slope of 0.84 (0.76, 0.92). As a sensitivity analysis, the internal validation was repeated using the first cross validation method described above. This resulted in an identical c-statistic of 0.74 (0.72, 0.76) but a higher calibration slope of 0.94 (0.82, 1.05). As a further sensitivity analysis, the internal validation was repeated using Harrell's bootstrap-based optimism-corrected method (Harrell et al., 1996). This again resulted in an identical c-statistic of 0.74 but a different calibration slope of 0.88. The heuristic shrinkage factor of Van Houwelingen (van Houwelingen & Le Cessie, 1990) was calculated as a final comparison which was 0.90.

It was the original cross-validation calibration slope (0.84) which was applied as a shrinkage factor before the model was tested at external validation. This resulted in an external validation calibration slope of 0.85 (0.42, 1.27). As such, an even lower value for the shrinkage factor would have been required to give an external validation calibration slope near the ideal (unity). If the higher

cross-validation calibration slope from the sensitivity analysis was used (0.94), this would have produced an even worse external validation slope further from unity.

Following discussions with my supervisors, for Chapter 4's internal validation analysis it was the first of the two cross-validation methods I described above that I used. This change was influenced by the recommendations in a paper by Forman and Scholz (Forman & Scholz, 2010) which I read after I had completed the analysis for Chapter 3 but before I had started the analysis for Chapter 4. However, this paper only comments on the effect of the cross-validation method on the c-statistic, not the estimation of the calibration slope.

Appendix 4 R code for Chaper 4

```
################################################
#Baseline comparisons code
#
#
#
################################################

library(readr)
library(pastecs)
library(ggpubr)
library(car)
library(gmodels)
library(WRS2)
library(Hmisc)
library(psych)
library(sjstats)
library(pwr)
library(ppcor)
library(coin)
library(rstatix)
library(boot)
library(effectsize)
library(compute.es)

options(max.print=1000000)

#don't use scientific notation (revert back with options(scipen=0)
options(scipen=999)
options(digits = 4)

glx_df = read_csv("glx_df.csv")
summary(as.factor(glx_df$PATIENT), maxsum = 9999)
glx_df$GENDER = as.factor(glx_df$GENDER)
glx_df$THERAPY = as.factor(glx_df$THERAPY)
glx_df$SMOKER = as.factor(glx_df$SMOKER)
glx_df$M3_Remission = as.factor(glx_df$M3_Remission)
glx_df$PATIENT = NULL
str(glx_df)
View(glx_df)

#Sign does not contain much info - just report absolute value
#p665 discovering statistics using R
#N = total sample size
rFromWilcox<-function(wilcoxModel, N){
  z<- qnorm(wilcoxModel$p.value/2)
  r<- z/ sqrt(N)
  cat(wilcoxModel$data.name, "Effect Size, r = ", abs(r))
}

################################################
#Demographic Comparisons by 3 month remission status
Rem_data = subset(glx_df, glx_df$M3_Remission == "Yes")
NRem_data = subset(glx_df, glx_df$M3_Remission == "No")

#therapy chi-squared test
by(glx_df$THERAPY, glx_df$M3_Remission, summary)
CrossTable(x=glx_df$THERAPY, y=glx_df$M3_Remission, chisq = T, expected = T, fisher = T, format = "SPSS")

#Age by Remision Mann-Whitney U Test
by(glx_df$AGE, glx_df$M3_Remission, stat.desc, basic = FALSE, norm = TRUE)
by(glx_df$AGE, glx_df$M3_Remission, summary)
by(glx_df$AGE, glx_df$M3_Remission, length)
plot(AGE ~ M3_Remission, data = glx_df)
leveneTest(AGE ~ M3_Remission, data = glx_df)
ggqqplot(Rem_data$AGE)
hist(Rem_data$AGE)
ggqqplot(NRem_data$AGE)
hist(NRem_data$AGE)
wilcox_test = wilcox.test(AGE ~ M3_Remission, data = glx_df, alternative = "two.sided", paired = FALSE)
wilcox_test
rFromWilcox(wilcox_test,N=168)
set.seed(987)
wilcox_eff = wilcox_effsize(AGE ~ M3_Remission, data = glx_df, alternative = "two.sided", paired = FALSE, ci = TRUE)
wilcox_eff$effsize
wilcox_eff$conf.low
```

```
wilcox_eff$conf.high

#GENDER chi-squared test
by(glx_df$GENDER, glx_df$M3_Remission, summary)
CrossTable(x=glx_df$GENDER, y=glx_df$M3_Remission, chisq = T, expected = T, fisher = T, format = "SPSS")

#SMOKER chi-squared test
by(glx_df$SMOKER, glx_df$M3_Remission, summary)
CrossTable(x=glx_df$SMOKER, y=glx_df$M3_Remission, chisq = T, expected = T, fisher = T, format = "SPSS")

#M0_FF_Positive t-test
by(glx_df$M0_FF_Pos, glx_df$M3_Remission, stat.desc, basic = FALSE, norm = TRUE)
by(glx_df$M0_FF_Pos, glx_df$M3_Remission, summary)
plot(M0_FF_Pos ~ M3_Remission, data = glx_df)
leveneTest(M0_FF_Pos ~ M3_Remission, data = glx_df)
ggqqplot(Rem_data$M0_FF_Pos)
hist(Rem_data$M0_FF_Pos)
ggqqplot(NRem_data$M0_FF_Pos)
hist(NRem_data$M0_FF_Pos)
t_test = t.test(M0_FF_Pos ~ M3_Remission, data = glx_df, alternative = "two.sided",paired = FALSE, var.equal = TRUE)
t_test
set.seed(987)
cohen = cohens_d(M0_FF_Pos ~ M3_Remission, data = glx_df, paired = FALSE, var.equal = TRUE, ci = TRUE)
cohen$effsize
cohen$conf.low
cohen$conf.high

#M0_FF_Neg Mann-Whitney U Test
by(glx_df$M0_FF_Neg, glx_df$M3_Remission, stat.desc, basic = FALSE, norm = TRUE)
by(glx_df$M0_FF_Neg, glx_df$M3_Remission, summary)
by(glx_df$M0_FF_Neg, glx_df$M3_Remission, length)
plot(M0_FF_Neg ~ M3_Remission, data = glx_df)
leveneTest(M0_FF_Neg ~ M3_Remission, data = glx_df)
ggqqplot(Rem_data$M0_FF_Neg)
hist(Rem_data$M0_FF_Neg)
ggqqplot(NRem_data$M0_FF_Neg)
hist(NRem_data$M0_FF_Neg)
wilcox_test = wilcox.test(M0_FF_Neg ~ M3_Remission, data = glx_df, alternative = "two.sided", paired = FALSE)
wilcox_test
rFromWilcox(wilcox_test,N=168)
set.seed(987)
wilcox_eff = wilcox_effsize(M0_FF_Neg ~ M3_Remission, data = glx_df, alternative = "two.sided", paired = FALSE, ci =
TRUE)
wilcox_eff$effsize
wilcox_eff$conf.low
wilcox_eff$conf.high

#M0_FF_Dis Mann-Whitney U Test
by(glx_df$M0_FF_Dis, glx_df$M3_Remission, stat.desc, basic = FALSE, norm = TRUE)
by(glx_df$M0_FF_Dis, glx_df$M3_Remission, summary)
by(glx_df$M0_FF_Dis, glx_df$M3_Remission, length)
plot(M0_FF_Dis ~ M3_Remission, data = glx_df)
leveneTest(M0_FF_Dis ~ M3_Remission, data = glx_df)
ggqqplot(Rem_data$M0_FF_Dis)
hist(Rem_data$M0_FF_Dis)
ggqqplot(NRem_data$M0_FF_Dis)
hist(NRem_data$M0_FF_Dis)
wilcox_test = wilcox.test(M0_FF_Dis ~ M3_Remission, data = glx_df, alternative = "two.sided", paired = FALSE)
wilcox_test
rFromWilcox(wilcox_test,N=168)
set.seed(987)
wilcox_eff = wilcox_effsize(M0_FF_Dis ~ M3_Remission, data = glx_df, alternative = "two.sided", paired = FALSE, ci =
TRUE)
wilcox_eff$effsize
wilcox_eff$conf.low
wilcox_eff$conf.high

#M0_FF_Exc Mann-Whitney U Test
by(glx_df$M0_FF_Exc, glx_df$M3_Remission, stat.desc, basic = FALSE, norm = TRUE)
by(glx_df$M0_FF_Exc, glx_df$M3_Remission, summary)
by(glx_df$M0_FF_Exc, glx_df$M3_Remission, length)
plot(M0_FF_Exc ~ M3_Remission, data = glx_df)
leveneTest(M0_FF_Exc ~ M3_Remission, data = glx_df)
ggqqplot(Rem_data$M0_FF_Exc)
hist(Rem_data$M0_FF_Exc)
ggqqplot(NRem_data$M0_FF_Exc)
hist(NRem_data$M0_FF_Exc)
wilcox_test = wilcox.test(M0_FF_Exc ~ M3_Remission, data = glx_df, alternative = "two.sided", paired = FALSE)
wilcox_test
```

```
rFromWilcox(wilcox_test,N=168)
set.seed(987)
wilcox_eff = wilcox_effsize(M0_FF_Exc ~ M3_Remission, data = glx_df, alternative = "two.sided", paired = FALSE, ci =
TRUE)
wilcox_eff$effsize
wilcox_eff$conf.low
wilcox_eff$conf.high

#M0_FF_Emo Mann-Whitney U Test
by(glx_df$M0_FF_Emo, glx_df$M3_Remission, stat.desc, basic = FALSE, norm = TRUE)
by(glx_df$M0_FF_Emo, glx_df$M3_Remission, summary)
by(glx_df$M0_FF_Emo, glx_df$M3_Remission, length)
plot(M0_FF_Emo ~ M3_Remission, data = glx_df)
leveneTest(M0_FF_Emo ~ M3_Remission, data = glx_df)
ggqqplot(Rem_data$M0_FF_Emo)
hist(Rem_data$M0_FF_Emo)
ggqqplot(NRem_data$M0_FF_Emo)
hist(NRem_data$M0_FF_Emo)
wilcox_test = wilcox.test(M0_FF_Emo ~ M3_Remission, data = glx_df, alternative = "two.sided", paired = FALSE)
wilcox_test
rFromWilcox(wilcox_test,N=168)
set.seed(987)
wilcox_eff = wilcox_effsize(M0_FF_Emo ~ M3_Remission, data = glx_df, alternative = "two.sided", paired = FALSE, ci =
TRUE)
wilcox_eff$effsize
wilcox_eff$conf.low
wilcox_eff$conf.high

##################################################

##################################################
#Demographic Comparisons by Olanzapine Vs Haloperidol
Olz_data = subset(glx_df, glx_df$THERAPY == "Olz")
Hal_data = subset(glx_df, glx_df$THERAPY == "Hal")

#remission chi-squared test
by(glx_df$M3_Remission, glx_df$THERAPY, summary)
CrossTable(x=glx_df$M3_Remission, y=glx_df$THERAPY, chisq = T, expected = T, fisher = T, format = "SPSS")

#Age by Therapy Mann-Whitney U Test
by(glx_df$AGE, glx_df$THERAPY, stat.desc, basic = FALSE, norm = TRUE)
by(glx_df$AGE, glx_df$THERAPY, summary)
by(glx_df$AGE, glx_df$THERAPY, length)
plot(AGE ~ THERAPY, data = glx_df)
leveneTest(AGE ~ THERAPY, data = glx_df)
ggqqplot(Olz_data$AGE)
hist(Olz_data$AGE)
ggqqplot(Hal_data$AGE)
hist(Hal_data$AGE)
wilcox_test = wilcox.test(AGE ~ THERAPY, data = glx_df, alternative = "two.sided", paired = FALSE)
wilcox_test
rFromWilcox(wilcox_test,N=263)
set.seed(987)
wilcox_eff = wilcox_effsize(AGE ~ THERAPY, data = glx_df, alternative = "two.sided", paired = FALSE, ci = TRUE)
wilcox_eff$effsize
wilcox_eff$conf.low
wilcox_eff$conf.high

#GENDER chi-squared test
by(glx_df$GENDER, glx_df$THERAPY, summary)
CrossTable(x=glx_df$GENDER, y=glx_df$THERAPY, chisq = T, expected = T, fisher = T, format = "SPSS")

#SMOKER chi-squared test
by(glx_df$SMOKER, glx_df$THERAPY, summary)
CrossTable(x=glx_df$SMOKER, y=glx_df$THERAPY, chisq = T, expected = T,fisher = T, format = "SPSS")

#M0_FF_Positive Mann-Whitney U Test
by(glx_df$M0_FF_Pos, glx_df$THERAPY, stat.desc, basic = FALSE, norm = TRUE)
by(glx_df$M0_FF_Pos, glx_df$THERAPY, summary)
by(glx_df$M0_FF_Pos, glx_df$THERAPY, length)
plot(M0_FF_Pos ~ THERAPY, data = glx_df)
leveneTest(M0_FF_Pos ~ THERAPY, data = glx_df)
ggqqplot(Olz_data$M0_FF_Pos)
hist(Olz_data$M0_FF_Pos)
ggqqplot(Hal_data$M0_FF_Pos)
hist(Hal_data$M0_FF_Pos)
wilcox_test = wilcox.test(M0_FF_Pos ~ THERAPY, data = glx_df, alternative = "two.sided", paired = FALSE)
wilcox_test
rFromWilcox(wilcox_test, N=262)
```

```
set.seed(987)
wilcox_eff = wilcox_effsize(M0_FF_Pos ~ THERAPY, data = glx_df, alternative = "two.sided", paired = FALSE, ci = TRUE)
wilcox_eff$effsize
wilcox_eff$conf.low
wilcox_eff$conf.high

#M0_FF_Neg welch t-test
by(glx_df$M0_FF_Neg, glx_df$THERAPY, stat.desc, basic = FALSE, norm = TRUE)
by(glx_df$M0_FF_Neg, glx_df$THERAPY, summary)
by(glx_df$M0_FF_Neg, glx_df$THERAPY, length)
plot(M0_FF_Neg ~ THERAPY, data = glx_df)
leveneTest(M0_FF_Neg ~ THERAPY, data = glx_df)
ggqqplot(Olz_data$M0_FF_Neg)
hist(Olz_data$M0_FF_Neg)
ggqqplot(Hal_data$M0_FF_Neg)
hist(Hal_data$M0_FF_Neg)
t_test = t.test(M0_FF_Neg ~ THERAPY, data = glx_df, alternative = "two.sided", paired = FALSE, var.equal = FALSE)
t_test
set.seed(987)
cohen = cohens_d(M0_FF_Neg ~ THERAPY, data = glx_df, paired = FALSE, var.equal = FALSE, ci = TRUE)
cohen$effsize
cohen$conf.low
cohen$conf.high

#M0_FF_Dis Mann-Whitney U Test
by(glx_df$M0_FF_Dis, glx_df$THERAPY, stat.desc, basic = FALSE, norm = TRUE)
by(glx_df$M0_FF_Dis, glx_df$THERAPY, summary)
by(glx_df$M0_FF_Dis, glx_df$THERAPY, length)
plot(M0_FF_Dis ~ THERAPY, data = glx_df)
leveneTest(M0_FF_Dis ~ THERAPY, data = glx_df)
ggqqplot(Olz_data$M0_FF_Dis)
hist(Olz_data$M0_FF_Dis)
ggqqplot(Hal_data$M0_FF_Dis)
hist(Hal_data$M0_FF_Dis)
wilcox_test = wilcox.test(M0_FF_Dis ~ THERAPY, data = glx_df, alternative = "two.sided", paired = FALSE)
wilcox_test
rFromWilcox(wilcox_test, N=262)
set.seed(987)
wilcox_eff = wilcox_effsize(M0_FF_Dis ~ THERAPY, data = glx_df, alternative = "two.sided", paired = FALSE, ci = TRUE)
wilcox_eff$effsize
wilcox_eff$conf.low
wilcox_eff$conf.high

#M0_FF_Exc Mann-Whitney U Test
by(glx_df$M0_FF_Exc, glx_df$THERAPY, stat.desc, basic = FALSE, norm = TRUE)
by(glx_df$M0_FF_Exc, glx_df$THERAPY, summary)
by(glx_df$M0_FF_Exc, glx_df$THERAPY, length)
plot(M0_FF_Exc ~ THERAPY, data = glx_df)
leveneTest(M0_FF_Exc ~ THERAPY, data = glx_df)
ggqqplot(Olz_data$M0_FF_Exc)
hist(Olz_data$M0_FF_Exc)
ggqqplot(Hal_data$M0_FF_Exc)
hist(Hal_data$M0_FF_Exc)
wilcox_test = wilcox.test(M0_FF_Exc ~ THERAPY, data = glx_df, alternative = "two.sided", paired = FALSE)
wilcox_test
rFromWilcox(wilcox_test, N=262)
set.seed(987)
wilcox_eff = wilcox_effsize(M0_FF_Exc ~ THERAPY, data = glx_df, alternative = "two.sided", paired = FALSE, ci = TRUE)
wilcox_eff$effsize
wilcox_eff$conf.low
wilcox_eff$conf.high

#M0_FF_Emo Mann-Whitney U Test
by(glx_df$M0_FF_Emo, glx_df$THERAPY, stat.desc, basic = FALSE, norm = TRUE)
by(glx_df$M0_FF_Emo, glx_df$THERAPY, summary)
by(glx_df$M0_FF_Emo, glx_df$THERAPY, length)
plot(M0_FF_Emo ~ THERAPY, data = glx_df)
leveneTest(M0_FF_Emo ~ THERAPY, data = glx_df)
ggqqplot(Olz_data$M0_FF_Emo)
hist(Olz_data$M0_FF_Emo)
ggqqplot(Hal_data$M0_FF_Emo)
hist(Hal_data$M0_FF_Emo)
wilcox_test = wilcox.test(M0_FF_Emo ~ THERAPY, data = glx_df, alternative = "two.sided", paired = FALSE)
wilcox_test
rFromWilcox(wilcox_test, N=262)
set.seed(987)
wilcox_eff = wilcox_effsize(M0_FF_Emo ~ THERAPY, data = glx_df, alternative = "two.sided", paired = FALSE, ci = TRUE)
wilcox_eff$effsize
wilcox_eff$conf.low
```

```
wilcox_eff$conf.high

################################################
#Main analysis code
#
#
#
################################################

#for multiple imputation
library(mice)
#for Riley sample size calculation
library(pmsampsize)
#for importing csv
library(readr)
#for standardisation, creating folds and ML
library(caret)
#parallel processing for caret
library(doParallel)
#for ROC curves
library(pROC)
#for converting probability to  LP (log-odds)
library(stats)
#for pool auc function which does rubin's rules with logit transforms for auc as per Steyerberg p150
#also editted the function to use it pool other performance metrics
library(psfmi)
#elastic net
library(glmnet)
library(Matrix)
#linear svm
library(e1071)
#naive bayes
library(naivebayes)
#random forest
library(party)
#radial SVM
library(kernlab)
#Levene's test before anova
library(car)
#qqplots
library(ggpubr)
#Games Howell Post-hoc Tests
library(rstatix)
#Welch's ANOVA effect size
library(statsExpressions)
#combine F across multiple imputations
library(miceadds)

#don't use scientific notation (revert back with options(scipen=0)
options(scipen = 999)
options(digits = 4)

#My custom functions
#############################################################################################
#change char cols to factor
csv_to_factor <- function(imported_csv)
{
  cols_char_csv = colnames(imported_csv[, sapply(imported_csv, class) == 'character'])
  for (i in seq(1:length(cols_char_csv)))
  {
    imported_csv[[cols_char_csv[i]]] = as.factor(imported_csv[[cols_char_csv[i]]])
  }
  return(imported_csv)
}

#amended pool_auc() function https://rdrr.io/cran/psfmi/src/R/pool_auc.R
#comment out code that binds result between 0 and 1 so can use to pool other performance metrics
#like citl and slope
pool_auc_2 <- function(est_auc,
                       est_se,
                       nimp = 5)
{

  RR_se <- function(est, se, nimp){
    m <- nimp
    w_auc <-
      mean(se^2) # within variance
    b_auc <-
      var(est) # between variance
```

```
  tv_auc <-
    w_auc + (1 + (1/m)) * b_auc # total variance
  se_total <-
    sqrt(tv_auc)
  r <- (1 + 1 / m) * (b_auc / w_auc)
  v <- (m - 1) * (1 + (1/r))^2
  t <- qt(0.975, v)
  res <- c(se_total, t)
  return(res)
 }

 est_auc <-
  unlist(est_auc)
 est_auc_se <-
  unlist(est_se)
 if(length(est_auc) != nimp)
  stop("Include value for each imputed dataset")

 mean_auc <-
  mean(est_auc)
 se_total <-
  RR_se(est_auc, est_auc_se, nimp=nimp)
 auc_u <-
  mean(est_auc) + (se_total[2]*se_total[1])
 #if(auc_u > 1) auc_u <- 1.00
 auc_l <- mean(est_auc) - (se_total[2]*se_total[1])
 #if(auc_l < 0) auc_l <- 0.00
 auc_res <-
  round(matrix(c(auc_l, mean_auc, auc_u),
               1, 3, byrow = T), 4)
 dimnames(auc_res) <-
  list(c("Metric"), c("95% Low", "Estimate", "95% Up"))

 return(auc_res)
}

data <- read_csv("final_logistic_df.csv" )
data <- csv_to_factor(data)

#sample size with 7 expert chosen predictors
#pmsampsize(
#  type = "b",
#  cstatistic = 0.8,
#  parameters = 7,
#  shrinkage = 0.9,
#  prevalence = 0.51
#)

#Sample size not large enough for development - put in limitation
#aiming for c-stat 0.74 like sbo model, 7 expert parameters and prevalence 0.51
#need sample size  of 385 with 197 events - epp 28.05

#https://pubmed.ncbi.nlm.nih.gov/30596876/
#PROBAST suggests an EPV of at least 10
#Events 86
#Try model with 7 clinical variables - 7 largest absolute regression coefficients from sbo study
#Compare to model with addition of MRS variable and model with addition blood variable
#MRS variable to use based on analysis - frontal GLX
#Blood variable to use based on literature - Neutrophil/Lymphocyte ratio
#https://academic.oup.com/schizophreniabulletin/advance-article/doi/10.1093/schbul/sbac089/6649676?login=false

#Data Prep and Imputation

#Top clinical variables from sbo model:
#DUP (not present), P3, PAS Highest, P2, N4, GAF Sx (CGI-S), male sex, Deprivation (not present), G6, Past Drug Use,
Insight, P1, Age, GAF Dis

#Get final dataset of 12 expert predictors, bloods and mrs and outcome for imputation including using auxillary variables
data_final <- data[,c(3:10,12,16:18,20,22:28,31,33,35,39)]

#impute using auxillary variables and outcome, 5 imputations
tempData <- mice(data_final,m=5,seed=5)

#####################################################
#Logistic regression code 1/6

#Clinical Variables only

#lists to store internal validated mean & SE aucs, CITL (intercept with LP as offset term), and calibration slopes
```

```
internal_val_aucs_mean_clin <- list()
internal_val_aucs_SE_clin <- list()
internal_val_citl_mean_clin <- list()
internal_val_citl_SE_clin <- list()
internal_val_slope_mean_clin <- list()
internal_val_slope_SE_clin <- list()

#list to store concatenated predictions for each MI to look at their distribution
internal_val_prob_clin <- list()
#list to store observations
internal_val_obs_clin <- list()
#list to store fold
internal_val_fold_clin <- list()

#Store model fit on entire dataset for each MI
finalModels_clin <- list()

#lists to store concatenated aucs, citl and slopes across all folds and multiple imputation
internal_val_aucs_clin <- list()
internal_val_citl_clin <- list()
internal_val_slope_clin <- list()

for (i in seq(1:tempData$m)){
  #Get imputed data
  lilly_imp <- complete(tempData,i)
  #just take the columns we are testing except outcome as standardising first
  #TODO change this to test new variables versus old
  #Add NLR, MLR, BGGLNCR, FCGLNCR, HCGLNCR
  lilly_imp_final <- lilly_imp[,c("CGISEV","GENDER",
                                  "PANSS_P2","PANSS_P3","PANSS_N4","PANSS_G6",
                                  "PAS_Highest")]
  #standardise the columns before building model
  preProcValues <- preProcess(lilly_imp_final, method = c("center", "scale"))
  lilly_imp_final_stand <- predict(preProcValues, lilly_imp_final)
  #Add factor outcome back in before imputed
  lilly_imp_final_stand$Non_Remission_12 <- data$Non_Remission_12
  #Remove rows with missing outcomes
  lilly_imp_final_stand_MID <- lilly_imp_final_stand[complete.cases(lilly_imp_final_stand), ]

  finalModels_clin[[i]] <- glm(Non_Remission_12 ~ ., data = lilly_imp_final_stand_MID, family = "binomial")

  set.seed(987)
  #10 fold CV repeated 50 times as per Frank Harrell "For 10-fold cv it is best to do 50 repeats"
  #https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation
  folds <-
    createMultiFolds(y = lilly_imp_final_stand_MID$Non_Remission_12,
                     #returns training data indices
                     k = 10,
                     times = 50) #TO TEST QUICKER REDUCE THIS NUMBER, e.g. to 1

  #vector to store per fold AUCs, CITL and calibration slope
  cv_aucs <- NULL
  cv_citl <- NULL
  cv_slope <- NULL

  #vector to store per fold probs, obs and folds
  cv_prob <- NULL
  cv_obs <- NULL
  cv_fold <- NULL

  for (j in seq(1:length(folds)))  {
    print(paste("Imputed dataset",i, "fold",j))

    trainCV <- lilly_imp_final_stand_MID[folds[[j]], ]
    testCV <- lilly_imp_final_stand_MID[-folds[[j]], ]

    model <- glm(Non_Remission_12 ~ ., data = trainCV, family = "binomial")

    prob_test <- predict.glm(model, testCV, type = "response")
    #store concatenated probs, obs and fold across internal validations
    cv_prob <- c(cv_prob, prob_test)
    cv_obs <- c(cv_obs, paste(testCV$Non_Remission_12))
    cv_fold <- c(cv_fold, rep(paste("fold",j),length(prob_test)))

    lp_test <- qlogis(prob_test)

    #Calculate AUC for test fold
    cv_aucs <- c(cv_aucs,
      roc(
```

```
      predictor = prob_test,
      response = testCV$Non_Remission_12,
      ci = F,
      levels = c("N", "Y"),
      direction = "<"
    )$auc
  )

  #LP as offset for citl
  cv_citl <- c(cv_citl,
           summary(glm(Non_Remission_12 ~ offset(lp_test),
                   data = testCV, family=binomial(link='logit')))$coefficients[1,1])

  #
  cv_slope <- c(cv_slope,
           summary(glm(Non_Remission_12 ~ lp_test,
                   data = testCV, family=binomial(link='logit')))$coefficients[2,1])

 }
 #store internal validated aucs for each imputed dataset
 internal_val_aucs_mean_clin[[i]] <- mean(cv_aucs)
 internal_val_aucs_SE_clin[[i]] <- sqrt(var(cv_aucs)) / sqrt(length(cv_aucs))

 #store internal validated citl for each imputed dataset
 internal_val_citl_mean_clin[[i]] <- mean(cv_citl)
 internal_val_citl_SE_clin[[i]] <- sqrt(var(cv_citl)) / sqrt(length(cv_citl))

 #store internal validated calibration slope for each imputed dataset
 internal_val_slope_mean_clin[[i]] <- mean(cv_slope)
 internal_val_slope_SE_clin[[i]] <- sqrt(var(cv_slope)) / sqrt(length(cv_slope))

 #store concatenated probs for each MI
 internal_val_prob_clin[[i]] <- cv_prob
 #store concatenated obs for each MI
 internal_val_obs_clin[[i]] <- cv_obs
 #store folds
 internal_val_fold_clin[[i]] <- cv_fold

 #Store concatenated aucs, citl and slopes across all folds and multiple imputation
 internal_val_aucs_clin[[i]] <- cv_aucs
 internal_val_citl_clin[[i]] <- cv_citl
 internal_val_slope_clin[[i]] <- cv_slope
}

pdf("supp_figure1A.pdf")
#look at distribution of probabilities for each imputed dataset
hist(internal_val_prob_clin[[1]], main = "Original Model", xlab="Predicted Probability",
    xlim=c(0,1))
dev.off()

pdf("supp_figure2A.pdf")
#look at distribution of probabilities for each imputed dataset
hist(internal_val_prob_clin[[1]], main = "MLE Logistic Regression", xlab="Predicted Probability",
    xlim=c(0,1))
dev.off()

#auc
pool_auc(
 est_auc = internal_val_aucs_mean_clin,
 est_se = internal_val_aucs_SE_clin,
 nimp = tempData$m,
 log_auc = T
)

#CITL - should be zero for internal validation
pool_auc_2(internal_val_citl_mean_clin, internal_val_citl_SE_clin,
        nimp = tempData$m)

#Calibration slope
pool_auc_2(internal_val_slope_mean_clin, internal_val_slope_SE_clin,
        nimp = tempData$m)

#final model
#Pool results to get predictor estimates based on Rubin's rule
#summary(pool(finalModels_clin), conf.int = T, exponentiate = F, conf.level = 0.95)
#View(exp(summary(pool(finalModels_clin), conf.int = T, exponentiate = F, conf.level = 0.95)[,c(2,7,8)]))

#####################################################
#Logistic regression code 2/6
```

```
#Clinical Variables + NLR

#lists to store internal validated mean & SE aucs, CITL (intercept with LP as offset term), and calibration slopes
internal_val_aucs_mean_clin_nlr <- list()
internal_val_aucs_SE_clin_nlr <- list()
internal_val_citl_mean_clin_nlr <- list()
internal_val_citl_SE_clin_nlr <- list()
internal_val_slope_mean_clin_nlr <- list()
internal_val_slope_SE_clin_nlr <- list()

#list to store concatenated predictions for each MI to look at their distribution
internal_val_prob_clin_nlr <- list()
#list to store observations
internal_val_obs_clin_nlr <- list()
#list to store fold
internal_val_fold_clin_nlr <- list()

#Store model fit on entire dataset for each MI
finalModels_clin_nlr <- list()

#lists to store concatenated aucs, citl and slopes across all folds and multiple imputation
internal_val_aucs_clin_nlr <- list()
internal_val_citl_clin_nlr <- list()
internal_val_slope_clin_nlr <- list()

for (i in seq(1:tempData$m)){
  #Get imputed data
  lilly_imp <- complete(tempData,i)
  #just take the columns we are testing except outcome as standardising first
  #TODO change this to test new variables versus old
  #Add NLR, MLR, BGGLNCR, FCGLNCR, HCGLNCR
  lilly_imp_final <- lilly_imp[,c("CGISEV","GENDER",
                                  "PANSS_P2","PANSS_P3","PANSS_N4","PANSS_G6",
                                  "PAS_Highest","NLR")]
  #standardise the columns before building model
  preProcValues <- preProcess(lilly_imp_final, method = c("center", "scale"))
  lilly_imp_final_stand <- predict(preProcValues, lilly_imp_final)
  #Add factor outcome back in before imputed
  lilly_imp_final_stand$Non_Remission_12 <- data$Non_Remission_12
  #Remove rows with missing outcomes
  lilly_imp_final_stand_MID <- lilly_imp_final_stand[complete.cases(lilly_imp_final_stand), ]

  finalModels_clin_nlr[[i]] <- glm(Non_Remission_12 ~ ., data = lilly_imp_final_stand_MID, family = "binomial")

  set.seed(987)
  #10 fold CV repeated 50 times as per Frank Harrell "For 10-fold cv it is best to do 50 repeats"
  #https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation
  folds <-
    createMultiFolds(y = lilly_imp_final_stand_MID$Non_Remission_12,
                     #returns training data indices
                     k = 10,
                     times = 50) #TO TEST QUICKER REDUCE THIS NUMBER, e.g. to 1

  #vector to store per fold AUCs, CITL and calibration slope
  cv_aucs <- NULL
  cv_citl <- NULL
  cv_slope <- NULL

  #vector to store per fold probs, obs and folds
  cv_prob <- NULL
  cv_obs <- NULL
  cv_fold <- NULL

  for (j in seq(1:length(folds)))  {
    print(paste("Imputed dataset",i, "fold",j))

    trainCV <- lilly_imp_final_stand_MID[folds[[j]], ]
    testCV <- lilly_imp_final_stand_MID[-folds[[j]], ]

    model <- glm(Non_Remission_12 ~ ., data = trainCV, family = "binomial")

    prob_test <- predict.glm(model, testCV, type = "response")
    #store concatenated probs, obs and fold across internal validations
    cv_prob <- c(cv_prob, prob_test)
    cv_obs <- c(cv_obs, paste(testCV$Non_Remission_12))
    cv_fold <- c(cv_fold, rep(paste("fold",j),length(prob_test)))

    lp_test <- qlogis(prob_test)
```

```
  #Calculate AUC for test fold
  cv_aucs <- c(cv_aucs,
           roc(
             predictor = prob_test,
             response = testCV$Non_Remission_12,
             ci = F,
             levels = c("N", "Y"),
             direction = "<"
           )$auc
  )

  #LP as offset for citl
  cv_citl <- c(cv_citl,
           summary(glm(Non_Remission_12 ~ offset(lp_test),
                    data = testCV, family=binomial(link='logit')))$coefficients[1,1])

  #
  cv_slope <- c(cv_slope,
            summary(glm(Non_Remission_12 ~ lp_test,
                     data = testCV, family=binomial(link='logit')))$coefficients[2,1])

 }
 #store internal validated aucs for each imputed dataset
 internal_val_aucs_mean_clin_nlr[[i]] <- mean(cv_aucs)
 internal_val_aucs_SE_clin_nlr[[i]] <- sqrt(var(cv_aucs)) / sqrt(length(cv_aucs))

 #store internal validated citl for each imputed dataset
 internal_val_citl_mean_clin_nlr[[i]] <- mean(cv_citl)
 internal_val_citl_SE_clin_nlr[[i]] <- sqrt(var(cv_citl)) / sqrt(length(cv_citl))

 #store internal validated calibration slope for each imputed dataset
 internal_val_slope_mean_clin_nlr[[i]] <- mean(cv_slope)
 internal_val_slope_SE_clin_nlr[[i]] <- sqrt(var(cv_slope)) / sqrt(length(cv_slope))

 #store concatenated probs for each MI
 internal_val_prob_clin_nlr[[i]] <- cv_prob
 #store concatenated obs for each MI
 internal_val_obs_clin_nlr[[i]] <- cv_obs
 #store folds
 internal_val_fold_clin_nlr[[i]] <- cv_fold

 #Store concatenated aucs, citl and slopes across all folds and multiple imputation
 internal_val_aucs_clin_nlr[[i]] <- cv_aucs
 internal_val_citl_clin_nlr[[i]] <- cv_citl
 internal_val_slope_clin_nlr[[i]] <- cv_slope
}

pdf("supp_figure1F.pdf")
#look at distribution of probabilities for each imputed dataset
hist(internal_val_prob_clin_nlr[[1]], main = "Original Model + NLR", xlab="Predicted Probability",
    xlim=c(0,1))
dev.off()

#auc
pool_auc(
 est_auc = internal_val_aucs_mean_clin_nlr,
 est_se = internal_val_aucs_SE_clin_nlr,
 nimp = tempData$m,
 log_auc = T
)

#CITL - should be zero for internal validation
pool_auc_2(internal_val_citl_mean_clin_nlr, internal_val_citl_SE_clin_nlr,
        nimp = tempData$m)

#Calibration slope
pool_auc_2(internal_val_slope_mean_clin_nlr, internal_val_slope_SE_clin_nlr,
        nimp = tempData$m)

#final model
#Pool results to get predictor estimates based on Rubin's rule
#summary(pool(finalModels_clin_nlr), conf.int = T, exponentiate = F, conf.level = 0.95)
#View(exp(summary(pool(finalModels_clin_nlr), conf.int = T, exponentiate = F, conf.level = 0.95)[,c(2,7,8)]))

#######################################################
#Logistic regression code 3/6

#Clinical Variables + MLR
```

```
#lists to store internal validated mean & SE aucs, CITL (intercept with LP as offset term), and calibration slopes
internal_val_aucs_mean_clin_mlr <- list()
internal_val_aucs_SE_clin_mlr <- list()
internal_val_citl_mean_clin_mlr <- list()
internal_val_citl_SE_clin_mlr <- list()
internal_val_slope_mean_clin_mlr <- list()
internal_val_slope_SE_clin_mlr <- list()

#list to store concatenated predictions for each MI to look at their distribution
internal_val_prob_clin_mlr <- list()
#list to store observations
internal_val_obs_clin_mlr <- list()
#list to store fold
internal_val_fold_clin_mlr <- list()

#Store model fit on entire dataset for each MI
finalModels_clin_mlr <- list()

#lists to store concatenated aucs, citl and slopes across all folds and multiple imputation
internal_val_aucs_clin_mlr <- list()
internal_val_citl_clin_mlr <- list()
internal_val_slope_clin_mlr <- list()

for (i in seq(1:tempData$m)){
 #Get imputed data
 lilly_imp <- complete(tempData,i)
 #just take the columns we are testing except outcome as standardising first
 #TODO change this to test new variables versus old
 #Add NLR, MLR, BGGLNCR, FCGLNCR, HCGLNCR
 lilly_imp_final <- lilly_imp[,c("CGISEV","GENDER",
                          "PANSS_P2","PANSS_P3","PANSS_N4","PANSS_G6",
                          "PAS_Highest","MLR")]
 #standardise the columns before building model
 preProcValues <- preProcess(lilly_imp_final, method = c("center", "scale"))
 lilly_imp_final_stand <- predict(preProcValues, lilly_imp_final)
 #Add factor outcome back in before imputed
 lilly_imp_final_stand$Non_Remission_12 <- data$Non_Remission_12
 #Remove rows with missing outcomes
 lilly_imp_final_stand_MID <- lilly_imp_final_stand[complete.cases(lilly_imp_final_stand), ]

 finalModels_clin_mlr[[i]] <- glm(Non_Remission_12 ~ ., data = lilly_imp_final_stand_MID, family = "binomial")

 set.seed(987)
 #10 fold CV repeated 50 times as per Frank Harrell "For 10-fold cv it is best to do 50 repeats"
 #https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation
 folds <-
  createMultiFolds(y = lilly_imp_final_stand_MID$Non_Remission_12,
                   #returns training data indices
                   k = 10,
                   times = 50) #TO TEST QUICKER REDUCE THIS NUMBER, e.g. to 1

 #vector to store per fold AUCs, CITL and calibration slope
 cv_aucs <- NULL
 cv_citl <- NULL
 cv_slope <- NULL

 #vector to store per fold probs, obs and folds
 cv_prob <- NULL
 cv_obs <- NULL
 cv_fold <- NULL

 for (j in seq(1:length(folds)))  {
  print(paste("Imputed dataset",i, "fold",j))

  trainCV <- lilly_imp_final_stand_MID[folds[[j]], ]
  testCV <- lilly_imp_final_stand_MID[-folds[[j]], ]

  model <- glm(Non_Remission_12 ~ ., data = trainCV, family = "binomial")

  prob_test <- predict.glm(model, testCV, type = "response")
  #store concatenated probs, obs and fold across internal validations
  cv_prob <- c(cv_prob, prob_test)
  cv_obs <- c(cv_obs, paste(testCV$Non_Remission_12))
  cv_fold <- c(cv_fold, rep(paste("fold",j),length(prob_test)))

  lp_test <- qlogis(prob_test)

  #Calculate AUC for test fold
```

```
  cv_aucs <- c(cv_aucs,
               roc(
                 predictor = prob_test,
                 response = testCV$Non_Remission_12,
                 ci = F,
                 levels = c("N", "Y"),
                 direction = "<"
               )$auc
  )

  #LP as offset for citl
  cv_citl <- c(cv_citl,
               summary(glm(Non_Remission_12 ~ offset(lp_test),
                           data = testCV, family=binomial(link='logit')))$coefficients[1,1])

  #
  cv_slope <- c(cv_slope,
                summary(glm(Non_Remission_12 ~ lp_test,
                            data = testCV, family=binomial(link='logit')))$coefficients[2,1])

 }
 #store internal validated aucs for each imputed dataset
 internal_val_aucs_mean_clin_mlr[[i]] <- mean(cv_aucs)
 internal_val_aucs_SE_clin_mlr[[i]] <- sqrt(var(cv_aucs)) / sqrt(length(cv_aucs))

 #store internal validated citl for each imputed dataset
 internal_val_citl_mean_clin_mlr[[i]] <- mean(cv_citl)
 internal_val_citl_SE_clin_mlr[[i]] <- sqrt(var(cv_citl)) / sqrt(length(cv_citl))

 #store internal validated calibration slope for each imputed dataset
 internal_val_slope_mean_clin_mlr[[i]] <- mean(cv_slope)
 internal_val_slope_SE_clin_mlr[[i]] <- sqrt(var(cv_slope)) / sqrt(length(cv_slope))

 #store concatenated probs for each MI
 internal_val_prob_clin_mlr[[i]] <- cv_prob
 #store concatenated obs for each MI
 internal_val_obs_clin_mlr[[i]] <- cv_obs
 #store folds
 internal_val_fold_clin_mlr[[i]] <- cv_fold

 #Store concatenated aucs, citl and slopes across all folds and multiple imputation
 internal_val_aucs_clin_mlr[[i]] <- cv_aucs
 internal_val_citl_clin_mlr[[i]] <- cv_citl
 internal_val_slope_clin_mlr[[i]] <- cv_slope
}

pdf("supp_figure1E.pdf")
#look at distribution of probabilities for each imputed dataset
hist(internal_val_prob_clin_mlr[[1]], main = "Original Model + MLR", xlab="Predicted Probability",
    xlim=c(0,1))
dev.off()

#auc
pool_auc(
 est_auc = internal_val_aucs_mean_clin_mlr,
 est_se = internal_val_aucs_SE_clin_mlr,
 nimp = tempData$m,
 log_auc = T
)

#CITL - should be zero for internal validation
pool_auc_2(internal_val_citl_mean_clin_mlr, internal_val_citl_SE_clin_mlr,
        nimp = tempData$m)

#Calibration slope
pool_auc_2(internal_val_slope_mean_clin_mlr, internal_val_slope_SE_clin_mlr,
        nimp = tempData$m)

#final model
#Pool results to get predictor estimates based on Rubin's rule
#summary(pool(finalModels_clin_mlr), conf.int = T, exponentiate = F, conf.level = 0.95)
#View(exp(summary(pool(finalModels_clin_mlr), conf.int = T, exponentiate = F, conf.level = 0.95)[,c(2,7,8)]))

#####################################################
#Logistic regression code 4/6

#Clinical Variables + BGGLNCR

#lists to store internal validated mean & SE aucs, CITL (intercept with LP as offset term), and calibration slopes
```

```
internal_val_aucs_mean_clin_bgglncr <- list()
internal_val_aucs_SE_clin_bgglncr <- list()
internal_val_citl_mean_clin_bgglncr <- list()
internal_val_citl_SE_clin_bgglncr <- list()
internal_val_slope_mean_clin_bgglncr <- list()
internal_val_slope_SE_clin_bgglncr <- list()

#list to store concatenated predictions for each MI to look at their distribution
internal_val_prob_clin_bgglncr <- list()
#list to store observations
internal_val_obs_clin_bgglncr <- list()
#list to store fold
internal_val_fold_clin_bgglncr <- list()

#Store model fit on entire dataset for each MI
finalModels_clin_bgglncr <- list()

#lists to store concatenated aucs, citl and slopes across all folds and multiple imputation
internal_val_aucs_clin_bgglncr <- list()
internal_val_citl_clin_bgglncr <- list()
internal_val_slope_clin_bgglncr <- list()

for (i in seq(1:tempData$m)){
  #Get imputed data
  lilly_imp <- complete(tempData,i)
  #just take the columns we are testing except outcome as standardising first
  #TODO change this to test new variables versus old
  #Add NLR, MLR, BGGLNCR, FCGLNCR, HCGLNCR
  lilly_imp_final <- lilly_imp[,c("CGISEV","GENDER",
                                  "PANSS_P2","PANSS_P3","PANSS_N4","PANSS_G6",
                                  "PAS_Highest","BGGLNCR")]
  #standardise the columns before building model
  preProcValues <- preProcess(lilly_imp_final, method = c("center", "scale"))
  lilly_imp_final_stand <- predict(preProcValues, lilly_imp_final)
  #Add factor outcome back in before imputed
  lilly_imp_final_stand$Non_Remission_12 <- data$Non_Remission_12
  #Remove rows with missing outcomes
  lilly_imp_final_stand_MID <- lilly_imp_final_stand[complete.cases(lilly_imp_final_stand), ]

  finalModels_clin_bgglncr[[i]] <- glm(Non_Remission_12 ~ ., data = lilly_imp_final_stand_MID, family = "binomial")

  set.seed(987)
  #10 fold CV repeated 50 times as per Frank Harrell "For 10-fold cv it is best to do 50 repeats"
  #https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation
  folds <-
    createMultiFolds(y = lilly_imp_final_stand_MID$Non_Remission_12,
                     #returns training data indices
                     k = 10,
                     times = 50) #TO TEST QUICKER REDUCE THIS NUMBER, e.g. to 1

  #vector to store per fold AUCs, CITL and calibration slope
  cv_aucs <- NULL
  cv_citl <- NULL
  cv_slope <- NULL

  #vector to store per fold probs, obs and folds
  cv_prob <- NULL
  cv_obs <- NULL
  cv_fold <- NULL

  for (j in seq(1:length(folds)))  {
    print(paste("Imputed dataset",i, "fold",j))

    trainCV <- lilly_imp_final_stand_MID[folds[[j]], ]
    testCV <- lilly_imp_final_stand_MID[-folds[[j]], ]

    model <- glm(Non_Remission_12 ~ ., data = trainCV, family = "binomial")

    prob_test <- predict.glm(model, testCV, type = "response")
    #store concatenated probs, obs and fold across internal validations
    cv_prob <- c(cv_prob, prob_test)
    cv_obs <- c(cv_obs, paste(testCV$Non_Remission_12))
    cv_fold <- c(cv_fold, rep(paste("fold",j),length(prob_test)))

    lp_test <- qlogis(prob_test)

    #Calculate AUC for test fold
    cv_aucs <- c(cv_aucs,
                 roc(
```

```
              predictor = prob_test,
              response = testCV$Non_Remission_12,
              ci = F,
              levels = c("N", "Y"),
              direction = "<"
             )$auc
  )

  #LP as offset for citl
  cv_citl <- c(cv_citl,
           summary(glm(Non_Remission_12 ~ offset(lp_test),
                    data = testCV, family=binomial(link='logit')))$coefficients[1,1])

  #
  cv_slope <- c(cv_slope,
            summary(glm(Non_Remission_12 ~ lp_test,
                     data = testCV, family=binomial(link='logit')))$coefficients[2,1])

 }
 #store internal validated aucs for each imputed dataset
 internal_val_aucs_mean_clin_bgglncr[[i]] <- mean(cv_aucs)
 internal_val_aucs_SE_clin_bgglncr[[i]] <- sqrt(var(cv_aucs)) / sqrt(length(cv_aucs))

 #store internal validated citl for each imputed dataset
 internal_val_citl_mean_clin_bgglncr[[i]] <- mean(cv_citl)
 internal_val_citl_SE_clin_bgglncr[[i]] <- sqrt(var(cv_citl)) / sqrt(length(cv_citl))

 #store internal validated calibration slope for each imputed dataset
 internal_val_slope_mean_clin_bgglncr[[i]] <- mean(cv_slope)
 internal_val_slope_SE_clin_bgglncr[[i]] <- sqrt(var(cv_slope)) / sqrt(length(cv_slope))

 #store concatenated probs for each MI
 internal_val_prob_clin_bgglncr[[i]] <- cv_prob
 #store concatenated obs for each MI
 internal_val_obs_clin_bgglncr[[i]] <- cv_obs
 #store folds
 internal_val_fold_clin_bgglncr[[i]] <- cv_fold

 #Store concatenated aucs, citl and slopes across all folds and multiple imputation
 internal_val_aucs_clin_bgglncr[[i]] <- cv_aucs
 internal_val_citl_clin_bgglncr[[i]] <- cv_citl
 internal_val_slope_clin_bgglncr[[i]] <- cv_slope
}

pdf("supp_figure1B.pdf")
#look at distribution of probabilities for each imputed dataset
hist(internal_val_prob_clin_bgglncr[[1]], main = "Original Model + Basal Ganglia GLX/Cr", xlab="Predicted Probability",
    xlim=c(0,1))
dev.off()

#auc
pool_auc(
 est_auc = internal_val_aucs_mean_clin_bgglncr,
 est_se = internal_val_aucs_SE_clin_bgglncr,
 nimp = tempData$m,
 log_auc = T
)

#CITL - should be zero for internal validation
pool_auc_2(internal_val_citl_mean_clin_bgglncr, internal_val_citl_SE_clin_bgglncr,
        nimp = tempData$m)

#Calibration slope
pool_auc_2(internal_val_slope_mean_clin_bgglncr, internal_val_slope_SE_clin_bgglncr,
        nimp = tempData$m)

#final model
#Pool results to get predictor estimates based on Rubin's rule
#summary(pool(finalModels_clin_bgglncr), conf.int = T, exponentiate = F, conf.level = 0.95)
#View(exp(summary(pool(finalModels_clin_bgglncr), conf.int = T, exponentiate = F, conf.level = 0.95)[,c(2,7,8)]))

#######################################################
#Logistic regression code 5/6

#Clinical Variables + FCGLNCR

#lists to store internal validated mean & SE aucs, CITL (intercept with LP as offset term), and calibration slopes
internal_val_aucs_mean_clin_fcglncr <- list()
internal_val_aucs_SE_clin_fcglncr <- list()
```

```
internal_val_citl_mean_clin_fcglncr <- list()
internal_val_citl_SE_clin_fcglncr <- list()
internal_val_slope_mean_clin_fcglncr <- list()
internal_val_slope_SE_clin_fcglncr <- list()

#list to store concatenated predictions for each MI to look at their distribution
internal_val_prob_clin_fcglncr <- list()
#list to store observations
internal_val_obs_clin_fcglncr <- list()
#list to store fold
internal_val_fold_clin_fcglncr <- list()

#Store model fit on entire dataset for each MI
finalModels_clin_fcglncr <- list()

#lists to store concatenated aucs, citl and slopes across all folds and multiple imputation
internal_val_aucs_clin_fcglncr <- list()
internal_val_citl_clin_fcglncr <- list()
internal_val_slope_clin_fcglncr <- list()

for (i in seq(1:tempData$m)){
  #Get imputed data
  lilly_imp <- complete(tempData,i)
  #just take the columns we are testing except outcome as standardising first
  #TODO change this to test new variables versus old
  #Add NLR, MLR, BGGLNCR, FCGLNCR, HCGLNCR
  lilly_imp_final <- lilly_imp[,c("CGISEV","GENDER",
                                  "PANSS_P2","PANSS_P3","PANSS_N4","PANSS_G6",
                                  "PAS_Highest","FCGLNCR")]
  #standardise the columns before building model
  preProcValues <- preProcess(lilly_imp_final, method = c("center", "scale"))
  lilly_imp_final_stand <- predict(preProcValues, lilly_imp_final)
  #Add factor outcome back in before imputed
  lilly_imp_final_stand$Non_Remission_12 <- data$Non_Remission_12
  #Remove rows with missing outcomes
  lilly_imp_final_stand_MID <- lilly_imp_final_stand[complete.cases(lilly_imp_final_stand), ]

  finalModels_clin_fcglncr[[i]] <- glm(Non_Remission_12 ~ ., data = lilly_imp_final_stand_MID, family = "binomial")

  set.seed(987)
  #10 fold CV repeated 50 times as per Frank Harrell "For 10-fold cv it is best to do 50 repeats"
  #https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation
  folds <-
    createMultiFolds(y = lilly_imp_final_stand_MID$Non_Remission_12,
                     #returns training data indices
                     k = 10,
                     times = 50) #TO TEST QUICKER REDUCE THIS NUMBER, e.g. to 1

  #vector to store per fold AUCs, CITL and calibration slope
  cv_aucs <- NULL
  cv_citl <- NULL
  cv_slope <- NULL

  #vector to store per fold probs, obs and folds
  cv_prob <- NULL
  cv_obs <- NULL
  cv_fold <- NULL

  for (j in seq(1:length(folds)))  {
    print(paste("Imputed dataset",i, "fold",j))

    trainCV <- lilly_imp_final_stand_MID[folds[[j]], ]
    testCV <- lilly_imp_final_stand_MID[-folds[[j]], ]

    model <- glm(Non_Remission_12 ~ ., data = trainCV, family = "binomial")

    prob_test <- predict.glm(model, testCV, type = "response")
    #store concatenated probs, obs and fold across internal validations
    cv_prob <- c(cv_prob, prob_test)
    cv_obs <- c(cv_obs, paste(testCV$Non_Remission_12))
    cv_fold <- c(cv_fold, rep(paste("fold",j),length(prob_test)))

    lp_test <- qlogis(prob_test)

    #Calculate AUC for test fold
    cv_aucs <- c(cv_aucs,
                 roc(
                   predictor = prob_test,
                   response = testCV$Non_Remission_12,
```

```
              ci = F,
              levels = c("N", "Y"),
              direction = "<"
             )$auc
  )

  #LP as offset for citl
  cv_citl <- c(cv_citl,
           summary(glm(Non_Remission_12 ~ offset(lp_test),
                    data = testCV, family=binomial(link='logit')))$coefficients[1,1])

  #
  cv_slope <- c(cv_slope,
            summary(glm(Non_Remission_12 ~ lp_test,
                     data = testCV, family=binomial(link='logit')))$coefficients[2,1])

 }
 #store internal validated aucs for each imputed dataset
 internal_val_aucs_mean_clin_fcglncr[[i]] <- mean(cv_aucs)
 internal_val_aucs_SE_clin_fcglncr[[i]] <- sqrt(var(cv_aucs)) / sqrt(length(cv_aucs))

 #store internal validated citl for each imputed dataset
 internal_val_citl_mean_clin_fcglncr[[i]] <- mean(cv_citl)
 internal_val_citl_SE_clin_fcglncr[[i]] <- sqrt(var(cv_citl)) / sqrt(length(cv_citl))

 #store internal validated calibration slope for each imputed dataset
 internal_val_slope_mean_clin_fcglncr[[i]] <- mean(cv_slope)
 internal_val_slope_SE_clin_fcglncr[[i]] <- sqrt(var(cv_slope)) / sqrt(length(cv_slope))

 #store concatenated probs for each MI
 internal_val_prob_clin_fcglncr[[i]] <- cv_prob
 #store concatenated obs for each MI
 internal_val_obs_clin_fcglncr[[i]] <- cv_obs
 #store folds
 internal_val_fold_clin_fcglncr[[i]] <- cv_fold

 #Store concatenated aucs, citl and slopes across all folds and multiple imputation
 internal_val_aucs_clin_fcglncr[[i]] <- cv_aucs
 internal_val_citl_clin_fcglncr[[i]] <- cv_citl
 internal_val_slope_clin_fcglncr[[i]] <- cv_slope
}

pdf("supp_figure1C.pdf")
#look at distribution of probabilities for each imputed dataset
hist(internal_val_prob_clin_fcglncr[[1]], main = "Original Model + Frontal GLX/Cr", xlab="Predicted Probability",
   xlim=c(0,1))
dev.off()

#auc
pool_auc(
 est_auc = internal_val_aucs_mean_clin_fcglncr,
 est_se = internal_val_aucs_SE_clin_fcglncr,
 nimp = tempData$m,
 log_auc = T
)

#CITL - should be zero for internal validation
pool_auc_2(internal_val_citl_mean_clin_fcglncr, internal_val_citl_SE_clin_fcglncr,
        nimp = tempData$m)

#Calibration slope
pool_auc_2(internal_val_slope_mean_clin_fcglncr, internal_val_slope_SE_clin_fcglncr,
        nimp = tempData$m)

#final model
#Pool results to get predictor estimates based on Rubin's rule
#summary(pool(finalModels_clin_fcglncr), conf.int = T, exponentiate = F, conf.level = 0.95)
#View(exp(summary(pool(finalModels_clin_fcglncr), conf.int = T, exponentiate = F, conf.level = 0.95)[,c(2,7,8)]))

#####################################################
#Logistic regression code 6/6

#Clinical Variables + HCGLNCR

#lists to store internal validated mean & SE aucs, CITL (intercept with LP as offset term), and calibration slopes
internal_val_aucs_mean_clin_hcglncr <- list()
internal_val_aucs_SE_clin_hcglncr <- list()
internal_val_citl_mean_clin_hcglncr <- list()
internal_val_citl_SE_clin_hcglncr <- list()
```

```
internal_val_slope_mean_clin_hcglncr <- list()
internal_val_slope_SE_clin_hcglncr <- list()

#list to store concatenated predictions for each MI to look at their distribution
internal_val_prob_clin_hcglncr <- list()
#list to store observations
internal_val_obs_clin_hcglncr <- list()
#list to store fold
internal_val_fold_clin_hcglncr <- list()

#Store model fit on entire dataset for each MI
finalModels_clin_hcglncr <- list()

#lists to store concatenated aucs, citl and slopes across all folds and multiple imputation
internal_val_aucs_clin_hcglncr <- list()
internal_val_citl_clin_hcglncr <- list()
internal_val_slope_clin_hcglncr <- list()

for (i in seq(1:tempData$m)){
 #Get imputed data
 lilly_imp <- complete(tempData,i)
 #just take the columns we are testing except outcome as standardising first
 #TODO change this to test new variables versus old
 #Add NLR, MLR, BGGLNCR, FCGLNCR, HCGLNCR
 lilly_imp_final <- lilly_imp[,c("CGISEV","GENDER",
                          "PANSS_P2","PANSS_P3","PANSS_N4","PANSS_G6",
                          "PAS_Highest","HCGLNCR")]
 #standardise the columns before building model
 preProcValues <- preProcess(lilly_imp_final, method = c("center", "scale"))
 lilly_imp_final_stand <- predict(preProcValues, lilly_imp_final)
 #Add factor outcome back in before imputed
 lilly_imp_final_stand$Non_Remission_12 <- data$Non_Remission_12
 #Remove rows with missing outcomes
 lilly_imp_final_stand_MID <- lilly_imp_final_stand[complete.cases(lilly_imp_final_stand), ]

 finalModels_clin_hcglncr[[i]] <- glm(Non_Remission_12 ~ ., data = lilly_imp_final_stand_MID, family = "binomial")

 set.seed(987)
 #10 fold CV repeated 50 times as per Frank Harrell "For 10-fold cv it is best to do 50 repeats"
 #https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation
 folds <-
  createMultiFolds(y = lilly_imp_final_stand_MID$Non_Remission_12,
                   #returns training data indices
                   k = 10,
                   times = 50) #TO TEST QUICKER REDUCE THIS NUMBER, e.g. to 1

 #vector to store per fold AUCs, CITL and calibration slope
 cv_aucs <- NULL
 cv_citl <- NULL
 cv_slope <- NULL

 #vector to store per fold probs, obs and folds
 cv_prob <- NULL
 cv_obs <- NULL
 cv_fold <- NULL

 for (j in seq(1:length(folds)))  {
  print(paste("Imputed dataset",i, "fold",j))

  trainCV <- lilly_imp_final_stand_MID[folds[[j]], ]
  testCV <- lilly_imp_final_stand_MID[-folds[[j]], ]

  model <- glm(Non_Remission_12 ~ ., data = trainCV, family = "binomial")

  prob_test <- predict.glm(model, testCV, type = "response")
  #store concatenated probs, obs and fold across internal validations
  cv_prob <- c(cv_prob, prob_test)
  cv_obs <- c(cv_obs, paste(testCV$Non_Remission_12))
  cv_fold <- c(cv_fold, rep(paste("fold",j),length(prob_test)))

  lp_test <- qlogis(prob_test)

  #Calculate AUC for test fold
  cv_aucs <- c(cv_aucs,
               roc(
                 predictor = prob_test,
                 response = testCV$Non_Remission_12,
                 ci = F,
                 levels = c("N", "Y"),
```

```
            direction = "<"
          )$auc
  )

  #LP as offset for citl
  cv_citl <- c(cv_citl,
          summary(glm(Non_Remission_12 ~ offset(lp_test),
                  data = testCV, family=binomial(link='logit')))$coefficients[1,1])

  #
  cv_slope <- c(cv_slope,
           summary(glm(Non_Remission_12 ~ lp_test,
                   data = testCV, family=binomial(link='logit')))$coefficients[2,1])

 }
 #store internal validated aucs for each imputed dataset
 internal_val_aucs_mean_clin_hcglncr[[i]] <- mean(cv_aucs)
 internal_val_aucs_SE_clin_hcglncr[[i]] <- sqrt(var(cv_aucs)) / sqrt(length(cv_aucs))

 #store internal validated citl for each imputed dataset
 internal_val_citl_mean_clin_hcglncr[[i]] <- mean(cv_citl)
 internal_val_citl_SE_clin_hcglncr[[i]] <- sqrt(var(cv_citl)) / sqrt(length(cv_citl))

 #store internal validated calibration slope for each imputed dataset
 internal_val_slope_mean_clin_hcglncr[[i]] <- mean(cv_slope)
 internal_val_slope_SE_clin_hcglncr[[i]] <- sqrt(var(cv_slope)) / sqrt(length(cv_slope))

 #store concatenated probs for each MI
 internal_val_prob_clin_hcglncr[[i]] <- cv_prob
 #store concatenated obs for each MI
 internal_val_obs_clin_hcglncr[[i]] <- cv_obs
 #store folds
 internal_val_fold_clin_hcglncr[[i]] <- cv_fold

 #Store concatenated aucs, citl and slopes across all folds and multiple imputation
 internal_val_aucs_clin_hcglncr[[i]] <- cv_aucs
 internal_val_citl_clin_hcglncr[[i]] <- cv_citl
 internal_val_slope_clin_hcglncr[[i]] <- cv_slope
}

pdf("supp_figure1D.pdf")
#look at distribution of probabilities for each imputed dataset
hist(internal_val_prob_clin_hcglncr[[1]], main = "Original Model + Hippocampus GLX/Cr", xlab="Predicted Probability",
    xlim=c(0,1))
dev.off()

#auc
pool_auc(
 est_auc = internal_val_aucs_mean_clin_hcglncr,
 est_se = internal_val_aucs_SE_clin_hcglncr,
 nimp = tempData$m,
 log_auc = T
)

#CITL - should be zero for internal validation
pool_auc_2(internal_val_citl_mean_clin_hcglncr, internal_val_citl_SE_clin_hcglncr,
        nimp = tempData$m)

#Calibration slope
pool_auc_2(internal_val_slope_mean_clin_hcglncr, internal_val_slope_SE_clin_hcglncr,
        nimp = tempData$m)

#final model
#Pool results to get predictor estimates based on Rubin's rule
#summary(pool(finalModels_clin_hcglncr), conf.int = T, exponentiate = F, conf.level = 0.95)
#View(exp(summary(pool(finalModels_clin_hcglncr), conf.int = T, exponentiate = F, conf.level = 0.95)[,c(2,7,8)]))

#####################################################
#Multiple imputations ANOVA combine F statistics using miceadds::micombine.F
#Dataset 1 to 5
logistic_df_mi1 <- data.frame(Group =
rep(c("clin","clin_nlr","clin_mlr","clin_bgglncr","clin_fcglncr","clin_hcglncr"),each=500),
                       AUCs = c(internal_val_aucs_clin[[1]], internal_val_aucs_clin_nlr[[1]], internal_val_aucs_clin_mlr[[1]],
                              internal_val_aucs_clin_bgglncr[[1]], internal_val_aucs_clin_fcglncr[[1]],
internal_val_aucs_clin_hcglncr[[1]]),
                       CITLs = c(internal_val_citl_clin[[1]], internal_val_citl_clin_nlr[[1]], internal_val_citl_clin_mlr[[1]],
                                internal_val_citl_clin_bgglncr[[1]], internal_val_citl_clin_fcglncr[[1]],
internal_val_citl_clin_hcglncr[[1]]),
```

```
                    SLOPEs = c(internal_val_slope_clin[[1]], internal_val_slope_clin_nlr[[1]],
internal_val_slope_clin_mlr[[1]],
                              internal_val_slope_clin_bgglncr[[1]], internal_val_slope_clin_fcglncr[[1]],
internal_val_slope_clin_hcglncr[[1]]))
logistic_df_mi1$Group <- as.factor(logistic_df_mi1$Group)

logistic_df_mi2 <- data.frame(Group =
rep(c("clin","clin_nlr","clin_mlr","clin_bgglncr","clin_fcglncr","clin_hcglncr"),each=500),
                    AUCs = c(internal_val_aucs_clin[[2]], internal_val_aucs_clin_nlr[[2]], internal_val_aucs_clin_mlr[[2]],
                          internal_val_aucs_clin_bgglncr[[2]], internal_val_aucs_clin_fcglncr[[2]],
internal_val_aucs_clin_hcglncr[[2]]),
                    CITLs = c(internal_val_citl_clin[[2]], internal_val_citl_clin_nlr[[2]], internal_val_citl_clin_mlr[[2]],
                           internal_val_citl_clin_bgglncr[[2]], internal_val_citl_clin_fcglncr[[2]],
internal_val_citl_clin_hcglncr[[2]]),
                    SLOPEs = c(internal_val_slope_clin[[2]], internal_val_slope_clin_nlr[[2]],
internal_val_slope_clin_mlr[[2]],
                              internal_val_slope_clin_bgglncr[[2]], internal_val_slope_clin_fcglncr[[2]],
internal_val_slope_clin_hcglncr[[2]]))
logistic_df_mi2$Group <- as.factor(logistic_df_mi2$Group)

logistic_df_mi3 <- data.frame(Group =
rep(c("clin","clin_nlr","clin_mlr","clin_bgglncr","clin_fcglncr","clin_hcglncr"),each=500),
                    AUCs = c(internal_val_aucs_clin[[3]], internal_val_aucs_clin_nlr[[3]], internal_val_aucs_clin_mlr[[3]],
                          internal_val_aucs_clin_bgglncr[[3]], internal_val_aucs_clin_fcglncr[[3]],
internal_val_aucs_clin_hcglncr[[3]]),
                    CITLs = c(internal_val_citl_clin[[3]], internal_val_citl_clin_nlr[[3]], internal_val_citl_clin_mlr[[3]],
                           internal_val_citl_clin_bgglncr[[3]], internal_val_citl_clin_fcglncr[[3]],
internal_val_citl_clin_hcglncr[[3]]),
                    SLOPEs = c(internal_val_slope_clin[[3]], internal_val_slope_clin_nlr[[3]],
internal_val_slope_clin_mlr[[3]],
                              internal_val_slope_clin_bgglncr[[3]], internal_val_slope_clin_fcglncr[[3]],
internal_val_slope_clin_hcglncr[[3]]))
logistic_df_mi3$Group <- as.factor(logistic_df_mi3$Group)


logistic_df_mi4 <- data.frame(Group =
rep(c("clin","clin_nlr","clin_mlr","clin_bgglncr","clin_fcglncr","clin_hcglncr"),each=500),
                    AUCs = c(internal_val_aucs_clin[[4]], internal_val_aucs_clin_nlr[[4]], internal_val_aucs_clin_mlr[[4]],
                          internal_val_aucs_clin_bgglncr[[4]], internal_val_aucs_clin_fcglncr[[4]],
internal_val_aucs_clin_hcglncr[[4]]),
                    CITLs = c(internal_val_citl_clin[[4]], internal_val_citl_clin_nlr[[4]], internal_val_citl_clin_mlr[[4]],
                           internal_val_citl_clin_bgglncr[[4]], internal_val_citl_clin_fcglncr[[4]],
internal_val_citl_clin_hcglncr[[4]]),
                    SLOPEs = c(internal_val_slope_clin[[4]], internal_val_slope_clin_nlr[[4]],
internal_val_slope_clin_mlr[[4]],
                              internal_val_slope_clin_bgglncr[[4]], internal_val_slope_clin_fcglncr[[4]],
internal_val_slope_clin_hcglncr[[4]]))
logistic_df_mi4$Group <- as.factor(logistic_df_mi4$Group)


logistic_df_mi5 <- data.frame(Group =
rep(c("clin","clin_nlr","clin_mlr","clin_bgglncr","clin_fcglncr","clin_hcglncr"),each=500),
                    AUCs = c(internal_val_aucs_clin[[5]], internal_val_aucs_clin_nlr[[5]], internal_val_aucs_clin_mlr[[5]],
                          internal_val_aucs_clin_bgglncr[[5]], internal_val_aucs_clin_fcglncr[[5]],
internal_val_aucs_clin_hcglncr[[5]]),
                    CITLs = c(internal_val_citl_clin[[5]], internal_val_citl_clin_nlr[[5]], internal_val_citl_clin_mlr[[5]],
                           internal_val_citl_clin_bgglncr[[5]], internal_val_citl_clin_fcglncr[[5]],
internal_val_citl_clin_hcglncr[[5]]),
                    SLOPEs = c(internal_val_slope_clin[[5]], internal_val_slope_clin_nlr[[5]],
internal_val_slope_clin_mlr[[5]],
                              internal_val_slope_clin_bgglncr[[5]], internal_val_slope_clin_fcglncr[[5]],
internal_val_slope_clin_hcglncr[[5]]))
logistic_df_mi5$Group <- as.factor(logistic_df_mi5$Group)


#AUCs
leveneTest(AUCs~Group, data = logistic_df_mi1) #homogeneous
ggqqplot(logistic_df_mi1$AUCs)

#ANOVA assume equal variances Imputation 1
df1 <- logistic_df_mi1
#Use logit transformation for AUC as per Steyerberg
df1$AUCs <- qlogis(df1$AUCs)
oneway.test(AUCs~Group, data = df1, var.equal = TRUE)
#F = 1.4, num df = 5, denom df = 2994, p-value = 0.2

#ANOVA assume equal variances Imputation 2
df2 <- logistic_df_mi2
#Use logit transformation for AUC as per Steyerberg
```

```
df2$AUCs <- qlogis(df2$AUCs)
oneway.test(AUCs~Group, data = df2, var.equal = TRUE)
#F = 1.5, num df = 5, denom df = 2994, p-value = 0.2

#ANOVA assume equal variances Imputation 3
df3 <- logistic_df_mi3
#Use logit transformation for AUC as per Steyerberg
df3$AUCs <- qlogis(df3$AUCs)
oneway.test(AUCs~Group, data = df3, var.equal = TRUE)
#F = 0.31, num df = 5, denom df = 2994, p-value = 0.9

#ANOVA assume equal variances Imputation 4
df4 <- logistic_df_mi4
#Use logit transformation for AUC as per Steyerberg
df4$AUCs <- qlogis(df4$AUCs)
oneway.test(AUCs~Group, data = df4, var.equal = TRUE)
#F = 1.1, num df = 5, denom df = 2994, p-value = 0.4

#ANOVA assume equal variances Imputation 5
df5 <- logistic_df_mi5
#Use logit transformation for AUC as per Steyerberg
df5$AUCs <- qlogis(df5$AUCs)
oneway.test(AUCs~Group, data = df5, var.equal = TRUE)
#F = 0.63, num df = 5, denom df = 2994, p-value = 0.7

#Combine F across multiple imputations
micombine.F(Fvalues = c(1.4,1.5,0.31,1.1,0.63),
        df1 = 5)
#AUCs
#F(5, 14.6)=0.184     p=0.9640

#CITLs
leveneTest(CITLs~Group, data = logistic_df_mi1) #homogeneous
ggqqplot(logistic_df_mi1$CITLs)

#ANOVA assume equal variances Imputation 1
oneway.test(CITLs~Group, data = logistic_df_mi1, var.equal = TRUE)
#F = 0.02, num df = 5, denom df = 2994, p-value = 1

#ANOVA assume equal variances Imputation 2
oneway.test(CITLs~Group, data = logistic_df_mi2, var.equal = TRUE)
#F = 0.028, num df = 5, denom df = 2994, p-value = 1

#ANOVA assume equal variances Imputation 3
oneway.test(CITLs~Group, data = logistic_df_mi3, var.equal = TRUE)
#F = 0.025, num df = 5, denom df = 2994, p-value = 1

#ANOVA assume equal variances Imputation 4
oneway.test(CITLs~Group, data = logistic_df_mi4, var.equal = TRUE)
#F = 0.035, num df = 5, denom df = 2994, p-value = 0.9

#ANOVA assume equal variances Imputation 5
oneway.test(CITLs~Group, data = logistic_df_mi5, var.equal = TRUE)
#F = 0.017, num df = 5, denom df = 2994, p-value = 1

#Combine F across multiple imputations
micombine.F(Fvalues = c(0.02, 0.028, 0.025, 0.035, 0.017),
        df1 = 5)
#CITLs
#F(5, 176512.09)=0.021     p=0.99983

#SLOPEs
leveneTest(SLOPEs~Group, data = logistic_df_mi1) #homogeneous
ggqqplot(logistic_df_mi1$SLOPEs)

#ANOVA assume equal variances Imputation 1
oneway.test(SLOPEs~Group, data = logistic_df_mi1, var.equal = TRUE)
#F = 0.39, num df = 5, denom df = 2994, p-value = 0.9

#ANOVA assume equal variances Imputation 2
oneway.test(SLOPEs~Group, data = logistic_df_mi2, var.equal = TRUE)
#F = 0.37, num df = 5, denom df = 2994, p-value = 0.9

#ANOVA assume equal variances Imputation 3
oneway.test(SLOPEs~Group, data = logistic_df_mi3, var.equal = TRUE)
#F = 0.3, num df = 5, denom df = 2994, p-value = 0.9

#ANOVA assume equal variances Imputation 4
oneway.test(SLOPEs~Group, data = logistic_df_mi4, var.equal = TRUE)
```

```
#F = 0.8, num df = 5, denom df = 2994, p-value = 0.5

#ANOVA assume equal variances Imputation 5
oneway.test(SLOPEs~Group, data = logistic_df_mi5, var.equal = TRUE)
#F = 0.28, num df = 5, denom df = 2994, p-value = 0.9

#Combine F across multiple imputations
micombine.F(Fvalues = c(0.39, 0.37, 0.3, 0.8, 0.28),
         df1 = 5)
#Slopes
#F(5, 114.8)=0.206     p=0.95941

#####################################################
#Machine Learning code
#Use 7 clinical variables and compare to performance of ml methods with logistic regression

#enable multicore which roughly halfs time for caret analysis runs
cl <- makeCluster(detectCores(), type='PSOCK')
registerDoParallel(cl)

#train control object for caret train method. Controls how we do parameter tuning
# 10-fold CV repeated 5 times
control <- trainControl(method="repeatedcv", number=10, repeats=5, classProbs=TRUE,
                  summaryFunction=twoClassSummary, selectionFunction = "best")

#####################################################
#ML 1/5 - Elastic net


#lists to store internal validated mean & SE aucs, CITL (intercept with LP as offset term), and calibration slopes
internal_val_aucs_ml_mean_en <- list()
internal_val_aucs_ml_SE_en <- list()
internal_val_citl_ml_mean_en <- list()
internal_val_citl_ml_SE_en <- list()
internal_val_slope_ml_mean_en <- list()
internal_val_slope_ml_SE_en <- list()

#list to store concatenated predictions for each MI to look at their distribution
internal_val_prob_ml_en <- list()
#list to store observations
internal_val_obs_ml_en <- list()
#list to store fold
internal_val_fold_ml_en <- list()

#Store model fit on entire dataset for each MI
finalModelsMl_en <- list()

internal_val_aucs_ml_en <- NULL
internal_val_citl_ml_en <- NULL
internal_val_slope_ml_en <- NULL

for (i in seq(1:tempData$m)){
  #Get imputed data
  lilly_ml_imp <- complete(tempData,i)
  #just take the columns we are testing except outcome as standardising first
  #Use 7 expert chosen variables
  lilly_imp_ml_final <- lilly_ml_imp[,c("CGISEV","GENDER",
                           "PANSS_P2","PANSS_P3","PANSS_N4","PANSS_G6",
                           "PAS_Highest")]
  #standardise the columns before building model
  preProcMlValues <- preProcess(lilly_imp_ml_final, method = c("center", "scale"))
  lilly_imp_ml_final_stand <- predict(preProcMlValues, lilly_imp_ml_final)
  #Add factor outcome back in before imputed
  lilly_imp_ml_final_stand$Non_Remission_12 <- data$Non_Remission_12
  #Remove rows with missing outcomes
  lilly_imp_ml_final_stand_MID <- lilly_imp_ml_final_stand[complete.cases(lilly_imp_ml_final_stand), ]

  set.seed(987)

  #caret train call to get model fitted on entire MI dataset
  #ML methods to try
  #logistic regression: glm (to check same results as above)
  #logistic regression:glmnet
  #naive bayes: naive_bayes
  #random forest: cforest
  #linear SVM: svmLinear2
  #radial SVM: svmRadial

  finalModelsMl_en[[i]] <- train(Non_Remission_12 ~ ., #Outcome against all predictors
```

```
                  data = lilly_imp_ml_final_stand_MID, #data
                  method ="glmnet", #TODO try different ML methods here
                  metric="ROC", #performance metric - ROC for classification problems
                  tuneLength = 5, #size of default tune grid #caret::getModelInfo("glmnet")
                  trControl = control) #Our tuning method rules

#10 fold CV repeated 50 times as per Frank Harrell "For 10-fold cv it is best to do 50 repeats"
#https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation
folds <-
  createMultiFolds(y = lilly_imp_ml_final_stand_MID$Non_Remission_12,
                   #returns training data indices
                   k = 10,
                   times = 50) #TO TEST QUICKER REDUCE THIS NUMBER, e.g. to 1

#vector to store per fold AUCs, CITL and calibration slope
cv_aucs <- NULL
cv_citl <- NULL
cv_slope <- NULL

#vector to store per fold probs, obs and folds
cv_prob <- NULL
cv_obs <- NULL
cv_fold <- NULL

for (j in seq(1:length(folds)))  {
  print(paste("Imputed dataset",i, "fold",j))

  trainCV <- lilly_imp_ml_final_stand_MID[folds[[j]], ]
  testCV <- lilly_imp_ml_final_stand_MID[-folds[[j]], ]

  #Caret train object
  #ML methods to try
  #logistic regression: glm (to check same results as above)
  #logistic regression:glmnet
  #naive bayes: naive_bayes
  #random forest: cforest
  #linear SVM: svmLinear2
  #radial SVM: svmRadial

  model <- train(Non_Remission_12 ~ ., #Outcome against all predictors
                 data = trainCV, #data
                 method ="glmnet", #TODO try different ML methods here
                 metric="ROC", #performance metric - ROC for classification problems
                 tuneLength = 5, #size of default tune grid #caret::getModelInfo("glmnet")
                 trControl = control) #Our tuning method rules

  #Caret train predict - get probabilities
  #for caret predict.train object - have to specify level to get raw probabilities
  prob_test <- predict.train(model, testCV, type = "prob")$Y

  #store concatenated probs, obs and fold across internal validations
  cv_prob <- c(cv_prob, prob_test)
  cv_obs <- c(cv_obs, paste(testCV$Non_Remission_12))
  cv_fold <- c(cv_fold, rep(paste("fold",j),length(prob_test)))

  lp_test <- qlogis(prob_test)

  #Calculate AUC for test fold
  cv_aucs <- c(cv_aucs,
               roc(
                 predictor = prob_test,
                 response = testCV$Non_Remission_12,
                 ci = F,
                 levels = c("N", "Y"),
                 direction = "<"
               )$auc
  )

  #LP as offset for citl
  cv_citl <- c(cv_citl,
               summary(glm(Non_Remission_12 ~ offset(lp_test),
                           data = testCV, family=binomial(link='logit')))$coefficients[1,1])

  #
  cv_slope <- c(cv_slope,
                summary(glm(Non_Remission_12 ~ lp_test,
                            data = testCV, family=binomial(link='logit')))$coefficients[2,1])

}
```

```
  #store internal validated aucs for each imputed dataset
  internal_val_aucs_ml_mean_en[[i]] <- mean(cv_aucs)
  internal_val_aucs_ml_SE_en[[i]] <- sqrt(var(cv_aucs)) / sqrt(length(cv_aucs))

  #store internal validated citl for each imputed dataset
  internal_val_citl_ml_mean_en[[i]] <- mean(cv_citl)
  internal_val_citl_ml_SE_en[[i]] <- sqrt(var(cv_citl)) / sqrt(length(cv_citl))

  #store internal validated calibration slope for each imputed dataset
  internal_val_slope_ml_mean_en[[i]] <- mean(cv_slope)
  internal_val_slope_ml_SE_en[[i]] <- sqrt(var(cv_slope)) / sqrt(length(cv_slope))


  #store concatenated probs for each MI
  internal_val_prob_ml_en[[i]] <- cv_prob
  #store concatenated obs for each MI
  internal_val_obs_ml_en[[i]] <- cv_obs
  #store folds
  internal_val_fold_ml_en[[i]] <- cv_fold

  #Store concatenated aucs, citl and slopes across all folds and multiple imputation
  internal_val_aucs_ml_en[[i]] <- cv_aucs
  internal_val_citl_ml_en[[i]] <- cv_citl
  internal_val_slope_ml_en[[i]] <- cv_slope
}

pdf("supp_figure2B.pdf")
#look at distribution of probabilities for each imputed dataset
hist(internal_val_prob_ml_en[[1]], main = "Elastic Net Logistic Regression", xlab="Predicted Probability",
    xlim=c(0,1))
dev.off()

#AUC
pool_auc(
  est_auc = internal_val_aucs_ml_mean_en,
  est_se = internal_val_aucs_ml_SE_en,
  nimp = tempData$m,
  log_auc = T
)

#CITL - should be zero for internal validation
pool_auc_2(internal_val_citl_ml_mean_en, internal_val_citl_ml_SE_en,
        nimp = tempData$m)

#Calibration slope
pool_auc_2(internal_val_slope_ml_mean_en, internal_val_slope_ml_SE_en,
        nimp = tempData$m)

#look at final model training
finalModelsMl_en[[1]]
#glmnet training grid alpha 0.1,0.325,0.55,0.775,1 by lambda 0.0007209, 0.0033461, 0.0155311, 0.0720888


#########################################################
#ML 2/5 - Naive Bayes

#lists to store internal validated mean & SE aucs, CITL (intercept with LP as offset term), and calibration slopes
internal_val_aucs_ml_mean_nb <- list()
internal_val_aucs_ml_SE_nb <- list()
internal_val_citl_ml_mean_nb <- list()
internal_val_citl_ml_SE_nb <- list()
internal_val_slope_ml_mean_nb <- list()
internal_val_slope_ml_SE_nb <- list()

#list to store concatenated predictions for each MI to look at their distribution
internal_val_prob_ml_nb <- list()
#list to store observations
internal_val_obs_ml_nb <- list()
#list to store fold
internal_val_fold_ml_nb <- list()

#Store model fit on entire dataset for each MI
finalModelsMl_nb <- list()

internal_val_aucs_ml_nb <- NULL
internal_val_citl_ml_nb <- NULL
internal_val_slope_ml_nb <- NULL
```

```
for (i in seq(1:tempData$m)){
  #Get imputed data
  lilly_ml_imp <- complete(tempData,i)
  #just take the columns we are testing except outcome as standardising first
  #Use 7 expert chosen variables
  lilly_imp_ml_final <- lilly_ml_imp[,c("CGISEV","GENDER",
                                        "PANSS_P2","PANSS_P3","PANSS_N4","PANSS_G6",
                                        "PAS_Highest")]
  #standardise the columns before building model
  preProcMlValues <- preProcess(lilly_imp_ml_final, method = c("center", "scale"))
  lilly_imp_ml_final_stand <- predict(preProcMlValues, lilly_imp_ml_final)
  #Add factor outcome back in before imputed
  lilly_imp_ml_final_stand$Non_Remission_12 <- data$Non_Remission_12
  #Remove rows with missing outcomes
  lilly_imp_ml_final_stand_MID <- lilly_imp_ml_final_stand[complete.cases(lilly_imp_ml_final_stand), ]

  set.seed(987)

  #caret train call to get model fitted on entire MI dataset
  #ML methods to try
  #logistic regression: glm (to check same results as above)
  #logistic regression:glmnet
  #naive bayes: naive_bayes
  #random forest: cforest
  #linear SVM: svmLinear2
  #radial SVM: svmRadial

  finalModelsMl_nb[[i]] <- train(Non_Remission_12 ~ ., #Outcome against all predictors
                                 data = lilly_imp_ml_final_stand_MID, #data
                                 method ="naive_bayes", #TODO try different ML methods here
                                 metric="ROC", #performance metric - ROC for classification problems
                                 tuneLength = 5, #size of default tune grid #caret::getModelInfo("glmnet")
                                 trControl = control) #Our tuning method rules

  #10 fold CV repeated 50 times as per Frank Harrell "For 10-fold cv it is best to do 50 repeats"
  #https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation
  folds <-
    createMultiFolds(y = lilly_imp_ml_final_stand_MID$Non_Remission_12,
                     #returns training data indices
                     k = 10,
                     times = 50) #TO TEST QUICKER REDUCE THIS NUMBER, e.g. to 1

  #vector to store per fold AUCs, CITL and calibration slope
  cv_aucs <- NULL
  cv_citl <- NULL
  cv_slope <- NULL

  #vector to store per fold probs, obs and folds
  cv_prob <- NULL
  cv_obs <- NULL
  cv_fold <- NULL

  for (j in seq(1:length(folds)))  {
    print(paste("Imputed dataset",i, "fold",j))

    trainCV <- lilly_imp_ml_final_stand_MID[folds[[j]], ]
    testCV <- lilly_imp_ml_final_stand_MID[-folds[[j]], ]

    #Caret train object
    #ML methods to try
    #logistic regression: glm (to check same results as above)
    #logistic regression:glmnet
    #naive bayes: naive_bayes
    #random forest: cforest
    #linear SVM: svmLinear2
    #radial SVM: svmRadial

    model <- train(Non_Remission_12 ~ ., #Outcome against all predictors
                   data = trainCV, #data
                   method ="naive_bayes", #TODO try different ML methods here
                   metric="ROC", #performance metric - ROC for classification problems
                   tuneLength = 5, #size of default tune grid #caret::getModelInfo("glmnet")
                   trControl = control) #Our tuning method rules

    #Caret train predict - get probabilities
    #for caret predict.train object - have to specify level to get raw probabilities
    prob_test <- predict.train(model, testCV, type = "prob")$Y

    #store concatenated probs, obs and fold across internal validations
```

```
    cv_prob <- c(cv_prob, prob_test)
    cv_obs <- c(cv_obs, paste(testCV$Non_Remission_12))
    cv_fold <- c(cv_fold, rep(paste("fold",j),length(prob_test)))

    lp_test <- qlogis(prob_test)

    #Calculate AUC for test fold
    cv_aucs <- c(cv_aucs,
                 roc(
                   predictor = prob_test,
                   response = testCV$Non_Remission_12,
                   ci = F,
                   levels = c("N", "Y"),
                   direction = "<"
                 )$auc
    )

    #LP as offset for citl
    cv_citl <- c(cv_citl,
                 summary(glm(Non_Remission_12 ~ offset(lp_test),
                             data = testCV, family=binomial(link='logit')))$coefficients[1,1])

    #
    cv_slope <- c(cv_slope,
                  summary(glm(Non_Remission_12 ~ lp_test,
                              data = testCV, family=binomial(link='logit')))$coefficients[2,1])

  }

  #store internal validated aucs for each imputed dataset
  internal_val_aucs_ml_mean_nb[[i]] <- mean(cv_aucs)
  internal_val_aucs_ml_SE_nb[[i]] <- sqrt(var(cv_aucs)) / sqrt(length(cv_aucs))

  #store internal validated citl for each imputed dataset
  internal_val_citl_ml_mean_nb[[i]] <- mean(cv_citl)
  internal_val_citl_ml_SE_nb[[i]] <- sqrt(var(cv_citl)) / sqrt(length(cv_citl))

  #store internal validated calibration slope for each imputed dataset
  internal_val_slope_ml_mean_nb[[i]] <- mean(cv_slope)
  internal_val_slope_ml_SE_nb[[i]] <- sqrt(var(cv_slope)) / sqrt(length(cv_slope))


  #store concatenated probs for each MI
  internal_val_prob_ml_nb[[i]] <- cv_prob
  #store concatenated obs for each MI
  internal_val_obs_ml_nb[[i]] <- cv_obs
  #store folds
  internal_val_fold_ml_nb[[i]] <- cv_fold

  #Store concatenated aucs, citl and slopes across all folds and multiple imputation
  internal_val_aucs_ml_nb[[i]] <- cv_aucs
  internal_val_citl_ml_nb[[i]] <- cv_citl
  internal_val_slope_ml_nb[[i]] <- cv_slope
}

pdf("supp_figure2C.pdf")
#look at distribution of probabilities for each imputed dataset
hist(internal_val_prob_ml_nb[[1]], main = "Naïve Bayes", xlab="Predicted Probability",
     xlim=c(0,1))
dev.off()

#AUC
pool_auc(
  est_auc = internal_val_aucs_ml_mean_nb,
  est_se = internal_val_aucs_ml_SE_nb,
  nimp = tempData$m,
  log_auc = T
)

#CITL - should be zero for internal validation
pool_auc_2(internal_val_citl_ml_mean_nb, internal_val_citl_ml_SE_nb,
           nimp = tempData$m)

#Calibration slope
pool_auc_2(internal_val_slope_ml_mean_nb, internal_val_slope_ml_SE_nb,
           nimp = tempData$m)

#look at final model training
finalModelsMl_nb[[1]]
```

```
#naive_bayes: laplace 0, adjust 1, usekernal TRUE, FALSE

#####################################################
#ML 3/5 - random forest

#lists to store internal validated mean & SE aucs, CITL (intercept with LP as offset term), and calibration slopes
internal_val_aucs_ml_mean_rf <- list()
internal_val_aucs_ml_SE_rf <- list()
internal_val_citl_ml_mean_rf <- list()
internal_val_citl_ml_SE_rf <- list()
internal_val_slope_ml_mean_rf <- list()
internal_val_slope_ml_SE_rf <- list()

#list to store concatenated predictions for each MI to look at their distribution
internal_val_prob_ml_rf <- list()
#list to store observations
internal_val_obs_ml_rf <- list()
#list to store fold
internal_val_fold_ml_rf <- list()

#Store model fit on entire dataset for each MI
finalModelsMl_rf <- list()

internal_val_aucs_ml_rf <- NULL
internal_val_citl_ml_rf <- NULL
internal_val_slope_ml_rf <- NULL

for (i in seq(1:tempData$m)){
  #Get imputed data
  lilly_ml_imp <- complete(tempData,i)
  #just take the columns we are testing except outcome as standardising first
  #Use 7 expert chosen variables
  lilly_imp_ml_final <- lilly_ml_imp[,c("CGISEV","GENDER",
                                        "PANSS_P2","PANSS_P3","PANSS_N4","PANSS_G6",
                                        "PAS_Highest")]
  #standardise the columns before building model
  preProcMlValues <- preProcess(lilly_imp_ml_final, method = c("center", "scale"))
  lilly_imp_ml_final_stand <- predict(preProcMlValues, lilly_imp_ml_final)
  #Add factor outcome back in before imputed
  lilly_imp_ml_final_stand$Non_Remission_12 <- data$Non_Remission_12
  #Remove rows with missing outcomes
  lilly_imp_ml_final_stand_MID <- lilly_imp_ml_final_stand[complete.cases(lilly_imp_ml_final_stand), ]

  set.seed(987)

  #caret train call to get model fitted on entire MI dataset
  #ML methods to try
  #logistic regression: glm (to check same results as above)
  #logistic regression:glmnet
  #naive bayes: naive_bayes
  #random forest: cforest
  #linear SVM: svmLinear2
  #radial SVM: svmRadial

  finalModelsMl_rf[[i]] <- train(Non_Remission_12 ~ ., #Outcome against all predictors
                                 data = lilly_imp_ml_final_stand_MID, #data
                                 method ="cforest", #TODO try different ML methods here
                                 metric="ROC", #performance metric - ROC for classification problems
                                 tuneLength = 5, #size of default tune grid #caret::getModelInfo("glmnet")
                                 trControl = control) #Our tuning method rules

  #10 fold CV repeated 50 times as per Frank Harrell "For 10-fold cv it is best to do 50 repeats"
  #https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation
  folds <-
    createMultiFolds(y = lilly_imp_ml_final_stand_MID$Non_Remission_12,
                     #returns training data indices
                     k = 10,
                     times = 50) #TO TEST QUICKER REDUCE THIS NUMBER, e.g. to 1

  #vector to store per fold AUCs, CITL and calibration slope
  cv_aucs <- NULL
  cv_citl <- NULL
  cv_slope <- NULL

  #vector to store per fold probs, obs and folds
  cv_prob <- NULL
  cv_obs <- NULL
  cv_fold <- NULL
```

```
for (j in seq(1:length(folds)))  {
  print(paste("Imputed dataset",i, "fold",j))

  trainCV <- lilly_imp_ml_final_stand_MID[folds[[j]], ]
  testCV <- lilly_imp_ml_final_stand_MID[-folds[[j]], ]

  #Caret train object
  #ML methods to try
  #logistic regression: glm (to check same results as above)
  #logistic regression:glmnet
  #naive bayes: naive_bayes
  #random forest: cforest
  #linear SVM: svmLinear2
  #radial SVM: svmRadial

  model <- train(Non_Remission_12 ~ ., #Outcome against all predictors
                 data = trainCV, #data
                 method ="cforest", #TODO try different ML methods here
                 metric="ROC", #performance metric - ROC for classification problems
                 tuneLength = 5, #size of default tune grid #caret::getModelInfo("glmnet")
                 trControl = control) #Our tuning method rules

  #Caret train predict - get probabilities
  #for caret predict.train object - have to specify level to get raw probabilities
  prob_test <- predict.train(model, testCV, type = "prob")$Y

  #store concatenated probs, obs and fold across internal validations
  cv_prob <- c(cv_prob, prob_test)
  cv_obs <- c(cv_obs, paste(testCV$Non_Remission_12))
  cv_fold <- c(cv_fold, rep(paste("fold",j),length(prob_test)))

  lp_test <- qlogis(prob_test)

  #Calculate AUC for test fold
  cv_aucs <- c(cv_aucs,
               roc(
                 predictor = prob_test,
                 response = testCV$Non_Remission_12,
                 ci = F,
                 levels = c("N", "Y"),
                 direction = "<"
               )$auc
  )

  #LP as offset for citl
  cv_citl <- c(cv_citl,
               summary(glm(Non_Remission_12 ~ offset(lp_test),
                           data = testCV, family=binomial(link='logit')))$coefficients[1,1])

  #
  cv_slope <- c(cv_slope,
                summary(glm(Non_Remission_12 ~ lp_test,
                            data = testCV, family=binomial(link='logit')))$coefficients[2,1])

}

#store internal validated aucs for each imputed dataset
internal_val_aucs_ml_mean_rf[[i]] <- mean(cv_aucs)
internal_val_aucs_ml_SE_rf[[i]] <- sqrt(var(cv_aucs)) / sqrt(length(cv_aucs))

#store internal validated citl for each imputed dataset
internal_val_citl_ml_mean_rf[[i]] <- mean(cv_citl)
internal_val_citl_ml_SE_rf[[i]] <- sqrt(var(cv_citl)) / sqrt(length(cv_citl))

#store internal validated calibration slope for each imputed dataset
internal_val_slope_ml_mean_rf[[i]] <- mean(cv_slope)
internal_val_slope_ml_SE_rf[[i]] <- sqrt(var(cv_slope)) / sqrt(length(cv_slope))


#store concatenated probs for each MI
internal_val_prob_ml_rf[[i]] <- cv_prob
#store concatenated obs for each MI
internal_val_obs_ml_rf[[i]] <- cv_obs
#store folds
internal_val_fold_ml_rf[[i]] <- cv_fold

#Store concatenated aucs, citl and slopes across all folds and multiple imputation
internal_val_aucs_ml_rf[[i]] <- cv_aucs
internal_val_citl_ml_rf[[i]] <- cv_citl
```

```
  internal_val_slope_ml_rf[[i]] <- cv_slope
}

pdf("supp_figure2D.pdf")
#look at distribution of probabilities for each imputed dataset
hist(internal_val_prob_ml_rf[[1]], main = "Random Forest", xlab="Predicted Probability",
    xlim=c(0,1))
dev.off()

#AUC
pool_auc(
  est_auc = internal_val_aucs_ml_mean_rf,
  est_se = internal_val_aucs_ml_SE_rf,
  nimp = tempData$m,
  log_auc = T
)

#CITL - should be zero for internal validation
pool_auc_2(internal_val_citl_ml_mean_rf, internal_val_citl_ml_SE_rf,
         nimp = tempData$m)

#Calibration slope
pool_auc_2(internal_val_slope_ml_mean_rf, internal_val_slope_ml_SE_rf,
         nimp = tempData$m)

#look at final model training
finalModelsMl_rf[[1]]
#cforest mtry 2,3,4,5,7

#######################################################
#ML 4/5 - linear SVM

#lists to store internal validated mean & SE aucs, CITL (intercept with LP as offset term), and calibration slopes
internal_val_aucs_ml_mean_lsvm <- list()
internal_val_aucs_ml_SE_lsvm <- list()
internal_val_citl_ml_mean_lsvm <- list()
internal_val_citl_ml_SE_lsvm <- list()
internal_val_slope_ml_mean_lsvm <- list()
internal_val_slope_ml_SE_lsvm <- list()

#list to store concatenated predictions for each MI to look at their distribution
internal_val_prob_ml_lsvm <- list()
#list to store observations
internal_val_obs_ml_lsvm <- list()
#list to store fold
internal_val_fold_ml_lsvm <- list()

#Store model fit on entire dataset for each MI
finalModelsMl_lsvm <- list()

internal_val_aucs_ml_lsvm <- NULL
internal_val_citl_ml_lsvm <- NULL
internal_val_slope_ml_lsvm <- NULL

for (i in seq(1:tempData$m)){
  #Get imputed data
  lilly_ml_imp <- complete(tempData,i)
  #just take the columns we are testing except outcome as standardising first
  #Use 7 expert chosen variables
  lilly_imp_ml_final <- lilly_ml_imp[,c("CGISEV","GENDER",
                                        "PANSS_P2","PANSS_P3","PANSS_N4","PANSS_G6",
                                        "PAS_Highest")]
  #standardise the columns before building model
  preProcMlValues <- preProcess(lilly_imp_ml_final, method = c("center", "scale"))
  lilly_imp_ml_final_stand <- predict(preProcMlValues, lilly_imp_ml_final)
  #Add factor outcome back in before imputed
  lilly_imp_ml_final_stand$Non_Remission_12 <- data$Non_Remission_12
  #Remove rows with missing outcomes
  lilly_imp_ml_final_stand_MID <- lilly_imp_ml_final_stand[complete.cases(lilly_imp_ml_final_stand), ]

  set.seed(987)

  #caret train call to get model fitted on entire MI dataset
  #ML methods to try
  #logistic regression: glm (to check same results as above)
  #logistic regression:glmnet
  #naive bayes: naive_bayes
  #random forest: cforest
  #linear SVM: svmLinear2
```

```
#radial SVM: svmRadial

finalModelsMl_lsvm[[i]] <- train(Non_Remission_12 ~ ., #Outcome against all predictors
                         data = lilly_imp_ml_final_stand_MID, #data
                         method ="svmLinear2", #TODO try different ML methods here
                         metric="ROC", #performance metric - ROC for classification problems
                         tuneLength = 5, #size of default tune grid #caret::getModelInfo("glmnet")
                         trControl = control) #Our tuning method rules

#10 fold CV repeated 50 times as per Frank Harrell "For 10-fold cv it is best to do 50 repeats"
#https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation
folds <-
  createMultiFolds(y = lilly_imp_ml_final_stand_MID$Non_Remission_12,
                #returns training data indices
                k = 10,
                times = 50) #TO TEST QUICKER REDUCE THIS NUMBER, e.g. to 1

#vector to store per fold AUCs, CITL and calibration slope
cv_aucs <- NULL
cv_citl <- NULL
cv_slope <- NULL

#vector to store per fold probs, obs and folds
cv_prob <- NULL
cv_obs <- NULL
cv_fold <- NULL

for (j in seq(1:length(folds)))  {
  print(paste("Imputed dataset",i, "fold",j))

  trainCV <- lilly_imp_ml_final_stand_MID[folds[[j]], ]
  testCV <- lilly_imp_ml_final_stand_MID[-folds[[j]], ]

  #Caret train object
  #ML methods to try
  #logistic regression: glm (to check same results as above)
  #logistic regression:glmnet
  #naive bayes: naive_bayes
  #random forest: cforest
  #linear SVM: svmLinear2
  #radial SVM: svmRadial

  model <- train(Non_Remission_12 ~ ., #Outcome against all predictors
                data = trainCV, #data
                method ="svmLinear2", #TODO try different ML methods here
                metric="ROC", #performance metric - ROC for classification problems
                tuneLength = 5, #size of default tune grid #caret::getModelInfo("glmnet")
                trControl = control) #Our tuning method rules

  #Caret train predict - get probabilities
  #for caret predict.train object - have to specify level to get raw probabilities
  prob_test <- predict.train(model, testCV, type = "prob")$Y

  #store concatenated probs, obs and fold across internal validations
  cv_prob <- c(cv_prob, prob_test)
  cv_obs <- c(cv_obs, paste(testCV$Non_Remission_12))
  cv_fold <- c(cv_fold, rep(paste("fold",j),length(prob_test)))

  lp_test <- qlogis(prob_test)

  #Calculate AUC for test fold
  cv_aucs <- c(cv_aucs,
              roc(
                predictor = prob_test,
                response = testCV$Non_Remission_12,
                ci = F,
                levels = c("N", "Y"),
                direction = "<"
              )$auc
  )

  #LP as offset for citl
  cv_citl <- c(cv_citl,
              summary(glm(Non_Remission_12 ~ offset(lp_test),
                          data = testCV, family=binomial(link='logit')))$coefficients[1,1])

  #
  cv_slope <- c(cv_slope,
                summary(glm(Non_Remission_12 ~ lp_test,
```

```
                    data = testCV, family=binomial(link='logit')))$coefficients[2,1])

  }

  #store internal validated aucs for each imputed dataset
  internal_val_aucs_ml_mean_lsvm[[i]] <- mean(cv_aucs)
  internal_val_aucs_ml_SE_lsvm[[i]] <- sqrt(var(cv_aucs)) / sqrt(length(cv_aucs))

  #store internal validated citl for each imputed dataset
  internal_val_citl_ml_mean_lsvm[[i]] <- mean(cv_citl)
  internal_val_citl_ml_SE_lsvm[[i]] <- sqrt(var(cv_citl)) / sqrt(length(cv_citl))

  #store internal validated calibration slope for each imputed dataset
  internal_val_slope_ml_mean_lsvm[[i]] <- mean(cv_slope)
  internal_val_slope_ml_SE_lsvm[[i]] <- sqrt(var(cv_slope)) / sqrt(length(cv_slope))


  #store concatenated probs for each MI
  internal_val_prob_ml_lsvm[[i]] <- cv_prob
  #store concatenated obs for each MI
  internal_val_obs_ml_lsvm[[i]] <- cv_obs
  #store folds
  internal_val_fold_ml_lsvm[[i]] <- cv_fold

  #Store concatenated aucs, citl and slopes across all folds and multiple imputation
  internal_val_aucs_ml_lsvm[[i]] <- cv_aucs
  internal_val_citl_ml_lsvm[[i]] <- cv_citl
  internal_val_slope_ml_lsvm[[i]] <- cv_slope
}

pdf("supp_figure2E.pdf")
#look at distribution of probabilities for each imputed dataset
hist(internal_val_prob_ml_lsvm[[1]], main = "Linear SVM", xlab="Predicted Probability",
    xlim=c(0,1))
dev.off()

#AUC
pool_auc(
  est_auc = internal_val_aucs_ml_mean_lsvm,
  est_se = internal_val_aucs_ml_SE_lsvm,
  nimp = tempData$m,
  log_auc = T
)

#CITL - should be zero for internal validation
pool_auc_2(internal_val_citl_ml_mean_lsvm, internal_val_citl_ml_SE_lsvm,
        nimp = tempData$m)

#Calibration slope
pool_auc_2(internal_val_slope_ml_mean_lsvm, internal_val_slope_ml_SE_lsvm,
        nimp = tempData$m)

#look at final model training
finalModelsMl_lsvm[[1]]
#svmLinear2 cost 0.25, 0.5, 1, 2, 4

#########################################################
#ML 5/5 - radial SVM

#lists to store internal validated mean & SE aucs, CITL (intercept with LP as offset term), and calibration slopes
internal_val_aucs_ml_mean_rsvm <- list()
internal_val_aucs_ml_SE_rsvm <- list()
internal_val_citl_ml_mean_rsvm <- list()
internal_val_citl_ml_SE_rsvm <- list()
internal_val_slope_ml_mean_rsvm <- list()
internal_val_slope_ml_SE_rsvm <- list()

#list to store concatenated predictions for each MI to look at their distribution
internal_val_prob_ml_rsvm <- list()
#list to store observations
internal_val_obs_ml_rsvm <- list()
#list to store fold
internal_val_fold_ml_rsvm <- list()

#Store model fit on entire dataset for each MI
finalModelsMl_rsvm <- list()

internal_val_aucs_ml_rsvm <- NULL
internal_val_citl_ml_rsvm <- NULL
```

```
internal_val_slope_ml_rsvm <- NULL

for (i in seq(1:tempData$m)){
 #Get imputed data
 lilly_ml_imp <- complete(tempData,i)
 #just take the columns we are testing except outcome as standardising first
 #Use 7 expert chosen variables
 lilly_imp_ml_final <- lilly_ml_imp[,c("CGISEV","GENDER",
                               "PANSS_P2","PANSS_P3","PANSS_N4","PANSS_G6",
                               "PAS_Highest")]
 #standardise the columns before building model
 preProcMlValues <- preProcess(lilly_imp_ml_final, method = c("center", "scale"))
 lilly_imp_ml_final_stand <- predict(preProcMlValues, lilly_imp_ml_final)
 #Add factor outcome back in before imputed
 lilly_imp_ml_final_stand$Non_Remission_12 <- data$Non_Remission_12
 #Remove rows with missing outcomes
 lilly_imp_ml_final_stand_MID <- lilly_imp_ml_final_stand[complete.cases(lilly_imp_ml_final_stand), ]

 set.seed(987)

 #caret train call to get model fitted on entire MI dataset
 #ML methods to try
 #logistic regression: glm (to check same results as above)
 #logistic regression:glmnet
 #naive bayes: naive_bayes
 #random forest: cforest
 #linear SVM: svmLinear2
 #radial SVM: svmRadial

 finalModelsMl_rsvm[[i]] <- train(Non_Remission_12 ~ ., #Outcome against all predictors
                          data = lilly_imp_ml_final_stand_MID, #data
                          method ="svmRadial", #TODO try different ML methods here
                          metric="ROC", #performance metric - ROC for classification problems
                          tuneLength = 5, #size of default tune grid #caret::getModelInfo("glmnet")
                          trControl = control) #Our tuning method rules

 #10 fold CV repeated 50 times as per Frank Harrell "For 10-fold cv it is best to do 50 repeats"
 #https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation
 folds <-
  createMultiFolds(y = lilly_imp_ml_final_stand_MID$Non_Remission_12,
               #returns training data indices
               k = 10,
               times = 50) #TO TEST QUICKER REDUCE THIS NUMBER, e.g. to 1

 #vector to store per fold AUCs, CITL and calibration slope
 cv_aucs <- NULL
 cv_citl <- NULL
 cv_slope <- NULL

 #vector to store per fold probs, obs and folds
 cv_prob <- NULL
 cv_obs <- NULL
 cv_fold <- NULL

 for (j in seq(1:length(folds)))  {
  print(paste("Imputed dataset",i, "fold",j))

  trainCV <- lilly_imp_ml_final_stand_MID[folds[[j]], ]
  testCV <- lilly_imp_ml_final_stand_MID[-folds[[j]], ]

  #Caret train object
  #ML methods to try
  #logistic regression: glm (to check same results as above)
  #logistic regression:glmnet
  #naive bayes: naive_bayes
  #random forest: cforest
  #linear SVM: svmLinear2
  #radial SVM: svmRadial

  model <- train(Non_Remission_12 ~ ., #Outcome against all predictors
              data = trainCV, #data
              method ="svmRadial", #TODO try different ML methods here
              metric="ROC", #performance metric - ROC for classification problems
              tuneLength = 5, #size of default tune grid #caret::getModelInfo("glmnet")
              trControl = control) #Our tuning method rules

  #Caret train predict - get probabilities
  #for caret predict.train object - have to specify level to get raw probabilities
  prob_test <- predict.train(model, testCV, type = "prob")$Y
```

```
  #store concatenated probs, obs and fold across internal validations
  cv_prob <- c(cv_prob, prob_test)
  cv_obs <- c(cv_obs, paste(testCV$Non_Remission_12))
  cv_fold <- c(cv_fold, rep(paste("fold",j),length(prob_test)))

  lp_test <- qlogis(prob_test)

  #Calculate AUC for test fold
  cv_aucs <- c(cv_aucs,
           roc(
             predictor = prob_test,
             response = testCV$Non_Remission_12,
             ci = F,
             levels = c("N", "Y"),
             direction = "<"
           )$auc
  )

  #LP as offset for citl
  cv_citl <- c(cv_citl,
           summary(glm(Non_Remission_12 ~ offset(lp_test),
                    data = testCV, family=binomial(link='logit')))$coefficients[1,1])

  #
  cv_slope <- c(cv_slope,
            summary(glm(Non_Remission_12 ~ lp_test,
                     data = testCV, family=binomial(link='logit')))$coefficients[2,1])

 }

 #store internal validated aucs for each imputed dataset
 internal_val_aucs_ml_mean_rsvm[[i]] <- mean(cv_aucs)
 internal_val_aucs_ml_SE_rsvm[[i]] <- sqrt(var(cv_aucs)) / sqrt(length(cv_aucs))

 #store internal validated citl for each imputed dataset
 internal_val_citl_ml_mean_rsvm[[i]] <- mean(cv_citl)
 internal_val_citl_ml_SE_rsvm[[i]] <- sqrt(var(cv_citl)) / sqrt(length(cv_citl))

 #store internal validated calibration slope for each imputed dataset
 internal_val_slope_ml_mean_rsvm[[i]] <- mean(cv_slope)
 internal_val_slope_ml_SE_rsvm[[i]] <- sqrt(var(cv_slope)) / sqrt(length(cv_slope))


 #store concatenated probs for each Ml
 internal_val_prob_ml_rsvm[[i]] <- cv_prob
 #store concatenated obs for each Ml
 internal_val_obs_ml_rsvm[[i]] <- cv_obs
 #store folds
 internal_val_fold_ml_rsvm[[i]] <- cv_fold

 #Store concatenated aucs, citl and slopes across all folds and multiple imputation
 internal_val_aucs_ml_rsvm[[i]] <- cv_aucs
 internal_val_citl_ml_rsvm[[i]] <- cv_citl
 internal_val_slope_ml_rsvm[[i]] <- cv_slope
}

pdf("supp_figure2F.pdf")
#look at distribution of probabilities for each imputed dataset
hist(internal_val_prob_ml_rsvm[[1]], main = "Radial SVM", xlab="Predicted Probability",
    xlim=c(0,1))
dev.off()

#AUC
pool_auc(
 est_auc = internal_val_aucs_ml_mean_rsvm,
 est_se = internal_val_aucs_ml_SE_rsvm,
 nimp = tempData$m,
 log_auc = T
)

#CITL - should be zero for internal validation
pool_auc_2(internal_val_citl_ml_mean_rsvm, internal_val_citl_ml_SE_rsvm,
        nimp = tempData$m)

#Calibration slope
pool_auc_2(internal_val_slope_ml_mean_rsvm, internal_val_slope_ml_SE_rsvm,
        nimp = tempData$m)
```

```
#look at final model training
finalModelsMl_rsvm[[1]]
#svmRadial sigma 0.1173, cost 0.25, 0.5, 1, 2, 4

#####################################################
#
##Multiple imputations ANOVA combine F statistics using miceadds::micombine.F
#Dataset 1 to 5
classifiers_df_mi1 <- data.frame(Group = rep(c("glm","glmnet","nb","rf","svml","svmr"),each=500),
                        AUCs = c(internal_val_aucs_clin[[1]], internal_val_aucs_ml_en[[1]], internal_val_aucs_ml_nb[[1]],
                                 internal_val_aucs_ml_rf[[1]], internal_val_aucs_ml_lsvm[[1]],
internal_val_aucs_ml_rsvm[[1]]),
                        CITLs = c(internal_val_citl_clin[[1]], internal_val_citl_ml_en[[1]], internal_val_citl_ml_nb[[1]],
                                  internal_val_citl_ml_rf[[1]], internal_val_citl_ml_lsvm[[1]], internal_val_citl_ml_rsvm[[1]]),
                        SLOPEs = c(internal_val_slope_clin[[1]], internal_val_slope_ml_en[[1]],
internal_val_slope_ml_nb[[1]],
                                   internal_val_slope_ml_rf[[1]], internal_val_slope_ml_lsvm[[1]],
internal_val_slope_ml_rsvm[[1]]))

classifiers_df_mi1$Group <- as.factor(classifiers_df_mi1$Group)

classifiers_df_mi2 <- data.frame(Group = rep(c("glm","glmnet","nb","rf","svml","svmr"),each=500),
                        AUCs = c(internal_val_aucs_clin[[2]], internal_val_aucs_ml_en[[2]], internal_val_aucs_ml_nb[[2]],
                                 internal_val_aucs_ml_rf[[2]], internal_val_aucs_ml_lsvm[[2]],
internal_val_aucs_ml_rsvm[[2]]),
                        CITLs = c(internal_val_citl_clin[[2]], internal_val_citl_ml_en[[2]], internal_val_citl_ml_nb[[2]],
                                  internal_val_citl_ml_rf[[2]], internal_val_citl_ml_lsvm[[2]], internal_val_citl_ml_rsvm[[2]]),
                        SLOPEs = c(internal_val_slope_clin[[2]], internal_val_slope_ml_en[[2]],
internal_val_slope_ml_nb[[2]],
                                   internal_val_slope_ml_rf[[2]], internal_val_slope_ml_lsvm[[2]],
internal_val_slope_ml_rsvm[[2]]))

classifiers_df_mi2$Group <- as.factor(classifiers_df_mi2$Group)

classifiers_df_mi3 <- data.frame(Group = rep(c("glm","glmnet","nb","rf","svml","svmr"),each=500),
                        AUCs = c(internal_val_aucs_clin[[3]], internal_val_aucs_ml_en[[3]], internal_val_aucs_ml_nb[[3]],
                                 internal_val_aucs_ml_rf[[3]], internal_val_aucs_ml_lsvm[[3]],
internal_val_aucs_ml_rsvm[[3]]),
                        CITLs = c(internal_val_citl_clin[[3]], internal_val_citl_ml_en[[3]], internal_val_citl_ml_nb[[3]],
                                  internal_val_citl_ml_rf[[3]], internal_val_citl_ml_lsvm[[3]], internal_val_citl_ml_rsvm[[3]]),
                        SLOPEs = c(internal_val_slope_clin[[3]], internal_val_slope_ml_en[[3]],
internal_val_slope_ml_nb[[3]],
                                   internal_val_slope_ml_rf[[3]], internal_val_slope_ml_lsvm[[3]],
internal_val_slope_ml_rsvm[[3]]))

classifiers_df_mi3$Group <- as.factor(classifiers_df_mi3$Group)

classifiers_df_mi4 <- data.frame(Group = rep(c("glm","glmnet","nb","rf","svml","svmr"),each=500),
                        AUCs = c(internal_val_aucs_clin[[4]], internal_val_aucs_ml_en[[4]], internal_val_aucs_ml_nb[[4]],
                                 internal_val_aucs_ml_rf[[4]], internal_val_aucs_ml_lsvm[[4]],
internal_val_aucs_ml_rsvm[[4]]),
                        CITLs = c(internal_val_citl_clin[[4]], internal_val_citl_ml_en[[4]], internal_val_citl_ml_nb[[4]],
                                  internal_val_citl_ml_rf[[4]], internal_val_citl_ml_lsvm[[4]], internal_val_citl_ml_rsvm[[4]]),
                        SLOPEs = c(internal_val_slope_clin[[4]], internal_val_slope_ml_en[[4]],
internal_val_slope_ml_nb[[4]],
                                   internal_val_slope_ml_rf[[4]], internal_val_slope_ml_lsvm[[4]],
internal_val_slope_ml_rsvm[[4]]))

classifiers_df_mi4$Group <- as.factor(classifiers_df_mi4$Group)

classifiers_df_mi5 <- data.frame(Group = rep(c("glm","glmnet","nb","rf","svml","svmr"),each=500),
                        AUCs = c(internal_val_aucs_clin[[5]], internal_val_aucs_ml_en[[5]], internal_val_aucs_ml_nb[[5]],
                                 internal_val_aucs_ml_rf[[5]], internal_val_aucs_ml_lsvm[[5]],
internal_val_aucs_ml_rsvm[[5]]),
                        CITLs = c(internal_val_citl_clin[[5]], internal_val_citl_ml_en[[5]], internal_val_citl_ml_nb[[5]],
                                  internal_val_citl_ml_rf[[5]], internal_val_citl_ml_lsvm[[5]], internal_val_citl_ml_rsvm[[5]]),
                        SLOPEs = c(internal_val_slope_clin[[5]], internal_val_slope_ml_en[[5]],
internal_val_slope_ml_nb[[5]],
                                   internal_val_slope_ml_rf[[5]], internal_val_slope_ml_lsvm[[5]],
internal_val_slope_ml_rsvm[[5]]))

classifiers_df_mi5$Group <- as.factor(classifiers_df_mi5$Group)

#AUCs
leveneTest(AUCs~Group, data = classifiers_df_mi1) #homogeneous
ggqqplot(classifiers_df_mi1$AUCs)

#ANOVA assume equal variances Imputation 1
df <- classifiers_df_mi1
```

```
#Use logit transformation for AUC as per Steyerberg but take away a constant as includes 1
#(logit 1 is infinity)
df$AUCs <- df$AUCs - 0.0001
df$AUCs <- qlogis(df$AUCs)
oneway.test(AUCs~Group, data = df, var.equal = TRUE)
#F = 7.1, num df = 5, denom df = 2994, p-value = 0.000001

#Tukey HSD Post-hoc Tests use median p-rule to combine
tukey_hsd(df, AUCs~Group)

#ANOVA assume equal variances Imputation 2
df <- classifiers_df_mi2
#Use logit transformation for AUC as per Steyerberg but take away a constant as includes 1
#(logit 1 is infinity)
df$AUCs <- df$AUCs - 0.0001
df$AUCs <- qlogis(df$AUCs)
oneway.test(AUCs~Group, data = df, var.equal = TRUE)
#F = 7.9, num df = 5, denom df = 2994, p-value = 0.0000002

#Tukey HSD Post-hoc Tests use median p-rule to combine
tukey_hsd(df, AUCs~Group)

#ANOVA assume equal variances Imputation 3
df <- classifiers_df_mi3
#Use logit transformation for AUC as per Steyerberg but take away a constant as includes 1
#(logit 1 is infinity)
df$AUCs <- df$AUCs - 0.0001
df$AUCs <- qlogis(df$AUCs)
oneway.test(AUCs~Group, data = df, var.equal = TRUE)
#F = 8.3, num df = 5, denom df = 2994, p-value = 0.00000009

#Tukey HSD Post-hoc Tests use median p-rule to combine
tukey_hsd(df, AUCs~Group)

#ANOVA assume equal variances Imputation 4
df <- classifiers_df_mi4
#Use logit transformation for AUC as per Steyerberg but take away a constant as includes 1
#(logit 1 is infinity)
df$AUCs <- df$AUCs - 0.0001
df$AUCs <- qlogis(df$AUCs)
oneway.test(AUCs~Group, data = df, var.equal = TRUE)
#F = 7.3, num df = 5, denom df = 2994, p-value = 0.0000009

#Tukey HSD Post-hoc Tests use median p-rule to combine
tukey_hsd(df, AUCs~Group)

#ANOVA assume equal variances Imputation 5
df <- classifiers_df_mi5
#Use logit transformation for AUC as per Steyerberg but take away a constant as includes 1
#(logit 1 is infinity)
df$AUCs <- df$AUCs - 0.0001
df$AUCs <- qlogis(df$AUCs)
oneway.test(AUCs~Group, data = df, var.equal = TRUE)
#F = 8.4, num df = 5, denom df = 2994, p-value = 0.00000007

#Tukey HSD Post-hoc Tests use median p-rule to combine
tukey_hsd(df, AUCs~Group)

#Combine F across multiple imputations
micombine.F(Fvalues = c(7.1,7.9,8.3,7.3,8.4),
         df1 = 5)
#AUCs
#F(5, 398.66)=7.225     p=0

#CITLs
leveneTest(CITLs~Group, data = classifiers_df_mi1) #heterogeneous
ggqqplot(classifiers_df_mi1$CITLs)

#ANOVA assume equal variances Imputation 1
oneway.test(CITLs~Group, data = classifiers_df_mi1, var.equal = FALSE)
#F = 104, num df = 5, denom df = 1382, p-value <0.0000000000000002

#Games Howell Post-hoc Tests
games_howell_test(CITLs~Group, data = classifiers_df_mi1, detailed = TRUE)
#just get p-values without scientific notation
games_howell_test(CITLs~Group, data = classifiers_df_mi1, detailed = TRUE)$p.adj

#ANOVA assume equal variances Imputation 2
oneway.test(CITLs~Group, data = classifiers_df_mi2, var.equal = FALSE)
```

```
#F = 118, num df = 5, denom df = 1381, p-value <0.0000000000000002

#Games Howell Post-hoc Tests
games_howell_test(CITLs~Group, data = classifiers_df_mi2, detailed = TRUE)
#just get p-values without scientific notation
games_howell_test(CITLs~Group, data = classifiers_df_mi2, detailed = TRUE)$p.adj

#ANOVA assume equal variances Imputation 3
oneway.test(CITLs~Group, data = classifiers_df_mi3, var.equal = FALSE)
#F = 125, num df = 5, denom df = 1383, p-value <0.0000000000000002

#Games Howell Post-hoc Tests
games_howell_test(CITLs~Group, data = classifiers_df_mi3, detailed = TRUE)
#just get p-values without scientific notation
games_howell_test(CITLs~Group, data = classifiers_df_mi3, detailed = TRUE)$p.adj

#ANOVA assume equal variances Imputation 4
oneway.test(CITLs~Group, data = classifiers_df_mi4, var.equal = FALSE)
#F = 111, num df = 5, denom df = 1383, p-value <0.0000000000000002

#Games Howell Post-hoc Tests
games_howell_test(CITLs~Group, data = classifiers_df_mi4, detailed = TRUE)
#just get p-values without scientific notation
games_howell_test(CITLs~Group, data = classifiers_df_mi4, detailed = TRUE)$p.adj

#ANOVA assume equal variances Imputation 5
oneway.test(CITLs~Group, data = classifiers_df_mi5, var.equal = FALSE)
#F = 124, num df = 5, denom df = 1383, p-value <0.0000000000000002

#Games Howell Post-hoc Tests
games_howell_test(CITLs~Group, data = classifiers_df_mi5, detailed = TRUE)
#just get p-values without scientific notation
games_howell_test(CITLs~Group, data = classifiers_df_mi5, detailed = TRUE)$p.adj

#Combine F across multiple imputations
micombine.F(Fvalues = c(104,118,125,111,124),
        df1 = 5)
#CITLs
#F(5, 5.87)=56.356     p=0.00007

#SLOPEs
leveneTest(SLOPEs~Group, data = classifiers_df_mi1) #homogeneous
ggqqplot(classifiers_df_mi1$SLOPEs)

#ANOVA assume equal variances Imputation 1
oneway.test(SLOPEs~Group, data = classifiers_df_mi1, var.equal = TRUE)
#F = 1.9, num df = 5, denom df = 2994, p-value = 0.09

#ANOVA assume equal variances Imputation 2
oneway.test(SLOPEs~Group, data = classifiers_df_mi2, var.equal = TRUE)
#F = 1.3, num df = 5, denom df = 2994, p-value = 0.2

#ANOVA assume equal variances Imputation 3
oneway.test(SLOPEs~Group, data = classifiers_df_mi3, var.equal = TRUE)
#F = 1.3, num df = 5, denom df = 2994, p-value = 0.3

#ANOVA assume equal variances Imputation 4
oneway.test(SLOPEs~Group, data = classifiers_df_mi4, var.equal = TRUE)
#F = 1.1, num df = 5, denom df = 2994, p-value = 0.4

#ANOVA assume equal variances Imputation 5
oneway.test(SLOPEs~Group, data = classifiers_df_mi5, var.equal = TRUE)
#F = 1.3, num df = 5, denom df = 2994, p-value = 0.3

#Combine F across multiple imputations
micombine.F(Fvalues = c(1.9,1.3,1.3,1.1,1.3),
        df1 = 5)
#CITLs
#F(5, 218.1)=1.139     p=0.3405


#Plots
df <- read_csv("df_1.csv")
pdf("figure1A.pdf")
ggplot(df, aes(x = Model, y = `c-statistic`)) + geom_point(size = 2) + geom_errorbar(aes(ymin = Low95CI, ymax =
High95CI), width =0.2)+
  theme_bw() + scale_y_continuous(minor_breaks = seq(0, 20, 1)) +
  theme(axis.text.x = element_text(angle=90, vjust=.5, hjust=1))
dev.off()
```

```
df <- read_csv("df_2.csv")
pdf("figure1B.pdf")
ggplot(df, aes(x = Model, y = CITL)) + geom_point(size = 2) + geom_errorbar(aes(ymin = Low95CI, ymax = High95CI),
width =0.2)+
  theme_bw() + scale_y_continuous(minor_breaks = seq(0, 20, 1)) +
  theme(axis.text.x = element_text(angle=90, vjust=.5, hjust=1))
dev.off()

df <- read_csv("df_3.csv")
pdf("figure1C.pdf")
ggplot(df, aes(x = Model, y = `Calibration Slope`)) + geom_point(size = 2) + geom_errorbar(aes(ymin = Low95CI, ymax =
High95CI), width =0.2)+
  theme_bw() + scale_y_continuous(minor_breaks = seq(0, 20, 1)) +
  theme(axis.text.x = element_text(angle=90, vjust=.5, hjust=1))
dev.off()

df <- read_csv("df_4.csv")
pdf("figure2A.pdf")
ggplot(df, aes(x = Model, y = `c-statistic`)) + geom_point(size = 2) + geom_errorbar(aes(ymin = Low95CI, ymax =
High95CI), width =0.2)+
  theme_bw() +
  theme(axis.text.x = element_text(angle=90, vjust=.5, hjust=1)) +
  geom_bracket(
    xmin = c("GLM", "GLMnet","GLMnet", "Naive Bayes","Naive Bayes"),
    xmax = c("Naive Bayes", "SVM (Linear)","SVM (Radial)", "SVM (Linear)", "SVM (Radial)"),
    y.position = c(0.72,0.73,0.735,0.725,0.74), label = c("**", "*","**", "****", "****"),
    tip.length = 0.01
  )
dev.off()

df <- read_csv("df_5.csv")
pdf("figure2B.pdf")
ggplot(df, aes(x = Model, y = CITL)) + geom_point(size = 2) + geom_errorbar(aes(ymin = Low95CI, ymax = High95CI),
width =0.2)+
  theme_bw() +
  theme(axis.text.x = element_text(angle=90, vjust=.5, hjust=1)) +
  geom_bracket(
    xmin = c("GLM","GLM","GLMnet","GLMnet","Naive Bayes","Naive Bayes","Naive Bayes","Random Forest","SVM (Linear)"),
    xmax = c("Naive Bayes","SVM (Radial)","Naive Bayes","SVM (Radial)","Random Forest","SVM (Linear)","SVM (Radial)","SVM
(Radial)","SVM (Radial)"),
    y.position = c(0.65,0.85,0.55,0.80,0.60,0.70,0.75,0.65,0.55), label =
c("****","***","****","****","****","****","****","****","****"),
    tip.length = 0.01
  )
dev.off()


df <- read_csv("df_6.csv")
pdf("figure2C.pdf",7,10)
ggplot(df, aes(x = Model, y = `Calibration Slope`)) + geom_point(size = 2) + geom_errorbar(aes(ymin = Low95CI, ymax =
High95CI), width =0.2)+
  theme_bw() + scale_y_continuous(minor_breaks = seq(0, 20, 1)) +
  theme(axis.text.x = element_text(angle=90, vjust=.5, hjust=1))
dev.off()


#library(gplots)
#plotmeans(AUCs~Group, data = classifiers_df_mi1, connect = FALSE)
#plot(AUCs~Group, data = classifiers_df_mi1, notch=TRUE)
```

Appendix 5 ICD-10 codes for Chapter 5

Delirium Coding

F05 Delirium not induced by alcohol and other psychoactive substances

Dementia Coding

F00 Dementia in Alzheimer's disease

F01 Vascular Dementia

F02 Dementia in other diseases classified elsewhere

F03 Unspecified dementia

F1073 Residual & Late-onset Psychotic Dementia Due to Use of Alcohol

F1173 Residual & Late-onset Psychotic Dementia Due to Use of Opioids

F1273 Residual & Late-onset Psychotic Dementia Due to Use of Cannabinoids

F1373 Residual & Late-onset Psychotic Dementia Due to Use of Sedatives/Hypnotics

F1473 Residual & Late-onset Psychotic Dementia Due to Use of Cocaine

F1573 Residual & Late-onset Psychotic Dementia Due to Use of Other Stimulants Including Caffeine

F1673 Residual & Late-onset Psychotic Dementia Due to Use of Hallucinogens

F1773 Residual & Late-onset Psychotic Dementia Due to Use of Tobacco

F1873 Residual & Late-onset Psychotic Dementia Due to Use of Volatile Solvents

F1973 Residual & Late-onset Psychotic Dementia Due to Use of Multiple Drugs/Psychoactive Substances

Appendix 6 Post-model assumption testing for Chapter 5

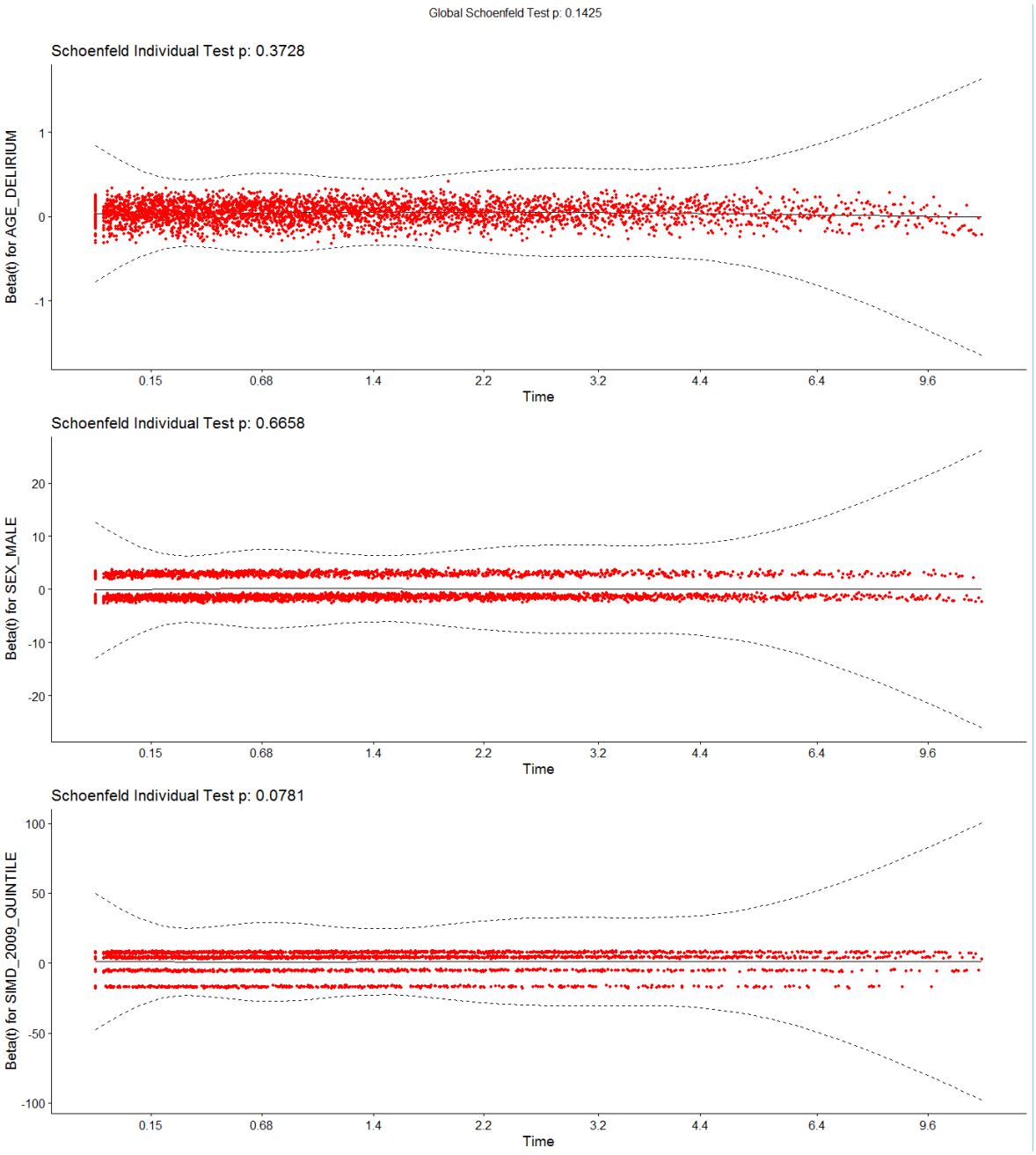


Figure 1 - The proportional hazard assumption is supported by a non-significant relationship between residuals and time for each of the covariates and the global test. The plots of the Schoenfeld residuals are independent of time.

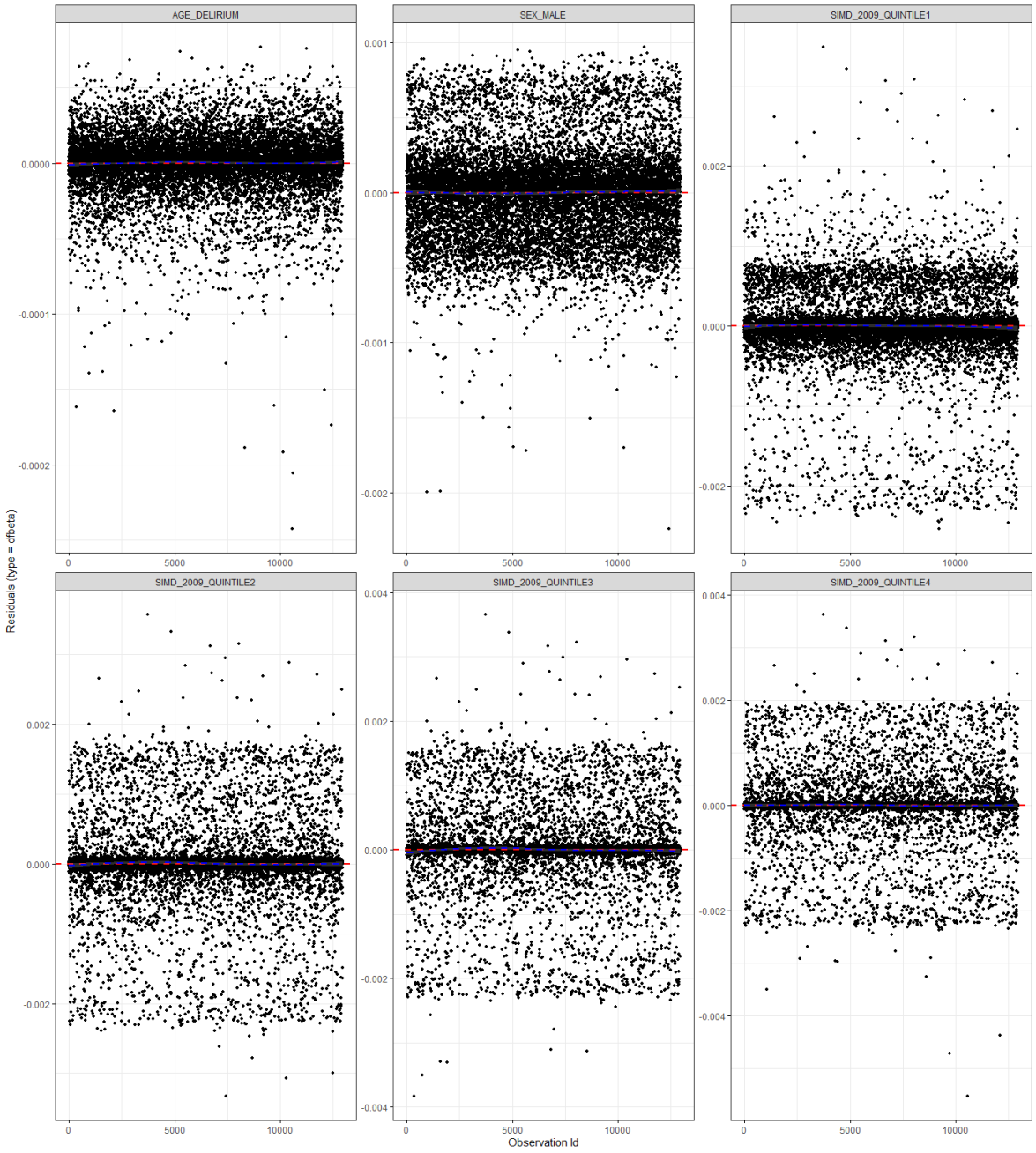


Figure 2 - The above index plots show that comparing the magnitudes of the largest DFBETA values to the regression coefficients suggests that none of the observations are influential individually according to the cut off proposed by Belsley, Kuh and Welch (values larger than 2/sqrt(n) are considered highly influential, in our case values larger than 0.018). (Belsey et al., 1980)

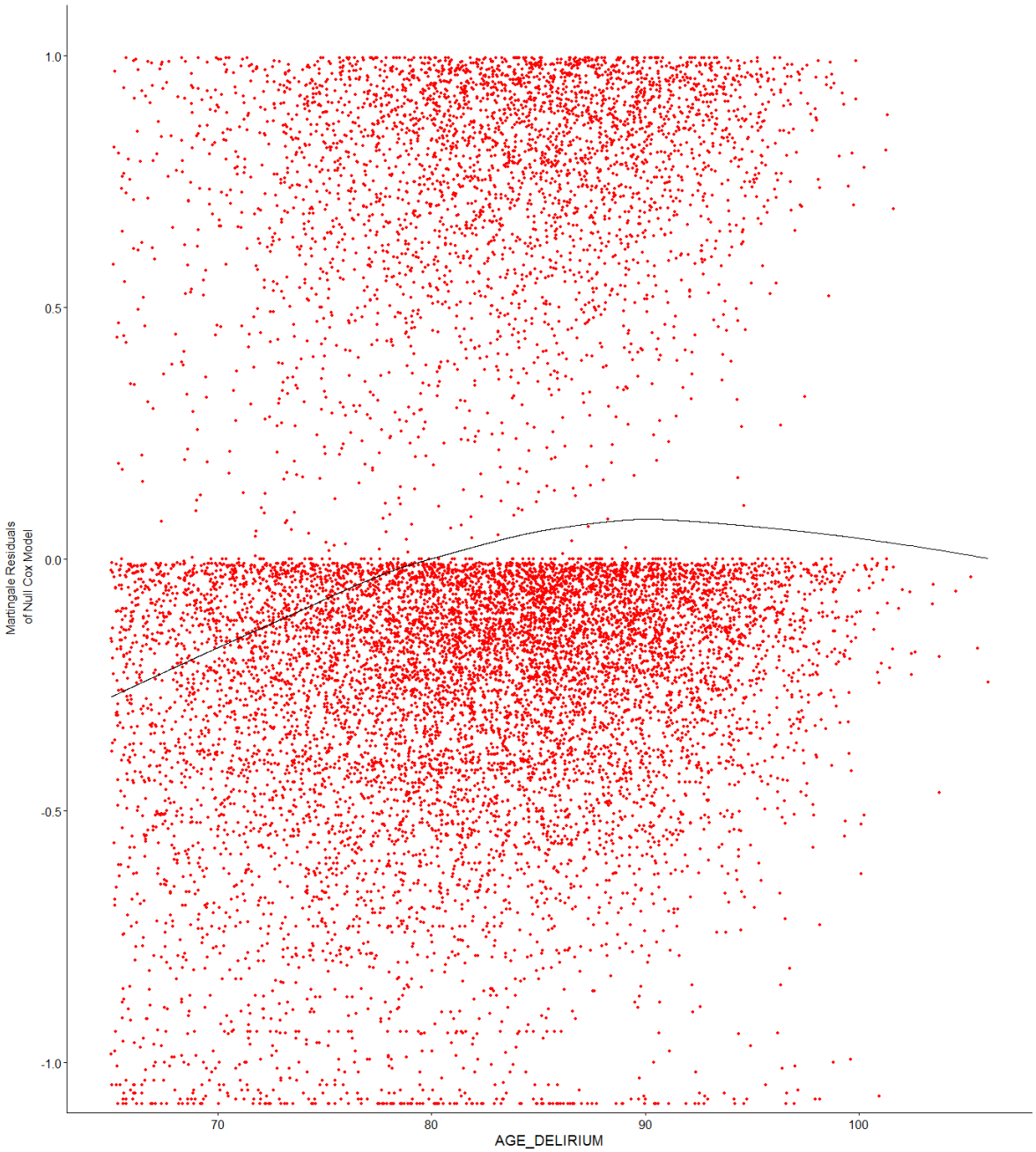


Figure 3 – We assessed for nonlinearity by plotting Martingale residuals of the null cox proportional hazards model against the continuous covariate, age (nonlinearity is not a concern for categorical variables). The fitted line suggests that age has a non-linear functional form, so the final Cox model was refitted using a penalised cubic spline term for age.

Appendix 7 R code for Chapter 5

```
library(cmprsk)
library(survminer)
library(readr)
library(ggplot2)
library(scales)
library(dotwhisker)
library(survival)

#working directory
setwd("S:/Sam/Analysis_3/")

options(scipen=999)

data = read_csv("final_ggc_65.csv")
data = data[which(data$TIME_DEMENTIA_DEATH>0),]
data$DELIRIUM_DATE = as.Date(data$DELIRIUM_DATE, format = "%d/%m/%Y")
data$DELIRIUM_YEAR = format(data$DELIRIUM_DATE, "%Y")
data$DEMENTIA_ON_DEATH = as.factor(data$DEMENTIA_ON_DEATH)
data$DEC = data$AGE_DELIRIUM/10
data$SIMD_2009_QUINTILE = as.factor(data$SIMD_2009_QUINTILE)
data$SIMD_2009_QUINTILE = relevel(data$SIMD_2009_QUINTILE, ref = 5)

# TIME_DEMENTIA_DEATH denotes the survival time to the occurrence of the first event
# STATUS_DEMENTIA_DEATH is the event type indicator:
# 1: Dementia diagnosis
# 2: Death before dementia diagnosis
# 0: Censored observation: alive at end of follow-up

# Demographics
summary(as.factor(data$STATUS_DEMENTIA_DEATH))
summary(as.factor(data$DELIRIUM_YEAR))
summary(as.factor(data$SIMD_2009_QUINTILE))
summary(as.factor(data$SEX_MALE))
summary(data$AGE_DELIRIUM)
sd(data$AGE_DELIRIUM)

by(data$SIMD_2009_QUINTILE,as.factor(data$STATUS_DEMENTIA_DEATH), summary)
by(as.factor(data$SEX_MALE),as.factor(data$STATUS_DEMENTIA_DEATH), summary)
by(data$AGE_DELIRIUM,as.factor(data$STATUS_DEMENTIA_DEATH), summary)
by(data$AGE_DELIRIUM,as.factor(data$STATUS_DEMENTIA_DEATH), sd)
by(data$DEMENTIA_ON_DEATH, as.factor(data$STATUS_DEMENTIA_DEATH), summary)

#Stacked bar chart of outcomes
Outcome = c(rep("Censored",2), rep("Dementia diagnosis",2), rep("Death without dementia",2))
dem_on_death = rep(c("No","Dementia diagnosed on death"),3)
Patients = c(3631,0,2887,643,5788,0)
graph_data = data.frame(Outcome, dem_on_death, Patients)
pdf("outcomes.pdf", 10, 7)
ggplot(graph_data, aes(fill=dem_on_death, y=Patients, x=Outcome)) +
  geom_bar(position="stack", stat = "identity", alpha = 0.5) +
  annotate("text", x=c(3,3), y =c(3250, 1000), label = c("On death","Before death"), color = "white", size = 6,
fontface="bold")+
  theme_bw()+theme(legend.position = "none", text = element_text(size=20), axis.title.x = element_blank())+
  scale_y_continuous(minor_breaks = seq(0,6000,500), breaks = seq(0,6000,1000))+
  scale_fill_manual(values = c("red","blue"))+
  ggtitle("Outcomes for patients with an index episode of delirium")
dev.off()

#hist(data$DELIRIUM_DATE, breaks = "months", freq = TRUE)
#grid()

#Count by year and month
new = data.frame(table(format(data$DELIRIUM_DATE, "%Y-%m")))
#Append a day
new$Var1 = paste0(new$Var1, "-1")
#TUrn back into a date
new$Var1 = as.Date(new$Var1, format = "%Y-%m-%d")
pdf("delirium.pdf",10,5)
#plot using scale_x_date with 6 month breaks
ggplot(data = new, aes(x=Var1, y=Freq)) + geom_bar(stat="identity", colour ="grey20", fill = "white", size = 0.25)+
  scale_x_date(labels = date_format("%b %Y"), breaks = date_breaks("12 months"))+theme_bw()+
  theme(axis.text.x = element_text(angle = 90, vjust =0.5, hjust=1))+xlab("Date")+ylab("Monthly Frequency of Delirium")+
  scale_y_continuous(minor_breaks = seq(0,225,5), breaks = seq(0,225,50))
dev.off()
```

```
summary(data$TIME_DEMENTIA_DEATH)

# Figure 1: Plot cumulative incidence functions for dementia and death without dementia
ci_fit = cuminc(ftime = data$TIME_DEMENTIA_DEATH/365.25,
            fstatus = data$STATUS_DEMENTIA_DEATH,
            cencode = 0)
ci_fit

#cumulative incidence 6 months
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >0.5)][1] #Dementia
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >0.5)][1]-1.96*sqrt(ci_fit$`1 1`$var[which(ci_fit$`1 1`$time >0.5)][1])
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >0.5)][1]+1.96*sqrt(ci_fit$`1 1`$var[which(ci_fit$`1 1`$time >0.5)][1])
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >0.5)][1] #Death without dementia
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >0.5)][1]-1.96*sqrt(ci_fit$`1 2`$var[which(ci_fit$`1 2`$time >0.5)][1])
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >0.5)][1]+1.96*sqrt(ci_fit$`1 2`$var[which(ci_fit$`1 2`$time >0.5)][1])
#cumulative incidence 1 year
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >1)][1] #Dementia
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >1)][1]-1.96*sqrt(ci_fit$`1 1`$var[which(ci_fit$`1 1`$time >1)][1])
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >1)][1]+1.96*sqrt(ci_fit$`1 1`$var[which(ci_fit$`1 1`$time >1)][1])
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >1)][1] #Death without dementia
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >1)][1]-1.96*sqrt(ci_fit$`1 2`$var[which(ci_fit$`1 2`$time >1)][1])
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >1)][1]+1.96*sqrt(ci_fit$`1 2`$var[which(ci_fit$`1 2`$time >1)][1])
#cumulative incidence 5 years
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >5)][1] #Dementia
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >5)][1]-1.96*sqrt(ci_fit$`1 1`$var[which(ci_fit$`1 1`$time >5)][1])
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >5)][1]+1.96*sqrt(ci_fit$`1 1`$var[which(ci_fit$`1 1`$time >5)][1])
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >5)][1] #Death without dementia
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >5)][1]-1.96*sqrt(ci_fit$`1 2`$var[which(ci_fit$`1 2`$time >5)][1])
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >5)][1]+1.96*sqrt(ci_fit$`1 2`$var[which(ci_fit$`1 2`$time >5)][1])
#cumulative incidence 10 years
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >10)][1] #Dementia
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >10)][1]-1.96*sqrt(ci_fit$`1 1`$var[which(ci_fit$`1 1`$time >10)][1])
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >10)][1]+1.96*sqrt(ci_fit$`1 1`$var[which(ci_fit$`1 1`$time >10)][1])
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >10)][1] #Death without dementia
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >10)][1]-1.96*sqrt(ci_fit$`1 2`$var[which(ci_fit$`1 2`$time >10)][1])
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >10)][1]+1.96*sqrt(ci_fit$`1 2`$var[which(ci_fit$`1 2`$time >10)][1])
#cumulative incidence 15 years
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >15)][1] #Dementia
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >15)][1]-1.96*sqrt(ci_fit$`1 1`$var[which(ci_fit$`1 1`$time >15)][1])
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >15)][1]+1.96*sqrt(ci_fit$`1 1`$var[which(ci_fit$`1 1`$time >15)][1])
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >15)][1] #Death without dementia
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >15)][1]-1.96*sqrt(ci_fit$`1 2`$var[which(ci_fit$`1 2`$time >15)][1])
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >15)][1]+1.96*sqrt(ci_fit$`1 2`$var[which(ci_fit$`1 2`$time >15)][1])
#cumulative incidence 20 years
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >20)][1] #Dementia
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >20)][1]-1.96*sqrt(ci_fit$`1 1`$var[which(ci_fit$`1 1`$time >20)][1])
ci_fit$`1 1`$est[which(ci_fit$`1 1`$time >20)][1]+1.96*sqrt(ci_fit$`1 1`$var[which(ci_fit$`1 1`$time >20)][1])
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >20)][1] #Death without dementia
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >20)][1]-1.96*sqrt(ci_fit$`1 2`$var[which(ci_fit$`1 2`$time >20)][1])
ci_fit$`1 2`$est[which(ci_fit$`1 2`$time >20)][1]+1.96*sqrt(ci_fit$`1 2`$var[which(ci_fit$`1 2`$time >20)][1])

pdf("CIF.pdf", 15, 10)
ggcompetingrisks(ci_fit, multiple_panels = F, conf.int = TRUE) + theme_bw()+
  guides(linetype = FALSE) + scale_y_continuous(minor_breaks = seq(0,1,0.05),breaks = seq(0,1,0.1))+
  scale_x_continuous(minor_breaks = seq(0,25,1), breaks = seq(0,25,5)) +xlab("Time (years)")+
  scale_color_manual(labels=c("Dementia","Death without dementia"), values = c("blue","red"))+
  scale_fill_manual(labels=c("Dementia","Death without dementia"), values = c("blue","red"))+
  theme(text = element_text(size = 20)) + labs(fill = "Outcome", colour ="Outcome")
dev.off()

# confidence intervals are 1.96+/- sqrt variance from cuminc function of cmprsk package as per Frank Harrell
# discourse.datamethods.org/t/95-ci-around-cumulative-incidence-estimate/3948

#Cause specific hazard model for dementia treats death before dementia as censored
#
#time in years
res.cox.dem = coxph(Surv(TIME_DEMENTIA_DEATH/365.25, STATUS_DEMENTIA) ~ AGE_DELIRIUM + SEX_MALE +
SIMD_2009_QUINTILE, data = data)
summary(res.cox.dem)

#sthda.com/english/wiki/cox-model-assumptions
#test proportional-hazards assumption
test.ph = cox.zph(res.cox.dem)
test.ph
ggcoxzph(test.ph)

#influential observations
ggcoxdiagnostics(res.cox.dem, type = "dfbeta", linear.predictions = FALSE, ggtheme = theme_bw())
#values larger than 2/sqrt(n) are considered highly influential = 0.018
```

```
ggcoxdiagnostics(res.cox.dem, type = "deviance", linear.predictions = FALSE, ggtheme = theme_bw())
#deviance residuals should be roughtly symetrically distributed around zero with a standard deviation of 1

ggcoxfunctional(Surv(TIME_DEMENTIA_DEATH/365.25, STATUS_DEMENTIA) ~ AGE_DELIRIUM, data = data, ylim = c(-1,1))
#Is the functional form of age linear. Assess using martingale's residuals from null model

#95%CI for loess fit
smoothSEcurve = function(yy, xx){
  # use after a call to "plot"
  # fit a lowess curve and 95% confidence interval curve
  # make list of x values
  xx.list = min(xx) + ((0:100)/100)*(max(xx) - min(xx))
  # Then fit loess function through the points (xx, yy)
  #  at the listed values
  yy.xx = predict(loess(yy ~ xx), se=T,
              newdata=data.frame(xx=xx.list))
  lines(yy.xx$fit ~ xx.list, lwd=2)
  lines(yy.xx$fit -
        qt(0.975, yy.xx$df)*yy.xx$se.fit ~ xx.list, lty=2)
  lines(yy.xx$fit +
        qt(0.975, yy.xx$df)*yy.xx$se.fit ~ xx.list, lty=2)
}

#SAME AS ABOVE ggcoxfunctional
res.cox.dem.0 = coxph(Surv(TIME_DEMENTIA_DEATH/365.25, STATUS_DEMENTIA) ~ 1, data = data)
rr.0 = residuals(res.cox.dem.0, type = "martingale")
plot(rr.0 ~ AGE_DELIRIUM, data=data)
smoothSEcurve(rr.0, data$AGE_DELIRIUM)
#Is the functional form of age linear. Assess using martingale's residuals from null model

rr.final = residuals(res.cox.dem, type = "martingale")
plot(rr.final~ AGE_DELIRIUM, data = data)
smoothSEcurve(rr.final, data$AGE_DELIRIUM)
#Is the functional form of age linear. Assess using martingale's residuals from fully adjusted model including age as
covariate

#Age does not have a linear functional form so remodel using a penalised spline term for age
#See https://stats.stackexchange.com/questions/362510 for interpretation of above martingale's residuals plots

#add spline setting df = 0 then the AIC = (loglik -df) is used to choose an "optimal" degrees of freedom
res.cox.dem.spline = coxph(Surv(TIME_DEMENTIA_DEATH/365.25, STATUS_DEMENTIA) ~ pspline(AGE_DELIRIUM, df=0) +
SEX_MALE + SIMD_2009_QUINTILE, data = data)
summary(res.cox.dem.spline)

#without age to do likelihood ratio test
res.cox.dem.noage = coxph(Surv(TIME_DEMENTIA_DEATH/365.25, STATUS_DEMENTIA) ~ SEX_MALE +
SIMD_2009_QUINTILE, data = data)
#p value for age term
#according to Robertson
anova(res.cox.dem.spline, res.cox.dem.noage)
#according to stack exchange https://stats.stackexchange.com/questions/197179
anova(res.cox.dem.spline, res.cox.dem)

#Plot the age term on a log-hazard scale
termplot(res.cox.dem.spline, se=T, terms = 1, ylabs = "Log hazard")

pdf("age_spline.pdf",10,7)
#Plot the age term on a hazard scale
predicted = predict(res.cox.dem.spline, type = "terms", se.fit = TRUE, terms = 1)
plot(data$AGE_DELIRIUM, exp(predicted$fit), type ="n", ylim = c(0,2), ylab = "Hazard ratio (df = 5.3)", xlab = "Age at
delirium diagnosis")
lines(smooth.spline(data$AGE_DELIRIUM, exp(predicted$fit)), col = "red", lty = 1)
lines(smooth.spline(data$AGE_DELIRIUM, exp(predicted$fit + 1.96 * predicted$se.fit)), col = "orange", lty = 2)
lines(smooth.spline(data$AGE_DELIRIUM, exp(predicted$fit - 1.96 * predicted$se.fit)), col = "orange", lty = 2)
abline(h=1, lty=2)
boxplot(add=T, data$AGE_DELIRIUM, horizontal = T, at = 0, boxwex = 0.1)
dev.off()

#forest plot
#set up data
term = c("Male", "Female", "SIMD2009 Quintile 1", "SIMD2009 Quintile 2", "SIMD2009 Quintile 3", "SIMD2009 Quintile 4",
"SIMD2009 Quintile 5")
estimate = c(1.057, 1, 1.386, 1.190, 1.454, 1.465,1)
conf.low = c(0.9858, 1, 1.2439, 1.0518, 1.2870, 1.2903, 1)
conf.high = c(1.134, 1, 1.544, 1.346, 1.643, 1.664, 1)
plot_data = data.frame(term, estimate, conf.low, conf.high)

pdf("forest.pdf", 10, 5)
```

```
#forest plot
dwplot(plot_data, vline = geom_vline(xintercept = 1, colour = "grey60", linetype = 2),
      dot_args = list(color = "#F8766D"), # color for the dot
      whisker_args = list(color = "Grey47") #color for the whisker
      ) +
  ggtitle("Multivariable Adjusted Hazard Ratios for Dementia") +
  theme_bw() +
  xlim(0.75, 1.75) +
  theme(plot.margin = unit(c(0,5,1,1), "cm")) +
  xlab("Hazard Ratio (95% CI)") +
  geom_text(x = 1.82,
            y = 7, label ="1.06 (0.99, 1.13); p = 0.12",
            hjust = 0,
            size = 3) +
  geom_text(x = 1.82,
            y = 6, label ="Reference",
            hjust = 0,
            size = 3) +
  geom_text(x = 1.82,
            y = 5, label ="1.39 (1.24, 1.54); p = <0.001",
            hjust = 0,
            size = 3) +
  geom_text(x = 1.82,
            y = 4, label ="1.19 (1.05, 1.35); p = 0.006",
            hjust = 0,
            size = 3) +
  geom_text(x = 1.82,
            y = 3, label ="1.45 (1.29, 1.64); p = <0.001",
            hjust = 0,
            size = 3) +
  geom_text(x = 1.82,
            y = 2, label ="1.47 (1.29, 1.66); p = <0.001",
            hjust = 0,
            size = 3) +
  geom_text(x = 1.82,
            y = 1, label ="Reference",
            hjust = 0,
            size = 3) +
  coord_cartesian(clip = "off")
dev.off()
```

List of References

Ajnakina, O., Agbedjro, D., Lally, J., Forti, M. D., Trotta, A., Mondelli, V., Pariante, C., Dazzan, P., Gaughran, F., Fisher, H. L., David, A., Murray, R. M., & Stahl, D. (2020). Predicting onset of early- and late-treatment resistance in first-episode schizophrenia patients using advanced shrinkage statistical methods in a small sample. *Psychiatry Res*, *294*, 113527. https://doi.org/10.1016/j.psychres.2020.113527

Alzheimer's Society. (2020). *How much does dementia care cost?* https://www.alzheimers.org.uk/blog/how-much-does-dementia-care-cost

Andreasen, N. C., Carpenter, W. T., Jr., Kane, J. M., Lasser, R. A., Marder, S. R., & Weinberger, D. R. (2005). Remission in schizophrenia: proposed criteria and rationale for consensus. *Am J Psychiatry*, *162*(3), 441-449. https://doi.org/10.1176/appi.ajp.162.3.441

Andreasson, S., Allebeck, P., Engstrom, A., & Rydberg, U. (1987). Cannabis and schizophrenia. A longitudinal study of Swedish conscripts. *Lancet*, *2*(8574), 1483-1486. https://doi.org/10.1016/s0140-6736(87)92620-1

Andrew, A., Knapp, M., McCrone, P., Parsonage, M., & Trachtenberg, M. (2012). *Effective Interventions in schizophrenia: the economic case*.

Austin, P. C., Lee, D. S., & Fine, J. P. (2016). Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation*, *133*(6), 601-609. https://doi.org/10.1161/CIRCULATIONAHA.115.017719

Barnett, J. H., Werners, U., Secher, S. M., Hill, K. E., Brazil, R., Masson, K., Pernet, D. E., Kirkbride, J. B., Murray, G. K., Bullmore, E. T., & Jones, P. B. (2007). Substance use in a population-based clinic sample of people with first-episode psychosis. *British Journal of Psychiatry*, *190*(6), 515-520. https://doi.org/10.1192/bjp.bp.106.024448

Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. *Jama-Journal of the American Medical Association*, *319*(13), 1317-1318. https://doi.org/10.1001/jama.2017.18391

Belsey, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley. https://doi.org/10.1002/0471725153

Benchimol, E. I., Smeeth, L., Guttmann, A., Harron, K., Moher, D., Petersen, I., Sorensen, H. T., von Elm, E., Langan, S. M., & Committee, R. W. (2015). The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med*, *12*(10), e1001885. https://doi.org/10.1371/journal.pmed.1001885

Bergink, V., Gibney, S. M., & Drexhage, H. A. (2014). Autoimmunity, inflammation, and psychosis: a search for peripheral markers. *Biol Psychiatry*, *75*(4), 324-331. https://doi.org/10.1016/j.biopsych.2013.09.037

Bertelsen, M., Jeppesen, P., Petersen, L., Thorup, A., Ohlenschlaeger, J., le Quach, P., Christensen, T. O., Krarup, G., Jorgensen, P., & Nordentoft, M. (2008). Five-year follow-up of a randomized multicenter trial of intensive early intervention vs standard treatment for patients with a first episode of psychotic illness: the OPUS trial. *Arch Gen Psychiatry*, *65*(7), 762-771. https://doi.org/10.1001/archpsyc.65.7.762

Bhattacharyya, S., Schoeler, T., Patel, R., di Forti, M., Murray, R. M., & McGuire, P. (2021). Individualized prediction of 2-year risk of relapse as indexed by psychiatric hospitalization following psychosis onset: Model development in two first episode samples. *Schizophr Res*, *228*, 483-492. https://doi.org/10.1016/j.schres.2020.09.016

Birchwood, M., Lester, H., McCarthy, L., Jones, P., Fowler, D., Amos, T., Freemantle, N., Sharma, V., Lavis, A., Singh, S., & Marshall, M. (2014). The UK national evaluation of the development and impact of Early Intervention Services (the National EDEN studies): study rationale, design and baseline characteristics. *Early Interv Psychiatry*, *8*(1), 59-67. https://doi.org/10.1111/eip.12007

Birchwood, M., Todd, P., & Jackson, C. (1998). Early intervention in psychosis - The critical period hypothesis. *British Journal of Psychiatry*, *172*, 53-59. https://doi.org/Doi 10.1192/S0007125000297663

Bird, V., Premkumar, P., Kendall, T., Whittington, C., Mitchell, J., & Kuipers, E. (2010). Early intervention services, cognitive–behavioural therapy and family intervention in early psychosis: systematic review. *British Journal of Psychiatry*, *197*(5), 350-356. https://doi.org/10.1192/bjp.bp.109.074526

Bone, C., Simmonds-Buckley, M., Thwaites, R., Sandford, D., Merzhvynska, M., Rubel, J., Deisenhofer, A. K., Lutz, W., & Delgadillo, J. (2021). Dynamic prediction of psychological treatment outcomes: development and validation of a prediction model using routinely collected symptom data. *Lancet Digital Health*, *3*(4), e231-e240. https://doi.org/10.1016/S2589-7500(21)00018-2

Breitborde, N. J. K., Srihari, V. H., & Woods, S. W. (2009). Review of the operational definition for first-episode psychosis. *Early Intervention in Psychiatry*, *3*(4), 259-265. https://doi.org/10.1111/j.1751-7893.2009.00148.x

Brown, A. S., & Patterson, P. H. (2011). Maternal infection and schizophrenia: implications for prevention. *Schizophr Bull*, *37*(2), 284-290. https://doi.org/10.1093/schbul/sbq146

Cadar, D., Lassale, C., Davies, H., Llewellyn, D. J., Batty, G. D., & Steptoe, A. (2018). Individual and Area-Based Socioeconomic Factors Associated With Dementia Incidence in England: Evidence From a 12-Year Follow-up in the English Longitudinal Study of Ageing. *JAMA Psychiatry*, *75*(7), 723-732. https://doi.org/10.1001/jamapsychiatry.2018.1012

Cakici, N., Sutterland, A. L., Penninx, B., Dalm, V. A., de Haan, L., & van Beveren, N. J. M. (2020). Altered peripheral blood compounds in drug-naive first-episode patients with either schizophrenia or major depressive disorder: a meta-analysis. *Brain Behav Immun*, *88*, 547-558. https://doi.org/10.1016/j.bbi.2020.04.039

Carvalho, A. F., Solmi, M., Sanches, M., Machado, M. O., Stubbs, B., Ajnakina, O., Sherman, C., Sun, Y. R., Liu, C. S., Brunoni, A. R., Pigato, G., Fernandes, B. S., Bortolato, B., Husain, M. I., Dragioti, E., Firth, J., Cosco, T. D., Maes, M., Berk, M., . . . Herrmann, N. (2020). Evidence-based umbrella review of 162 peripheral biomarkers for major mental disorders. *Transl Psychiatry*, *10*(1), 152. https://doi.org/10.1038/s41398-020-0835-5

Catalan, A., Richter, A., Salazar de Pablo, G., Vaquerizo-Serrano, J., Mancebo, G., Pedruzo, B., Aymerich, C., Solmi, M., González-Torres, M. Á., Gil, P., McGuire, P., & Fusar-Poli, P. (2021). Proportion and predictors of remission and recovery in first-episode psychosis: Systematic review and meta-analysis. *European Psychiatry*, *64*(1), e69, Article e69. https://doi.org/10.1192/j.eurpsy.2021.2246

Chavan, S. S., Pavlov, V. A., & Tracey, K. J. (2017). Mechanisms and Therapeutic Relevance of Neuro-immune Communication. *Immunity*, *46*(6), 927-942. https://doi.org/10.1016/j.immuni.2017.06.008

Chen, J. H., & Asch, S. M. (2017). Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl J Med*, *376*(26), 2507-2509. https://doi.org/10.1056/NEJMp1702071

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12-22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Chua, Y. C., Abdin, E., Tang, C., Subramaniam, M., & Verma, S. (2019). First-episode psychosis and vocational outcomes: A predictive model. *Schizophr Res*, *211*, 63-68. https://doi.org/10.1016/j.schres.2019.07.009

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*, *350*, g7594. https://doi.org/10.1136/bmj.g7594

Cook, C. E., & Thigpen, C. A. (2019). Five good reasons to be disappointed with randomized trials. *J Man Manip Ther*, *27*(2), 63-65. https://doi.org/10.1080/10669817.2019.1589697

Correll, C. U., Galling, B., Pawar, A., Krivko, A., Bonetto, C., Ruggeri, M., Craig, T. J., Nordentoft, M., Srihari, V. H., Guloksuz, S., Hui, C. L. M., Chen, E. Y. H., Valencia, M., Juarez, F., Robinson, D. G., Schooler, N. R., Brunette, M. F., Mueser, K. T., Rosenheck, R. A., . . . Kane, J. M. (2018). Comparison of Early Intervention Services vs Treatment as Usual for Early-Phase Psychosis: A Systematic Review, Meta-analysis, and Meta-regression. *JAMA Psychiatry*, *75*(6), 555-565. https://doi.org/10.1001/jamapsychiatry.2018.0623

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, *45*(3-4), 562-565. https://doi.org/10.1093/biomet/45.3-4.562

Cunningham, C. (2011). Systemic inflammation and delirium: important co-factors in the progression of dementia. *Biochem Soc Trans*, *39*(4), 945-953. https://doi.org/10.1042/BST0390945

Darcy, A. M., Louie, A. K., & Roberts, L. W. (2016). Machine Learning and the Profession of Medicine. *JAMA*, *315*(6), 551-552. https://doi.org/10.1001/jama.2015.18421

Davis, D. H., Muniz Terrera, G., Keage, H., Rahkonen, T., Oinas, M., Matthews, F. E., Cunningham, C., Polvikoski, T., Sulkava, R., MacLullich, A. M., & Brayne, C. (2012). Delirium is a strong risk factor for dementia in the oldest-old: a population-based cohort study. *Brain*, *135*(Pt 9), 2809-2816. https://doi.org/10.1093/brain/aws190

Davis, S. E., Greevy, R. A., Jr., Lasko, T. A., Walsh, C. G., & Matheny, M. E. (2020). Detection of calibration drift in clinical prediction models to inform model updating. *J Biomed Inform*, *112*, 103611. https://doi.org/10.1016/j.jbi.2020.103611

de Hond, A. A. H., Leeuwenberg, A. M., Hooft, L., Kant, I. M. J., Nijman, S. W. J., van Os, H. J. A., Aardoom, J. J., Debray, T. P. A., Schuit, E., van Smeden,

M., Reitsma, J. B., Steyerberg, E. W., Chavannes, N. H., & Moons, K. G. M. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digital Medicine*, *5*(1), 2. https://doi.org/10.1038/s41746-021-00549-7

de Nijs, J. (2019). *The outcome of psychosis* Utrecht University]. https://dspace.library.uu.nl/bitstream/1874/376436/1/22_01_3_jessica_de_nijs_compleet_final.pdf

de Nijs, J., Burger, T. J., Janssen, R. J., Kia, S. M., van Opstal, D. P. J., de Koning, M. B., de Haan, L., Alizadeh, B. Z., Bartels-Velthuis, A. A., van Beveren, N. J., Bruggeman, R., de Haan, L., Delespaul, P., Luykx, J. J., Myin-Germeys, I., Kahn, R. S., Schirmbeck, F., Simons, C. J. P., van Amelsvoort, T., . . . investigators, G. (2021). Individualized prediction of three- and six-year outcomes of psychosis in a longitudinal multicenter study: a machine learning approach. *npj Schizophrenia*, *7*(1), 34. https://doi.org/10.1038/s41537-021-00162-3

Demjaha, A., Lappin, J. M., Stahl, D., Patel, M. X., MacCabe, J. H., Howes, O. D., Heslin, M., Reininghaus, U. A., Donoghue, K., Lomas, B., Charalambides, M., Onyejiaka, A., Fearon, P., Jones, P., Doody, G., Morgan, C., Dazzan, P., & Murray, R. M. (2017). Antipsychotic treatment resistance in first-episode psychosis: prevalence, subtypes and predictors. *Psychol Med*, *47*(11), 1981-1989. https://doi.org/10.1017/S0033291717000435

Derks, E. M., Fleischhacker, W. W., Boter, H., Peuskens, J., Kahn, R. S., & Group, E. S. (2010). Antipsychotic drug treatment in first-episode psychosis: should patients be switched to a different antipsychotic drug after 2, 4, or 6 weeks of nonresponse? *J Clin Psychopharmacol*, *30*(2), 176-180. https://doi.org/10.1097/JCP.0b013e3181d2193c

Diana O. Perkins, M.D., M.P.H. ,, Hongbin Gu, Ph.D. ,, Kalina Boteva, M.D. , and, & Jeffrey A. Lieberman, M.D. (2005). Relationship Between Duration of Untreated Psychosis and Outcome in First-Episode Schizophrenia: A Critical Review and Meta-Analysis. *American Journal of Psychiatry*, *162*(10), 1785-1804. https://doi.org/10.1176/appi.ajp.162.10.1785

Díaz-Caneja, C. M., Pina-Camacho, L., Rodríguez-Quiroga, A., Fraguas, D., Parellada, M., & Arango, C. (2015). Predictors of outcome in early-onset psychosis: a systematic review. *npj Schizophrenia*, *1*(1), 14005. https://doi.org/10.1038/npjschz.2014.5

Drake, R. J., Husain, N., Marshall, M., Lewis, S. W., Tomenson, B., Chaudhry, I. B., Everard, L., Singh, S., Freemantle, N., Fowler, D., Jones, P. B., Amos, T., Sharma, V., Green, C. D., Fisher, H., Murray, R. M., Wykes, T., Buchan, I., & Birchwood, M. (2020). Effect of delaying treatment of first-episode psychosis on symptoms and social outcomes: a longitudinal analysis and modelling study. *Lancet Psychiatry*, *7*(7), 602-610. https://doi.org/10.1016/S2215-0366(20)30147-4

Dwyer, D., & Krishnadas, R. (2022). Five points to consider when reading a translational machine-learning paper. *Br J Psychiatry*, *220*(4), 169-171. https://doi.org/10.1192/bjp.2022.29

Eekhout, I., van de Wiel, M. A., & Heymans, M. W. (2017). Methods for significance testing of categorical covariates in logistic regression models after multiple imputation: power and applicability analysis. *BMC Medical Research Methodology*, *17*(1), 129. https://doi.org/10.1186/s12874-017-0404-7

Farooq, S., Choudry, A., Cohen, D., Naeem, F., & Ayub, M. (2019). Barriers to using clozapine in treatment-resistant schizophrenia: systematic review. *Bjpsych Bulletin*, *43*(1), 8-16. https://doi.org/10.1192/bjb.2018.67

Farooq, S., Large, M., Nielssen, O., & Waheed, W. (2009). The relationship between the duration of untreated psychosis and outcome in low-and-middle income countries: a systematic review and meta analysis. *Schizophrenia Research*, *109*(1-3), 15-23. https://doi.org/10.1016/j.schres.2009.01.008

Flyckt, L., Mattsson, M., Edman, G., Carlsson, R., & Cullberg, J. (2006). Predicting 5-year outcome in first-episode psychosis: construction of a prognostic rating scale. *J Clin Psychiatry*, *67*(6), 916-924. https://doi.org/10.4088/jcp.v67n0608

Fong, T. G., Davis, D., Growdon, M. E., Albuquerque, A., & Inouye, S. K. (2015). The interface between delirium and dementia in elderly adults. *Lancet Neurol*, *14*(8), 823-832. https://doi.org/10.1016/S1474-4422(15)00101-5

Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *SIGKDD Explor. Newsl.*, *12*(1), 49–57. https://doi.org/10.1145/1882471.1882479

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*, *33*(1), 1-22. https://www.ncbi.nlm.nih.gov/pubmed/20808728

Gandal, M. J., Zhang, P., Hadjimichael, E., Walker, R. L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., Shieh, A. W., Haney, J., Parhami, S., Belmont, J., Kim, M., Moran Losada, P., Khan, Z., Mleczko, J., Xia, Y., . . . Geschwind, D. H. (2018). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science*, *362*(6420). https://doi.org/10.1126/science.aat8127

Goldberg, T. E., Chen, C., Wang, Y., Jung, E., Swanson, A., Ing, C., Garcia, P. S., Whittington, R. A., & Moitra, V. (2020). Association of Delirium With Long-term Cognitive Decline: A Meta-analysis. *JAMA Neurol*, *77*(11), 1373-1381. https://doi.org/10.1001/jamaneurol.2020.2273

Goldsmith, D. R., Rapaport, M. H., & Miller, B. J. (2016). A meta-analysis of blood cytokine network alterations in psychiatric patients: comparisons between schizophrenia, bipolar disorder and depression. *Mol Psychiatry*, *21*(12), 1696-1709. https://doi.org/10.1038/mp.2016.3

Gonzalez-Blanch, C., Perez-Iglesias, R., Pardo-Garcia, G., Rodriguez-Sanchez, J. M., Martinez-Garcia, O., Vazquez-Barquero, J. L., & Crespo-Facorro, B. (2010). Prognostic value of cognitive functioning for global functional recovery in first-episode schizophrenia. *Psychol Med*, *40*(6), 935-944. https://doi.org/10.1017/S0033291709991267

Gray, B. (2020). *cmprsk: Subdistribution Analysis of Competing Risks*. In https://CRAN.R-project.org/package=cmprsk

Hall, C. B., Verghese, J., Sliwinski, M., Chen, Z., Katz, M., Derby, C., & Lipton, R. B. (2005). Dementia incidence may increase more slowly after age 90: results from the Bronx Aging Study. *Neurology*, *65*(6), 882-886. https://doi.org/10.1212/01.wnl.0000176053.98907.3f

Hansen, H. G., Starzer, M., Nilsson, S. F., Hjorthøj, C., Albert, N., & Nordentoft, M. (2023). Clinical Recovery and Long-Term Association of Specialized Early Intervention Services vs Treatment as Usual Among Individuals With First-Episode Schizophrenia Spectrum Disorder: 20-Year Follow-up of the OPUS Trial. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2022.5164

Harrell, F. E., Jr. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (2 ed.). Springer International Publishing. https://doi.org/2015.10.1007/978-3-319-19425-7

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, *15 4*, 361-387.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference and prediction. In (2 ed., pp. 245-247). Springer. hastie.su.domains/ElemStatLearn

Healthcare Improvement Scotland. (2014). *Improving the care for older people delirium toolkit think delirium*. https://www.healthcareimprovementscotland.org/

Hemingway, H., Croft, P., Perel, P., Hayden, J. A., Abrams, K., Timmis, A., Briggs, A., Udumyan, R., Moons, K. G., Steyerberg, E. W., Roberts, I., Schroter, S., Altman, D. G., Riley, R. D., & Group, P. (2013). Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ*, *346*, e5595. https://doi.org/10.1136/bmj.e5595

Hilker, R., Helenius, D., Fagerlund, B., Skytthe, A., Christensen, K., Werge, T. M., Nordentoft, M., & Glenthoj, B. (2018). Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. *Biol Psychiatry*, *83*(6), 492-498. https://doi.org/10.1016/j.biopsych.2017.08.017

Hingorani, A. D., Windt, D. A., Riley, R. D., Abrams, K., Moons, K. G., Steyerberg, E. W., Schroter, S., Sauerbrei, W., Altman, D. G., Hemingway, H., & Group, P. (2013). Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ*, *346*, e5793. https://doi.org/10.1136/bmj.e5793

Hippisley-Cox, J., Coupland, C., & Brindle, P. (2017). Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *Bmj-British Medical Journal*, *357*. https://doi.org/10.1136/bmj.j2099

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., Minhas, R., Sheikh, A., & Brindle, P. (2008). Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*, *336*(7659), 1475-1482. https://doi.org/10.1136/bmj.39609.449676.25

Hoeijmakers, L., Heinen, Y., van Dam, A. M., Lucassen, P. J., & Korosi, A. (2016). Microglial Priming and Alzheimer's Disease: A Possible Role for (Early) Immune Challenges and Epigenetics? *Front Hum Neurosci*, *10*, 398. https://doi.org/10.3389/fnhum.2016.00398

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651-674. https://doi.org/10.1198/106186006X133933

Howes, O., McCutcheon, R., & Stone, J. (2015). Glutamate and dopamine in schizophrenia: An update for the 21st century. *Journal of Psychopharmacology*, *29*(2), 97-115. https://doi.org/10.1177/0269881114563634

Howes, O. D., Vergunst, F., Gee, S., McGuire, P., Kapur, S., & Taylor, D. (2012). Adherence to treatment guidelines in clinical practice: study of antipsychotic

treatment prior to clozapine initiation. *Br J Psychiatry*, *201*(6), 481-485. https://doi.org/10.1192/bjp.bp.111.105833

Howes, O. D., Whitehurst, T., Shatalina, E., Townsend, L., Onwordi, E. C., Mak, T. L. A., Arumuham, A., O'Brien, O., Lobo, M., Vano, L., Zahid, U., Butler, E., & Osugo, M. (2021). The clinical significance of duration of untreated psychosis: an umbrella review and random-effects meta-analysis. *World Psychiatry*, *20*(1), 75-95. https://doi.org/10.1002/wps.20822

Huang, Y., Li, W., Macheret, F., Gabriel, R. A., & Ohno-Machado, L. (2020). A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc*, *27*(4), 621-633. https://doi.org/10.1093/jamia/ocz228

Iadecola, C. (2013). The pathobiology of vascular dementia. *Neuron*, *80*(4), 844-866. https://doi.org/10.1016/j.neuron.2013.10.008

Inouye, S. K., Bogardus, S. T., Jr., Charpentier, P. A., Leo-Summers, L., Acampora, D., Holford, T. R., & Cooney, L. M., Jr. (1999). A multicomponent intervention to prevent delirium in hospitalized older patients. *N Engl J Med*, *340*(9), 669-676. https://doi.org/10.1056/NEJM199903043400901

Inouye, S. K., Westendorp, R. G., & Saczynski, J. S. (2014). Delirium in elderly people. *Lancet*, *383*(9920), 911-922. https://doi.org/10.1016/S0140-6736(13)60688-1

Institute for Health Metrics and Evaluation (IHME). (2020). *GBD Compare Data Visualization*. IHME, University of Washington. http://vizhub.healthdata.org/gbd-compare

Iyer, S. N., Mangala, R., Anitha, J., Thara, R., & Malla, A. K. (2011). An examination of patient-identified goals for treatment in a first-episode programme in Chennai, India. *Early Interv Psychiatry*, *5*(4), 360-365. https://doi.org/10.1111/j.1751-7893.2011.00289.x

Jaaskelainen, E., Juola, P., Hirvonen, N., McGrath, J. J., Saha, S., Isohanni, M., Veijola, J., & Miettunen, J. (2013). A systematic review and meta-analysis of recovery in schizophrenia. *Schizophr Bull*, *39*(6), 1296-1306. https://doi.org/10.1093/schbul/sbs130

Jackson, T. A., Gladman, J. R., Harwood, R. H., MacLullich, A. M., Sampson, E. L., Sheehan, B., & Davis, D. H. (2017). Challenges and opportunities in understanding dementia and delirium in the acute hospital. *PLoS Med*, *14*(3), e1002247. https://doi.org/10.1371/journal.pmed.1002247

Jaffe, A. E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T. M., Weinberger, D. R., & Kleinman, J. E. (2016). Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat Neurosci*, *19*(1), 40-47. https://doi.org/10.1038/nn.4181

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer. https://doi.org/https://doi.org/10.1007/978-1-0716-1418-1

Jauhar, S., Nour, M. M., Veronese, M., Rogdaki, M., Bonoldi, I., Azis, M., Turkheimer, F., McGuire, P., Young, A. H., & Howes, O. D. (2017). A Test of the Transdiagnostic Dopamine Hypothesis of Psychosis Using Positron Emission Tomographic Imaging in Bipolar Affective Disorder and Schizophrenia. *JAMA Psychiatry*, *74*(12), 1206-1213. https://doi.org/10.1001/jamapsychiatry.2017.2943

Jauhar, S., Veronese, M., Nour, M. M., Rogdaki, M., Hathway, P., Turkheimer, F. E., Stone, J., Egerton, A., McGuire, P., Kapur, S., & Howes, O. D. (2019). Determinants of treatment response in first-episode psychosis: an (18)F-DOPA PET study. *Mol Psychiatry*, *24*(10), 1502-1512. https://doi.org/10.1038/s41380-018-0042-4

Jenkins, D. A., Martin, G. P., Sperrin, M., Riley, R. D., Debray, T. P. A., Collins, G. S., & Peek, N. (2021). Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn Progn Res*, *5*(1), 1. https://doi.org/10.1186/s41512-020-00090-3

Jenkins, D. A., Sperrin, M., Martin, G. P., & Peek, N. (2018). Dynamic models to predict health outcomes: current status and methodological challenges. *Diagn Progn Res*, *2*, 23. https://doi.org/10.1186/s41512-018-0045-2

Jia, L., Du, Y., Chu, L., Zhang, Z., Li, F., Lyu, D., Li, Y., Li, Y., Zhu, M., Jiao, H., Song, Y., Shi, Y., Zhang, H., Gong, M., Wei, C., Tang, Y., Fang, B., Guo, D., Wang, F., . . . Group, C. (2020). Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: a cross-sectional study. *Lancet Public Health*, *5*(12), e661-e671. https://doi.org/10.1016/S2468-2667(20)30185-7

Jones, P. B. (2013). Adult mental health disorders and their age at onset. *The British Journal of Psychiatry*, *202*(s54), s5-s10. https://doi.org/10.1192/bjp.bp.112.119164

Jongsma, H. E., Gayer-Anderson, C., Lasalvia, A., Quattrone, D., Mule, A., Szoke, A., Selten, J. P., Turner, C., Arango, C., Tarricone, I., Berardi, D., Tortelli, A., Llorca, P. M., de Haan, L., Bobes, J., Bernardo, M., Sanjuan, J., Santos, J. L., Arrojo, M., . . . European Network of National Schizophrenia Networks Studying Gene-Environment Interactions Work Package, G. (2018). Treated Incidence of Psychotic Disorders in the Multinational EU-GEI Study. *JAMA Psychiatry*, *75*(1), 36-46. https://doi.org/10.1001/jamapsychiatry.2017.3554

Jordan, G., Lutgens, D., Joober, R., Lepage, M., Iyer, S. N., & Malla, A. (2014). The relative contribution of cognition and symptomatic remission to functional outcome following treatment of a first episode of psychosis. *J Clin Psychiatry*, *75*(6), e566-572. https://doi.org/10.4088/JCP.13m08606

Kane, J. M., Kishimoto, T., & Correll, C. U. (2013). Non-adherence to medication in patients with psychotic disorders: epidemiology, contributing factors and management strategies. *World Psychiatry*, *12*(3), 216-226. https://doi.org/10.1002/wps.20060

Kappen, T. H., van Klei, W. A., van Wolfswinkel, L., Kalkman, C. J., Vergouwe, Y., & Moons, K. G. M. (2018). Evaluating the impact of prediction models:

lessons learned, challenges, and recommendations. *Diagn Progn Res*, *2*, 11. https://doi.org/10.1186/s41512-018-0033-6

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, *11*(9), 1 - 20. https://doi.org/10.18637/jss.v011.i09

Karhade, A. V., Thio, Q., Ogink, P., Kim, J., Lozano-Calderon, S., Raskin, K., & Schwab, J. H. (2018). Development of Machine Learning Algorithms for Prediction of 5-Year Spinal Chordoma Survival. *World Neurosurg*, *119*, e842-e847. https://doi.org/10.1016/j.wneu.2018.07.276

Kassambara, A., Kosinski, M., Biecek, P., & Fabian, S. (2021). *survminer: Drawing Survival Curves using 'ggplot2'*. In https://CRAN.R-project.org/package=survminer

Khachaturian, A. S., Hayden, K. M., Devlin, J. W., Fleisher, L. A., Lock, S. L., Cunningham, C., Oh, E. S., Fong, T. G., Fick, D. M., Marcantonio, E. R., Iyengar, V., Rockwood, K., Kuchel, G. A., Eckenhoff, R. G., MacLullich, A. M. J., Jones, R. N., Davis, D., D'Antonio, P. M., Fargo, K. N., . . . Inouye, S. K. (2020). International drive to illuminate delirium: A developing public health blueprint for action. *Alzheimers Dement*, *16*(5), 711-725. https://doi.org/10.1002/alz.12075

Kinney, J. W., Bemiller, S. M., Murtishaw, A. S., Leisgang, A. M., Salazar, A. M., & Lamb, B. T. (2018). Inflammation as a central mechanism in Alzheimer's disease. *Alzheimers Dement (N Y)*, *4*, 575-590. https://doi.org/10.1016/j.trci.2018.06.014

Kirkbride, J. B., Errazuriz, A., Croudace, T. J., Morgan, C., Jackson, D., Boydell, J., Murray, R. M., & Jones, P. B. (2012). Incidence of schizophrenia and other psychoses in England, 1950-2009: a systematic review and meta-analyses. *PLoS One*, *7*(3), e31660. https://doi.org/10.1371/journal.pone.0031660

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., Derks, E. M., Fleischhacker, W. W., & Hasan, A. (2016). Multisite prediction

of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *Lancet Psychiatry*, *3*(10), 935-946. https://doi.org/10.1016/S2215-0366(16)30171-7

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

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1 - 26. https://doi.org/10.18637/jss.v028.i05

Lally, J., Ajnakina, O., Stubbs, B., Cullinane, M., Murphy, K. C., Gaughran, F., & Murray, R. M. (2017). Remission and recovery from first-episode psychosis in adults: systematic review and meta-analysis of long-term outcome studies. *Br J Psychiatry*, *211*(6), 350-358. https://doi.org/10.1192/bjp.bp.117.201475

Lane, N., & Broome, M. (2022). Towards personalised predictive psychiatry in clinical practice: an ethical perspective. *Br J Psychiatry*, 1-3. https://doi.org/10.1192/bjp.2022.37

Lau, B., Cole, S. R., & Gange, S. J. (2009). Competing risk regression models for epidemiologic data. *Am J Epidemiol*, *170*(2), 244-256. https://doi.org/10.1093/aje/kwp107

Lee, R., Leighton, S. P., Thomas, L., Gkoutos, G. V., Wood, S. J., Fenton, S.-J. H., Deligianni, F., Cavanagh, J., & Mallikarjun, P. K. (2022). Prediction models in first-episode psychosis: systematic review and critical appraisal. *The British Journal of Psychiatry*, *220*(4), 179-191. https://doi.org/10.1192/bjp.2021.219

Leighton, S. P., Krishnadas, R., Chung, K., Blair, A., Brown, S., Clark, S., Sowerbutts, K., Schwannauer, M., Cavanagh, J., & Gumley, A. I. (2019). Predicting one-year outcome in first episode psychosis using machine learning. *PLoS One*, *14*(3). https://doi.org/10.1371/journal.pone.0212846

Leighton, S. P., Krishnadas, R., Upthegrove, R., Marwaha, S., Steyerberg, E. W., Gkoutos, G. V., Broome, M. R., Liddle, P. F., Everard, L., Singh, S. P., Freemantle, N., Fowler, D., Jones, P. B., Sharma, V., Murray, R., Wykes, T., Drake, R. J., Buchan, I., Rogers, S., . . . Mallikarjun, P. K. (2021). Development and Validation of a Nonremission Risk Prediction Model in First-Episode Psychosis: An Analysis of 2 Longitudinal Studies. *Schizophr Bull Open*, *2*(1), sgab041. https://doi.org/10.1093/schizbullopen/sgab041

Leighton, S. P., Upthegrove, R., Krishnadas, R., Benros, M. E., Broome, M. R., Gkoutos, G. V., Liddle, P. F., Singh, S. P., Everard, L., Jones, P. B., Fowler, D., Sharma, V., Freemantle, N., Christensen, R. H. B., Albert, N., Nordentoft, M., Schwannauer, M., Cavanagh, J., Gumley, A. I., . . . Mallikarjun, P. K. (2019). Development and validation of multivariable prediction models of remission, recovery, and quality of life outcomes in people with first episode psychosis: a machine learning approach. *Lancet Digital Health*, *1*(6), E261-E270. https://doi.org/10.1016/S2589-7500(19)30121-9

Lenert, M. C., Matheny, M. E., & Walsh, C. G. (2019). Prognostic models will be victims of their own success, unless. *J Am Med Inform Assoc*, *26*(12), 1645-1650. https://doi.org/10.1093/jamia/ocz145

Lieberman, J. A., & First, M. B. (2018). Psychotic Disorders. *N Engl J Med*, *379*(3), 270-280. https://doi.org/10.1056/NEJMra1801490

Lieberman, J. A., Tollefson, G. D., Charles, C., Zipursky, R., Sharma, T., Kahn, R. S., Keefe, R. S., Green, A. I., Gur, R. E., McEvoy, J., Perkins, D., Hamer, R. M., Gu, H., Tohen, M., & Group, H. S. (2005). Antipsychotic drug effects on brain morphology in first-episode psychosis. *Arch Gen Psychiatry*, *62*(4), 361-370. https://doi.org/10.1001/archpsyc.62.4.361

Lindhiem, O., Petersen, I. T., Mentch, L. K., & Youngstrom, E. A. (2020). The Importance of Calibration in Clinical Psychology. *Assessment*, *27*(4), 840-854. https://doi.org/10.1177/1073191117752055

Majka, M. (2020). *_naivebayes: High Performance Implementation of the Naive Bayes Algorithm in R_. R package version 0.9.7.* In https://CRAN.R-project.org/package=naivebayes

Marcantonio, E. R., Flacker, J. M., Wright, R. J., & Resnick, N. M. (2001). Reducing delirium after hip fracture: a randomized trial. *J Am Geriatr Soc*, *49*(5), 516-522. https://doi.org/10.1046/j.1532-5415.2001.49108.x

Marshall, M., Lewis, S., Lockwood, A., Drake, R., Jones, P., & Croudace, T. (2005). Association between duration of untreated psychosis and outcome in cohorts of first-episode patients: a systematic review. *Arch Gen Psychiatry*, *62*(9), 975-983. https://doi.org/10.1001/archpsyc.62.9.975

Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med*, *3*(11), e442. https://doi.org/10.1371/journal.pmed.0030442

Matthew Drury (https://stats.stackexchange.com/users/74500/matthew-drury). (26th June 2019). *How to apply standardization/normalization to train- and testset if prediction is the goal?* https://stats.stackexchange.com/q/174865

Mazza, M. G., Lucchi, S., Rossetti, A., & Clerici, M. (2020). Neutrophil-lymphocyte ratio, monocyte-lymphocyte ratio and platelet-lymphocyte ratio in non-affective psychosis: A meta-analysis and systematic review. *World J Biol Psychiatry*, *21*(5), 326-338. https://doi.org/10.1080/15622975.2019.1583371

McCutcheon, R. A., Krystal, J. H., & Howes, O. D. (2020). Dopamine and glutamate in schizophrenia: biology, symptoms and treatment. *World Psychiatry*, *19*(1), 15-33. https://doi.org/10.1002/wps.20693

Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., & Danese, A. (2022). Clinical prediction models in psychiatry: a systematic

review of two decades of progress and challenges. *Mol Psychiatry*, *27*(6), 2700-2708. https://doi.org/10.1038/s41380-022-01528-4

Merritt, K., Egerton, A., Kempton, M. J., Taylor, M. J., & McGuire, P. K. (2016). Nature of Glutamate Alterations in Schizophrenia: A Meta-analysis of Proton Magnetic Resonance Spectroscopy Studies. *JAMA Psychiatry*, *73*(7), 665-674. https://doi.org/10.1001/jamapsychiatry.2016.0442

Merritt, K., McGuire, P. K., Egerton, A., & Investigators, H.-M. S. (2021). Association of Age, Antipsychotic Medication, and Symptom Severity in Schizophrenia With Proton Magnetic Resonance Spectroscopy Brain Glutamate Level A Mega-analysis of Individual Participant-Level Data. *JAMA Psychiatry*, *78*(6), 667-681. https://doi.org/10.1001/jamapsychiatry.2021.0380

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel , A., & Leisch, F. (2021). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-7.* In https://CRAN.R-project.org/package=e1071

Miech, R. A., Breitner, J. C., Zandi, P. P., Khachaturian, A. S., Anthony, J. C., & Mayer, L. (2002). Incidence of AD may decline in the early 90s for men, later for women: The Cache County study. *Neurology*, *58*(2), 209-218. https://doi.org/10.1212/wnl.58.2.209

Miller, B. J., Buckley, P., Seabolt, W., Mellor, A., & Kirkpatrick, B. (2011). Meta-analysis of cytokine alterations in schizophrenia: clinical status and antipsychotic effects. *Biol Psychiatry*, *70*(7), 663-671. https://doi.org/10.1016/j.biopsych.2011.04.013

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*, *162*(1), W1-73. https://doi.org/10.7326/M14-0698

Moons, K. G., de Groot, J. A., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., & Collins, G. S. (2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*, *11*(10), e1001744. https://doi.org/10.1371/journal.pmed.1001744

Moons, K. G., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G., & Woodward, M. (2012). Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*, *98*(9), 691-698. https://doi.org/10.1136/heartjnl-2011-301247

Moons, K. G. M., Altman, D. G., Vergouwe, Y., & Royston, P. (2009). Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *British Medical Journal*, *338*. https://doi.org/10.1136/bmj.b606

Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med*, *170*(1), W1-W33. https://doi.org/10.7326/M18-1377

Moreno-Kustner, B., Martin, C., & Pastor, L. (2018). Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses. *PLoS One*, *13*(4), e0195687. https://doi.org/10.1371/journal.pone.0195687

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

Nakamura, T., & Takata, A. (2023). The molecular pathology of schizophrenia: an overview of existing knowledge and new directions for future research. *Mol Psychiatry*, *28*(5), 1868-1889. https://doi.org/10.1038/s41380-023-02005-2

National Institute for Health and Care Excellence (NICE). (2014, 01 March 2014). *Psychosis and schizophrenia in adults: prevention and management. (Clinical Guideline CG178)*. NICE. https://www.nice.org.uk/guidance/cg178

National Institute for Health and Care Excellence (NICE). (2016a, 27 September 2016). *Cardiovascular disease: risk assessment and reduction, including lipid modification. (Clinical Guideline CG181)*. NICE. https://www.nice.org.uk/guidance/cg181/

National Institute for Health and Care Excellence (NICE). (2016b). *Implementing the Early Intervention in Psychosis Access and Waiting Time Standard: Guidance*. NICE. https://www.nice.org.uk/guidance/qs80/resources/implementing-the-early-intervention-in-psychosis-access-and-waiting-time-standard-guidance-2487749725

National Institute for Health and Care Excellence (NICE). (2018). *Early and locally advanced breast cancer: diagnosis and management. NICE guideline [NG101]*. NICE. https://www.nice.org.uk/guidance/ng101

National Institute for Health and Care Excellence (NICE). (2021a). *Dementia: What are the risk factors?* NICE. https://cks.nice.org.uk/topics/dementia/background-information/risk-factors/

National Institute for Health and Care Excellence (NICE). (2021b). *Psychosis and schizophrenia*. NICE. https://cks.nice.org.uk/topics/psychosis-schizophrenia/

NHS Greater Glasgow & Clyde. (2023). *West Of Scotland Safe Haven*. https://www.nhsggc.scot/hospitals-services/services-a-to-z/west-of-scotland-safe-haven/

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*, n160. https://doi.org/10.1136/bmj.n160

Penttilä, M., Jääskeläinen, E., Hirvonen, N., Isohanni, M., & Miettunen, J. (2014). Duration of untreated psychosis as predictor of long-term outcome in schizophrenia: systematic review and meta-analysis. *British Journal of Psychiatry*, *205*(2), 88-94. https://doi.org/10.1192/bjp.bp.113.127753

Pereira, J. V., Aung Thein, M. Z., Nitchingham, A., & Caplan, G. A. (2021). Delirium in older adults is associated with development of new dementia: a systematic review and meta-analysis. *Int J Geriatr Psychiatry*, *36*(7), 993-1003. https://doi.org/10.1002/gps.5508

Perlis, R. H. (2013). A Clinical Risk Stratification Tool for Predicting Treatment Resistance in Major Depressive Disorder. *Biological Psychiatry*, *74*(1), 7-14. https://doi.org/https://doi.org/10.1016/j.biopsych.2012.12.007

Petersen, L., Jeppesen, P., Thorup, A., Abel, M. B., Ohlenschlaeger, J., Christensen, T. O., Krarup, G., Jorgensen, P., & Nordentoft, M. (2005). A randomised multicentre trial of integrated versus standard treatment for patients with a first episode of psychotic illness. *BMJ*, *331*(7517), 602. https://doi.org/10.1136/bmj.38565.415000.E01

Posselt, C. M., Albert, N., Nordentoft, M., & Hjorthoj, C. (2021). The Danish OPUS Early Intervention Services for First-Episode Psychosis: A Phase 4 Prospective Cohort Study With Comparison of Randomized Trial and Real-World Data. *Am J Psychiatry*, *178*(10), 941-951. https://doi.org/10.1176/appi.ajp.2021.20111596

Prince, M., Knapp, M., Guerchet, M., McCrone, P., Prina, M., Comas-Herrera, A., Wittenberg, R., Adelaja, B., Hu, B., King, D., Rehil, A., & Salimkumar, D. (2014). *Dementia UK: Update Second Edition Report*. A. s. Society. http://eprints.lse.ac.uk/59437/1/Dementia_UK_Second_edition_-_Overview.pdf

Puntis, S., Whiting, D., Pappa, S., & Lennox, B. (2021). Development and external validation of an admission risk prediction model after treatment from early intervention in psychosis services. *Translational Psychiatry*, *11*(1). https://doi.org/10.1038/s41398-020-01172-y

Qiu, C., & Fratiglioni, L. (2018). Aging without Dementia is Achievable: Current Evidence from Epidemiological Research. *J Alzheimers Dis*, *62*(3), 933-942. https://doi.org/10.3233/JAD-171037

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing.

Riley, R. D., Hayden, J. A., Steyerberg, E. W., Moons, K. G., Abrams, K., Kyzas, P. A., Malats, N., Briggs, A., Schroter, S., Altman, D. G., Hemingway, H., & Group, P. (2013). Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med*, *10*(2), e1001380. https://doi.org/10.1371/journal.pmed.1001380

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell, F. E., Jr., Moons, K. G., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med*, *38*(7), 1276-1296. https://doi.org/10.1002/sim.7992

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H. L., Pers, T. H., Agartz, I., Agerbo, E., Albus, M., Alexander, M., Amin, F., Bacanu, S. A., Begemann, M., Belliveau, R. A., . . . Consor, W. T. C.-C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421-+. https://doi.org/10.1038/nature13595

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

Rosen, M., Betz, L. T., Schultze-Lutter, F., Chisholm, K., Haidl, T. K., Kambeitz-Ilankovic, L., Bertolino, A., Borgwardt, S., Brambilla, P., Lencer, R., Meisenzahl, E., Ruhrmann, S., Salokangas, R. K. R., Upthegrove, R., Wood, S. J.,

Koutsouleris, N., Kambeitz, J., & Consortium, P. (2021). Towards clinical application of prediction models for transition to psychosis: A systematic review and external validation study in the PRONIA sample. *Neurosci Biobehav Rev*, *125*, 478-492. https://doi.org/10.1016/j.neubiorev.2021.02.032

Royal College of Psychiatrists. (2019). *National Audit of Dementia Care in General Hospitals 2018-2019: Round Four Audit Report*. R. C. o. Psychiatrists.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons. https://doi.org/10.1002/9780470316696

Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., Baldwin, H., Danese, A., Fazel, S., Steyerberg, E. W., Stahl, D., & Fusar-Poli, P. (2021). Implementing Precision Psychiatry: A Systematic Review of Individualized Prediction Models for Clinical Practice. *Schizophr Bull*, *47*(2), 284-297. https://doi.org/10.1093/schbul/sbaa120

Santesteban-Echarri, O., Paino, M., Rice, S., Gonzalez-Blanch, C., McGorry, P., Gleeson, J., & Alvarez-Jimenez, M. (2017). Predictors of functional recovery in first-episode psychosis: A systematic review and meta-analysis of longitudinal studies. *Clin Psychol Rev*, *58*, 59-75. https://doi.org/10.1016/j.cpr.2017.09.007

Schunemann, H. J., Oxman, A. D., Brozek, J., Glasziou, P., Jaeschke, R., Vist, G. E., Williams, J. W., Kunz, R., Craig, J., Montori, V. M., Bossuyt, P., Guyatt, G. H., & Grp, G. W. (2008). GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *Bmj-British Medical Journal*, *336*(7653), 1106-1110. https://doi.org/10.1136/bmj.39500.677199.AE

Schwarcz, R., Bruno, J. P., Muchowski, P. J., & Wu, H. Q. (2012). Kynurenines in the mammalian brain: when physiology meets pathology. *Nat Rev Neurosci*, *13*(7), 465-477. https://doi.org/10.1038/nrn3257

Scottish Intercollegiate Guidelines Network (SIGN). (2013). *SIGN 131: Management of schizophrenia*. SIGN. https://www.sign.ac.uk/assets/sign131.pdf

Secher, R. G., Hjorthoj, C. R., Austin, S. F., Thorup, A., Jeppesen, P., Mors, O., & Nordentoft, M. (2015). Ten-year follow-up of the OPUS specialized early intervention trial for patients with a first episode of psychosis. *Schizophr Bull*, *41*(3), 617-626. https://doi.org/10.1093/schbul/sbu155

Sedgwick, P. (2015). Understanding the ecological fallacy. *BMJ : British Medical Journal*, *351*, h4773. https://doi.org/10.1136/bmj.h4773

Seeman, P., & Lee, T. (1975). Antipsychotic drugs: direct correlation between clinical potency and presynaptic action on dopamine neurons. *Science*, *188*(4194), 1217-1219. https://doi.org/10.1126/science.1145194

Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., Genovese, G., Rose, S. A., Handsaker, R. E., Schizophrenia Working Group of the Psychiatric Genomics, C., Daly, M. J., Carroll, M. C., Stevens, B., & McCarroll, S. A. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature*, *530*(7589), 177-183. https://doi.org/10.1038/nature16549

Shivashankar, S., Telfer, S., Arunagiriraj, J., McKinnon, M., Jauhar, S., Krishnadas, R., & McCreadie, R. (2013). Has the prevalence, clinical presentation and social functioning of schizophrenia changed over the last 25 years? Nithsdale schizophrenia survey revisited. *Schizophr Res*, *146*(1-3), 349-356. https://doi.org/10.1016/j.schres.2013.02.006

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289-310, 222. https://doi.org/10.1214/10-STS330

Singh, S. P., Javed, A., & Psychosis, W. P. A. E. I. A. P. f. E. I. i. (2020). Early intervention in psychosis in low- and middle-income countries: a WPA initiative. *World Psychiatry*, *19*(1), 122. https://doi.org/10.1002/wps.20708

Siskind, D., Orr, S., Sinha, S., Yu, O., Brijball, B., Warren, N., MacCabe, J. H., Smart, S. E., & Kisely, S. (2022). Rates of treatment-resistant schizophrenia from first-episode cohorts: systematic review and meta-analysis. *Br J Psychiatry*, *220*(3), 115-120. https://doi.org/10.1192/bjp.2021.61

Slooter, A. J. C., Otte, W. M., Devlin, J. W., Arora, R. C., Bleck, T. P., Claassen, J., Duprey, M. S., Ely, E. W., Kaplan, P. W., Latronico, N., Morandi, A., Neufeld, K. J., Sharshar, T., MacLullich, A. M. J., & Stevens, R. D. (2020). Updated nomenclature of delirium and acute encephalopathy: statement of ten Societies. *Intensive Care Med*, *46*(5), 1020-1022. https://doi.org/10.1007/s00134-019-05907-4

Smart, S. E., Kepinska, A. P., Murray, R. M., & MacCabe, J. H. (2021). Predictors of treatment resistant schizophrenia: a systematic review of prospective observational studies. *Psychol Med*, *51*(1), 44-53. https://doi.org/10.1017/S0033291719002083

Snell, K. I., Ensor, J., Debray, T. P., Moons, K. G., & Riley, R. D. (2018). Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Statistical Methods in Medical Research*, *27*(11), 3505-3522. https://doi.org/10.1177/0962280217705678

Srireddy, P., Agnihotri, A., Park, J., Taylor, J., Connolly, M., & Krishnadas, R. (2012). Ethnicity, deprivation and psychosis: the Glasgow experience. *Epidemiol Psychiatr Sci*, *21*(3), 311-316. https://doi.org/10.1017/S2045796012000352

Stafford, M., & Marmot, M. (2003). Neighbourhood deprivation and health: does it affect us all equally? *Int J Epidemiol*, *32*(3), 357-366. https://doi.org/10.1093/ije/dyg084

Stauffer, V. L., Case, M., Kinon, B. J., Conley, R., Ascher-Svanum, H., Kollack-Walker, S., Kane, J., McEvoy, J., & Lieberman, J. (2011). Early response to antipsychotic therapy as a clinical marker of subsequent response in the treatment of patients with first-episode psychosis. *Psychiatry Research*, *187*(1), 42-48. https://doi.org/https://doi.org/10.1016/j.psychres.2010.11.017

Steyerberg, E. W. (2019). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* (2 ed.). Springer International Publishing.

Steyerberg, E. W., & Harrell, F. E., Jr. (2016). Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*, *69*, 245-247. https://doi.org/10.1016/j.jclinepi.2015.04.005

Steyerberg, E. W., Moons, K. G., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., Altman, D. G., & Group, P. (2013). Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*, *10*(2), e1001381. https://doi.org/10.1371/journal.pmed.1001381

Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*, *35*(29), 1925-1931. https://doi.org/10.1093/eurheartj/ehu207

Studerus, E., Ramyead, A., & Riecher-Rossler, A. (2017). Prediction of transition to psychosis in patients with a clinical high risk for psychosis: a systematic review of methodology and reporting. *Psychol Med*, *47*(7), 1163-1178. https://doi.org/10.1017/S0033291716003494

Sullivan, S., Northstone, K., Gadd, C., Walker, J., Margelyte, R., Richards, A., & Whiting, P. (2017). Models to predict relapse in psychosis: A systematic review. *PLoS One*, *12*(9), e0183998. https://doi.org/10.1371/journal.pone.0183998

Sullivan, T. R., Salter, A. B., Ryan, P., & Lee, K. J. (2015). Bias and Precision of the "Multiple Imputation, Then Deletion" Method for Dealing With Missing Outcome Data. *American Journal of Epidemiology*, *182*(6), 528-534. https://doi.org/10.1093/aje/kwv100

The Scottish Government. (2012). *Mental Health Strategy for Scotland: 2012-2015*. https://www.gov.scot/publications/mental-health-strategy-scotland-2012-2015/

The Scottish Government. (2019). *Early intervention in psychosis: action plan*. https://www.gov.scot/publications/vision-improve-early-intervention-psychosis-scotland/

Therneau, T. M. (2020). A package for survival analysis in R. https://CRAN.R-project.org/package=survival

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer.

Upthegrove, R., Manzanares-Teson, N., & Barnes, N. M. (2014). Cytokine function in medication-naive first episode psychosis: a systematic review and meta-analysis. *Schizophr Res*, *155*(1-3), 101-108. https://doi.org/10.1016/j.schres.2014.03.005

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1 - 67. https://doi.org/10.18637/jss.v045.i03

van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., Topic Group 'Evaluating diagnostic, t., & prediction models' of the, S. i. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Med*, *17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*, *74*, 167-176. https://doi.org/10.1016/j.jclinepi.2015.12.005

van der Gaag, M., Hoffman, T., Remijsen, M., Hijman, R., de Haan, L., van Meijel, B., van Harten, P. N., Valmaggia, L., de Hert, M., Cuijpers, A., & Wiersma, D. (2006). The five-factor model of the Positive and Negative Syndrome Scale II: A ten-fold cross-validation of a revised model. *Schizophrenia Research*, *85*(1), 280-287. https://doi.org/https://doi.org/10.1016/j.schres.2006.03.021

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*, *14*, 137. https://doi.org/10.1186/1471-2288-14-137

van der Ploeg, T., Nieboer, D., & Steyerberg, E. W. (2016). Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. *J Clin Epidemiol*, *78*, 83-89. https://doi.org/10.1016/j.jclinepi.2016.03.002

van Hoorde, K., van Huffel, S., Timmerman, D., Bourne, T., & van Calster, B. (2015). A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *J Biomed Inform*, *54*, 283-293. https://doi.org/10.1016/j.jbi.2014.12.016

van Houwelingen, J. C., & Le Cessie, S. (1990). Predictive value of statistical models. *Stat Med*, *9*(11), 1303-1325. https://doi.org/10.1002/sim.4780091109

van Os, J., Bak, M., Hanssen, M., Bijl, R. V., de Graaf, R., & Verdoux, H. (2002). Cannabis use and psychosis: a longitudinal population-based study. *Am J Epidemiol*, *156*(4), 319-327. https://doi.org/10.1093/aje/kwf043

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gotzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., Egger, M., & Initiative, S. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med*, *4*(10), e297. https://doi.org/10.1371/journal.pmed.0040297

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), 565-574. https://doi.org/10.1177/0272989x06295361

Vickers, A. J., van Calster, B., & Steyerberg, E. W. (2019). A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res*, *3*, 18. https://doi.org/10.1186/s41512-019-0064-7

von Hippel, P. T. (2007). Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data. *Sociological Methodology*, *37*(1), 83-117. https://doi.org/10.1111/j.1467-9531.2007.00180.x

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical applications of machine learning algorithms: beyond the black box. *BMJ*, *364*, l886. https://doi.org/10.1136/bmj.l886

Weibell, M. A., Hegelstad, W. T., Auestad, B., Bramness, J., Evensen, J., Haahr, U., Joa, I., Johannessen, J. O., Larsen, T. K., Melle, I., Opjordsmoen, S., Rund, B. R., Simonsen, E., Vaglum, P., McGlashan, T., McGorry, P., & Friis, S. (2017). The Effect of Substance Use on 10-Year Outcome in First-Episode Psychosis. *Schizophrenia Bulletin*, *43*(4), 843-851. https://doi.org/10.1093/schbul/sbw179

Wishart, G. C., Azzato, E. M., Greenberg, D. C., Rashbass, J., Kearins, O., Lawrence, G., Caldas, C., & Pharoah, P. D. P. (2010). PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer (vol 12, R1, 2010). *Breast Cancer Research*, *12*(2). https://doi.org/10.1186/bcr2480

Witlox, J., Eurelings, L. S., de Jonghe, J. F., Kalisvaart, K. J., Eikelenboom, P., & van Gool, W. A. (2010). Delirium in elderly patients and the risk of postdischarge mortality, institutionalization, and dementia: a meta-analysis. *JAMA*, *304*(4), 443-451. https://doi.org/10.1001/jama.2010.1013

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., Mallett, S., & Groupdagger, P. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med*, *170*(1), 51-58. https://doi.org/10.7326/M18-1376

World Health Organization (WHO). (1992). *The Tenth Revision of the International Classification of Diseases (ICD-10) Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. World Health Organization (WHO),.

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M. J., Dahly, D. L., Damen, J. A. A., Debray, T. P. A., de Jong, V. M. T., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Heus, P., Kammer, M., Kreuzberger, N., . . . van Smeden, M. (2020). Prediction models

for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*, *369*, m1328. https://doi.org/10.1136/bmj.m1328

Yaffe, K., Kanaya, A., Lindquist, K., Simonsick, E. M., Harris, T., Shorr, R. I., Tylavsky, F. A., & Newman, A. B. (2004). The metabolic syndrome, inflammation, and risk of cognitive decline. *JAMA*, *292*(18), 2237-2242. https://doi.org/10.1001/jama.292.18.2237

Yoshimura, B., Yada, Y., So, R., Takaki, M., & Yamada, N. (2017). The critical treatment window of clozapine in treatment-resistant schizophrenia: Secondary analysis of an observational study. *Psychiatry Res*, *250*, 65-70. https://doi.org/10.1016/j.psychres.2017.01.064

Yuan, N., Chen, Y., Xia, Y., Dai, J., & Liu, C. (2019). Inflammation-related biomarkers in major psychiatric disorders: a cross-disorder assessment of reproducibility and specificity in 43 meta-analyses. *Transl Psychiatry*, *9*(1), 233. https://doi.org/10.1038/s41398-019-0570-y

Yuen, H. P., Mackinnon, A., Hartmann, J., Amminger, G. P., Markulev, C., Lavoie, S., Schafer, M. R., Polari, A., Mossaheb, N., Schlogelhofer, M., Smesny, S., Hickie, I. B., Berger, G., Chen, E. Y. H., de Haan, L., Nieman, D. H., Nordentoft, M., Riecher-Rossler, A., Verma, S., . . . Nelson, B. (2018). Dynamic prediction of transition to psychosis using joint modelling. *Schizophr Res*, *202*, 333-340. https://doi.org/10.1016/j.schres.2018.07.002

Yuen, H. P., Mackinnon, A., & Nelson, B. (2020). Dynamic prediction systems of transition to psychosis using joint modelling: extensions to the base system. *Schizophr Res*, *216*, 207-212. https://doi.org/10.1016/j.schres.2019.11.059

Zammit, S., Allebeck, P., Andreasson, S., Lundberg, I., & Lewis, G. (2002). Self reported cannabis use as a risk factor for schizophrenia in Swedish conscripts of 1969: historical cohort study. *BMJ*, *325*(7374), 1199. https://doi.org/10.1136/bmj.325.7374.1199

Zou, K. H., Liu, A., Bandos, A. I., Ohno-Machado, L., & Rockette, H. E. (2012). Statitical Evaluation of Diagnostic Performance: Topics in ROC Analysis. In (pp. 204). CRC Press.