1.Background

Up to 35% of first-episode schizophrenia patients are estimated to meet criteria for treatment resistant schizophrenia (TRS) over the five years after illness onset (Lally et al., 2016). Emerging evidence demonstrates that around 70% of TRS patients did not respond to antipsychotic medications from the start of first treatment (i.e., early treatment resistant (E-TR) schizophrenia); whereas, the remaining 30% of TRS patients, broadly defined as late-treatment resistant (L-TR) schizophrenia, gradually transition to treatment resistance during the 5-year period, having initially responded to antipsychotic medications (Lally et al., 2016). Given TRS is a major cause of disability and is associated with high social and economic costs (Kennedy et al., 2014), having reliable prediction models for estimating an individual risk for E-TR and L-TR would advance our understanding of patients' risk of developing treatment resistance, especially after a period of favourable response to the ongoing treatment with antipsychotic medications.

Development of accurate prediction models for estimating individual, rather than average, risk for a disorder outcome during the illness (Steyerberg et al., 2010), based on available patient characteristics and clinical findings, is generally thought to require large datasets (Califf et al., 1997), mainly when the outcome of interest is a rare event. As they are often unavailable in schizophrenia research, stepwise selection techniques, particularly with the traditional *p*-value of 0.05, are frequently employed, which tend to provide readily interpretable models (Steyerberg et al., 2000). However, these stepwise selection techniques increase the risk of biased regression coefficients and overfitting (Derksen and Keselman, 1992) potentially leading to development of a prediction model with inaccurate predictions in new cases (Osborne et al., 2012; Steyerberg et al., 1999).

Computer intensive statistical learning methods, particularly regularised regression methods (RRMs), have been suggested as optimal methods for clinical and personalised risk prediction (Steyerberg, 2019), especially for small datasets (Steyerberg et al., 2000). Through an

introduction of penalty, RRMs produce a model with good interpretability and overcome problems of overfitting, multicollinearity and poor prediction of new cases (Hastie et al., 2009). As RRMs have not been applied to small datasets for estimating an individual risk of TRS subtypes, their usefulness for these important outcomes is unknown. Therefore, in this study, employing RRMs, we aimed to develop robust prediction models for estimating an individual risk for E-TR and L-TR in a sample of first-episode schizophrenia patients who were followed-up during the first 5 years of their illness (Ajnakina et al., 2017; Lally et al., 2016).

2.Methods

2.1.Sample

The study comprised 282 participants aged 18-65 meeting criteria for schizophrenia spectrum (FES) disorders (International Classification of Diseases, 10th-Revision (ICD-10) diagnoses: F20.0, F25.0, F28.0, F29.0)(World Health Organization, 1992), validated by administration of the Schedules for Clinical Assessment in Neuropsychiatry (WHO, 1994). All cases had been admitted to psychiatric inpatient units or seen by community-based mental health teams within the South London and Maudsley (SLaM) NHS Foundation Trust between December 2005 and October 2010(Di Forti et al., 2013). The study exclusion criteria were evidence of 1) psychotic symptoms precipitated by an organic cause; 2) evidence of transient psychotic symptoms resulting from acute intoxication as defined by ICD-10; 3) moderate or severe learning disabilities as defined by ICD-10; or 4) head injury causing clinically significant loss of consciousness.

2.2.Baseline predictors

Overall, 13 predictors were included in the models as described below. These predictors were chosen because they have consistently been implicated in the risk for schizophrenia onset and development of TRS (Ajnakina et al., 2018a; Di Forti at al., 2009; Fisher et al., 2010; Lally et al, 2016; Smart et al., 2019; Trotta et al., 2015).

2.2.1.Sociodemographic characteristics

Information on sociodemographic characteristics was collected using the Medical Research Council Socio-demographic Schedule (Mallett et al., 2002). Ethnicity was self-ascribed using the 16 categories employed by the UK Census in 2001. Cannabis use was measured with the Cannabis Experience Questionnaire modified version (Di Forti et al., 2009).

2.2.2.Clinical assessments

Baseline diagnoses were made from Schedules for Clinical Assessment in Neuropsychiatry (WHO, 1994) interviews and mental health records utilising the Operational Criteria Checklists (McGuffin et al., 1991) and ICD-10 (WHO, 1992). The degree of psychopathology at first presentation to mental health services was measured with the Positive and Negative Syndrome Scale (Kay et al., 1987). Duration of untreated psychosis (DUP) was defined as the difference between the date of onset of psychotic symptoms, as ascertained during face-to-face interview with trained researchers, and the date of initiation of treatment with antipsychotic medications (Malla et al., 2006; Singh et al., 2005). As DUP was skewed, we log-transformed it before including this variable in the analyses. The presence of either current or a previous diagnosis of psychosis in at least one first-degree relative was ascertained with the Family Interview for Genetic Studies (https://www. nimhgenetics.org/interviews/figs).

2.2.3.Childhood adversity

Childhood adversity (CA) occurring before 17 years of age was assessed using the Childhood Experiences of Care and Abuse Questionnaire (Bifulco et al., 2005). The presence of least one of the following six forms of CA was defined a presence of CA: i) physical abuse inflicted by either one or both parent-figures; ii) sexual abuse perpetrated by an individual at least 5 years senior to the recipient; iii) separation from either or both parent-figures for ≥6 months; iv) death of either or both biological parents; v) taken into care by authorities; and vi) number of changes in family arrangements (Fisher et al., 2010). The CECA.Q has been shown to have good internal consistency (Smith et al. 2002), satisfactory levels of test-retest reliability and

reasonable concurrent validity with the CECA.Q interview and Parental Bonding Instrument(Smith et al. 2002; Bifulco et al. 2005; Fisher et al. 2011).

2.3.Tracing patients at follow-up

Approximately five years after first contact with mental health services, we successfully traced 239 (84.5%) of the original FES cohort (Supplementary Table 1). Information at follow-up was collated from the electronic psychiatric record-keeping system within the SLaM Trust (Stewart et al., 2009) using the WHO Life Chart Schedule extended version (Sartorius et al., 1996). All deaths and emigrations up to and including those that occurred during the final year of follow-up were identified by a case-tracing procedure with the Office for National Statistics for England and Wales and the General Register Office for Scotland.

2.3.1.Outcomes

Following the National Institute for Health and Clinical Excellence (NICE) guideline (NICE guideline, 2014), patients were defined as having TRS if during the follow-up period they showed little or no symptomatic improvement to at least two consecutive treatments with antipsychotic medications of adequate dose and duration (≥6 weeks), as ascertained from the clinical records. A non-response to antipsychotic treatment was defined if 1) patients, having been treated with an antipsychotic medication of adequate dose and for an adequate duration did not show improvements in their clinical presentation as recorded by treating clinicians, and/or 2) the documented reason for switching antipsychotic medication was due to a lack of therapeutic response. An adequate daily dose of antipsychotic medication was defined according to a daily dose of ≥400mg chlorpromazine equivalence (Leucht et al., 2014). We only included as TRS cases those patients who failed to respond and not those who were intolerant of antipsychotic medications or those who self-discontinued antipsychotic medication (Ajnakina et al., 2018b; Lally et al., 2016). FES patients who met the criteria for TRS were divided into two groups: 1) “early-resistant” treatment resistance (E-TR) subgroup encompassed those cases who met criteria for TRS and who did not experience a

symptomatic remission from the time of the first presentation to the end of the follow-up period, and 2) “late-resistant” treatment resistance (L-TR) subgroup included those cases who had experienced a response to antipsychotics and attained a symptomatic remission (≥6 months duration), but at a later stage failed to respond to the ongoing use of antipsychotics, meeting the criteria for TRS (Ajnakina et al., 2018b; Lally et al., 2016).

2.4.Statistical analyses

The process of model development, validation and calibration was carried out according to methodological standards outlined by Steyerberg et al. (2009, 2019); the results were reported according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines (Collins et al., 2015).

2.4.1.Imputation of missing values

In the present study, the total missing values across the dataset was 32%; the proportion of missingness in each included variable is provided in Supplementary Table 2. Given that analyses of the subset of participants with no missing data in any of the variables can lead to either biased estimates or reduced precision of the predicted estimates (Steyerberg, 2009), we imputed the missing values. Assuming missing values did not depend on unobserved data (Sterne et al., 2009), we employed missForest for imputation of missing values, which is an iterative imputation method based on Random Forest (Stekhoven and Bühlmann, 2012). MissForest outperforms the well-known imputation methods, such as *k*-nearest neighbours and parametric multivariate imputation by chained equations (MICE) in the presence of large proportion of missingness, non-linearity and variable interactions (Stekhoven and Bühlmann, 2012). As recommended for prediction models (Kontopantelis et al, 2017), the outcomes were included in the imputation. To evaluate the imputation in the dataset, we estimated the imputation error with Normalized Root Mean Squared Error (NRMSE) for continuous variables and proportion of falsely classified (PFC) for categorical variables (Stekhoven and Bühlmann, 2012; Oba et al., 2003) where a value close to 0 represents an excellent performance, and a

value of 1 indicates poor performance. Accordingly, the imputation quality of missing values was good for categorical variables and moderate for continuous variables (NRMSE=0.498; PFC=0.253).

2.4.2. Model fitting

To develop prediction models for estimating an individual risk of E-TR and L-TR at 5-year follow-up, we employed regularized regression methods (RRMs), which compare competitively with more complex machine learning methods, such as random forest or support vector machines (Kuhn and Johnson, 2013; Salvador et al., 2017). Specifically, we employed penalized maximum likelihood (RIDGE) logistic regression model for binary outcomes (Hoerl and Kennard, 1970). While RIDGE penalises coefficients for overfitting, it does not do variable selection. Therefore, to see if some of the included predictors might have been more important for predicting each TRS subgroup than others, we also ran the least absolute shrinkage and selection operator (LASSO) model, which enabled variable selection simultaneously dealing with overfitting (Tibshirani, 1996); though, these RRMs do not produce confidence intervals. RIDGE and LASSO achieve their functions through regularisation, which entails imposing penalty (λ) on the size of regression parameter estimates. λ is a non-negative tuning parameter that controls the amount of penalisation, with increased penalisation for higher λ values (Hastie et al 2009). The tuning parameter λ optimising the model performance was selected separately for RIDGE and LASSO using repeated cross-validation methods; this method is comparable in terms of its efficiency and practicality to nested cross validation (see below).

2.4.3.Model estimation using repeated cross-validation

The tuning parameter λ optimising the model performance as measured with area under the receiver operating characteristic curve (AUC), which is recommended for imbalanced data, was chosen from a grid of 100 λ values through 5-fold repeated cross-validation (CV) (Tibshirani, 1997). Although CV produces nearly unbiased estimates of accuracy, the resulting

estimates may still have high variance (Efron, 1997). Therefore, we repeated the process 200 times for each λ value and computed the average AUC (Hastie et al 2009). The optimal λ was chosen as the one that had an AUC within one-standard error of the maximum (Hastie et al., 2009).

2.4.4.Model performance

The predictive ability of our models was assessed through discrimination and calibration. Discrimination was assessed using AUC (Bernardini et al., 2017); AUC value of 0.5 indicates that a model does not discriminate better than chance, while 1 indicates that a model discriminates perfectly. Calibration reflects the agreement between the predicted probabilities produced by the model and the observed outcome frequencies (Moons et al., 2012) and can be described as a measure of prediction bias in a model (Harrell et al., 1996). Calibration was assessed via calibration slope β, which should ideally be 1, and the calibration-in-the-large α, which ideally should be zero (Steyerberg, 2009). As discrimination and calibration values do not provide information about the distribution of true and false positives and negatives (Bernardini et al., 2017), we further derived sensitivity, specificity, positive predictive values (PPV) and negative predictive values (NPV) for each model (Altman and Bland, 1994a,b). To make optimal decisions for the classification based on the prediction models, the optimal cut-off point for the predicted probability (i.e., “decision thresholds”) (Steyerberg, 2009) was defined at the threshold which maximised overall correct classification rates and minimised misclassification rates, while choosing the point on the receiver operating characteristic curve farthest from chance (Perkins and Schisterman, 2006).

2.4.5.Model validity and re-calibration

To correct measures of predictive performance for optimism, defined as difference in test performance and apparent performance (Steyerberg et al., 2010), which occurs when a model’s predictions are more extreme than they should be for individuals in a new dataset from the same target population, we carried out internal validation of each model separately

using Harrell's optimism-correction procedure through 5-fold repeated CV iterated 200 times (Harrell et al., 1996). Accordingly, the whole model building process from imputing the missing values with missForest, selecting tuning parameter λ through repeated CV to fitting each model (i.e., LASSO and RIDGE) was repeated 1000 times on the 200×5 different resamples. We then estimated the overall optimism for each measure of performance as the median of the 1000 estimated optimisms (Supplementary Tables 5-6). To account for overfitting during the model development process, for each measure of performance (p), we obtained the optimism-corrected performance ($p_{corrected}$), by using the formula: $p_{corrected}=p_{apparent}-p_{optimism}$ (Steyerberg, 2019). We further recalibrated the models by updating the baseline betas for the entire dataset using $\beta_{corrected}$xb formula; similarly, we used $\alpha_{corrected}+\beta_{corrected}\times$intercept formula to obtain the recalibrated intercept; here, b are the uncalibrated log-odds ratios. The uncalibrated and recalibrated coefficients including intercepts from RIDGE and LASSO models are presented in Supplementary Table 7-10.

2.4.6.Power calculations

The current "rule of thumb" for sample size is to include at least 10 events per candidate predictor. However, it does not take into account several important aspects for accurate the needed sample size calculations, such as the magnitude of predictor effects, the overall outcome risk, the distribution of predictors, and the number of events for each category of categorical predictors (Riley et al., 2019). Therefore, we estimated the needed number of events (e.g., number of patients who will meet the criteria for either of treatment resistance schizophrenia subtype) per variable (e.g. degree of freedoms of predictors of treatment resistant schizophrenia subtypes (EPV) (Austin et al., 2017)) necessary for robust prediction models using calculations developed by Riley et al. (2019). Accordingly, based on an estimated prevalence of E-TR events in our sample (0.252) that occurred during the 5-year follow-up period, and 13 prognostic factors included in the model development, we needed 5.1 EPV which was slightly higher compared to 4.5 EPV available in our sample. Similarly, considering the prevalence of L-TR in the following 5 years was 0.126 and including 13

predictors, we needed to have 3.24 EPV; this was substantially higher compared to the available EPV (i.e., 1.8) in our sample.

3.Results

3.1.Sample characteristics

The core analytic sample comprised 239 FES patients with a mean length of follow-up of 5-years (SD=2.5 years). Of these, *n*=56 (25.2%) were defined as E-TR and *n*=24 (12.6%) were defined as L-TR. There were no significant differences between E-TR group and non-TR group in terms of ethnicity, socio-demographic and clinical characteristics at baseline (Supplementary Table 3). Cases in the L-TR group were significantly younger at the time of first contact with mental health services ($mean_{years}$=23.5, SD=4.8) compared to the non-TR group ($mean_{years}$=27.8, SD=8.3) ($t_{(188)}$=6.21, p=0.014), and a lower proportion of L-TR cases (43.5%) were men compared to non-TR group (66.3%) ($x_{(1)}^2$=4.51, p=0.034) (Supplementary Table 4).

3.2.Predicting E-TR: Performance of prediction models

Having retained 12 out 13 predictors included in the analyses (Supplementary Tables 7 and 9), LASSO demonstrated a good discrimination ($AUC_{corrected}$=0.74) and a good calibration (calibration slope β_{LASSO}=1.204 and calibration-in-the-large α_{LASSO}=0.188) (**Table 1**). Similarly, RIDGE model had high discrimination ($AUC_{corrected}$=0.77) and good calibration (β_{RIDGE}=1.264 and calibration-in-the-large α_{RIDGE}=0.028). To classify individuals at the high risk for E-TR, for LASSO model the decision threshold was estimated at 28.1%, and for RIDGE the decision threshold was estimated at 33.9%. Based on these thresholds, LASSO and RIDGE obtained excellent NPVs; specificity was lower for LASSO ($specificity_{corrected}$=0.71) compared to RIDGE model ($specificity_{corrected}$=1.00) (**Table 1**). Due to a very large estimated optimism for sensitivity in RIDGE model ($optimism_{sensitivity(RIDGE)}$=0.578) (Supplementary Table 6), after optimism-correction procedure sensitivity for RIDGE was shown to be equal to zero ($sensitivity_{corrected}$=0.00) (**Table 1**).

3.3.Predicting L-TR: Performance of prediction models

For predicting L-TR onset, LASSO retained all 13 predictors as important factors contributing to this outcome (Supplementary Tables 8 and 10). LASSO model had high discrimination ($AUC_{corrected}$=0.77) and a relatively good calibration (calibration slope β_{LASSO}=1.838 and calibration-in-the-large α_{LASSO}=0.504) (**Table 2**). RIDGE model also had good discrimination ($AUC_{corrected}$=0.75) and a relatively good calibration (calibration slope β_{RIDGE}=1.658 and calibration-in-the-large α_{RIDGE}=0.394). To classify individuals at the high risk for L-TR, for LASSO model the decision threshold was estimated at 12.8%, and for RIDGE the decision threshold was estimated at 12.6%. Using these thresholds, NPV for both RIDGE and LASSO was within an excellent range; PPV was higher for RIDGE model ($PPV_{corrected}$=0.59) than for LASSO model ($PPV_{corrected}$=0.42) (**Table 2**). LASSO had moderate sensitivity (0.62); sensitivity estimated for RIDGE model extremely low ($sensitivity_{corrected}$=0.00), which was due to having a very large estimated optimism ($optimism_{sensitivity(RIDGE)}$=0.579) (Supplementary Table 6).

4.Discussion

In the present study, we attempted to develop prediction models to predict an individual risk for meeting criteria for (i) early treatment resistant (E-TR) schizophrenia and (ii) late-treatment resistant (L-TR) schizophrenia during the first 5 years after the first contact with mental health services for first episode schizophrenia spectrum (FES) disorders. Having followed TRIPOD recommendations and employed factors known to be associated with poor schizophrenia outcomes, we utilised methods that are recommended for model development and validation when accuracy and interpretability are the priority for implementation of prediction models in practice (Fusar-Poli and Meyer-Lindenberg, 2016).

Encompassing information on ethnicity, cannabis use, low socio-economic status, family history of psychosis, childhood adversity, living arrangements and clinical presentation all collated on the first contact with mental health services for FES, our models had good

discriminative ability to identify those FES patients who were at a greater vs lesser risk to meet criteria for E-TR and L-TR. The obtained discriminative ability for each model is on par with prediction models for coronary heart disease (Wilson et al., 1998), breast cancer (Costantino et al., 1999) and cardiovascular disease (Hippisley-Cox et al., 2010), which are now included in clinical guidelines for therapeutic management. Moreover, the developed prediction models for estimating an individual risk for developing E-TR and L-TR following onset of FES showed excellent negative predictive values highlighting the models had strong ability to identify those FES patients who will not develop these TRS subtypes in the following 5 years; whereas, the high specificity implies these models are likely to correctly classify a higher proportion of FES patients as high risk for TRS subtypes in the following 5 years. Thus, utilising these models in practice can reduce the chances for FES patients being exposed to inappropriate intervention plans designed for treatment resistance in schizophrenia. The positive predictive values for E-TR ranged from 44% to 48% and for L-TR ranged from 42% to 59%. Considering that the prevalence of E-TR was 25% and 13% for L-TR in our sample, the obtained positive predictive values indicate that our models will be of advantage for clinical trials recruiting patients at risk for these TRS subtypes as they will require on average 19-46% less patients at risk for those outcomes before their onset. It is feasible, however, that improvements in these measures of prediction accuracy in our models might have been more significant if more complex machine learning methods, such as random forest or support vector machines, had been used (Hastie, 2009). Nonetheless, it has recently been shown that the more complex machine learning methods led to only minor improvements, if any at all, in prediction accuracy at the expense of reduced interpretability and lack of variable selection when compared with simpler statistical models (Christodoulou et al., 2019; van der Ploeg et al., 2014).

Because we aimed to develop models that would be likely accepted and implemented in clinical care, we included only those predictors that were consistently highlighted as risk factors for schizophrenia and TRS onset in the literature (Ajnakina et al., 2018a; Di Forti at al., 2009; Fisher et al., 2010; Lally et al, 2016; Smart et al., 2019; Trotta et al., 2015). Using *a*

priori knowledge to identify the most robust predictors to be included in prediction models is also a recommended approach for ensuring EPV is adequate for the analyses (de Jong et al., 2019; Fusar-Poli et al., 2018). This is because the selection of such predictors would be limited in number (preserving the EPV) and independent of the data on which the model is then tested (Studerus et al., 2017). However, having chosen variables based on *a priori* knowledge meant we had to retain a higher number of predictors in models than was advisable through sample size calculations (Riley et al., 2019), especially when developing a prediction model for L-TR, leading to reduced power.

This reduced power may explain why LASSO selected almost all predictors. However, having utilised only those variables that have been implicated in risk for poorer schizophrenia outcomes, it is equally feasible that LASSO retained all, or almost all, predictors in the models because all included predictors play an important role in the onset of E-TR and L-TR in the following 5 years in patients with FES. If this assertion is accurate, then our results reiterate the important role that such factors as cannabis use, low socio-economic status, family history of psychosis, childhood adversity and adverse living arrangements play in increasing risk for TRS subtypes providing avenues for prevention strategies. Although preventing childhood abuse is currently beyond our powers, as a means to reduce risk for onset of TRS subtypes, we can certainly advocate public health campaigns to educate people about the harms of cannabis use, while early intervention service could aim to improve or alleviate adverse environmental circumstances for patients with FES (Murray et al., 2020).

Although the reduced power may further explain the presence of very large optimism observed in sensitivity, especially in RIDGE model, leading to low internally validated sensitivity in predicting E-TR and L-TR in the following 5 years, the low sensitivity might, at least in part, be due to not having included other variables implicated in treatment response, or lack of thereof, in FES patients. The fact that approximately 70% of TRS patients were shown to be resistant to available treatment from their first contact with mental health services for schizophrenia

(Lally et al., 2016) in combination with accumulated evidence demonstrating that younger age of illness onset is associated with a greater risk for a poor outcome and TRS (Lally et al., 2016, Meltzer et al., 1997), suggests that neurodevelopmental disruption may play a crucial role in TRS onset (Murray et al., 1992). Indeed, localised differences in gyrification between TRS and non-TRS groups were previous observed (Palaniyappan et al., 2016); though, many other biological defects, such as glutamatergic or dopaminergic abnormalities and reductions in grey matter tissue (Gillespie et al., 2017), are likely to be associated with an increased risk for treatment resistance in patients with schizophrenia (Gillespie et al., 2017, Molent et al., 2019). However, a good prediction model ought to be based on data reflecting the real-life clinical information available to a physician and a participant when they need to make decisions on the likely risk of TRS subtypes for an individual during the next 5 years. Therefore, models developed with variables on neuroanatomical domains would be constrained by logistical and financial challenges that can impede the ability to implement them in everyday clinical practice.

4.1. Methodological consideration

Strengths of this study include the longitudinal design (Wynants et al., 2017), which allowed for the follow-up of FES cases over the first five years of their illness. We have employed rigorous methodology for model development and evaluation, including power calculation. To maximise the predictive accuracy (Cowley et al., 2019), we catered for incomplete data, which is a common but serious limitation in psychiatric research but generally not addressed sufficiently (Moons et al., 2006). The predictors included in our models are often collected in epidemiological studies and can be ascertainable during a brief patient-physician discussion.

Limitations include the small population of female patients and the lack of a robust measure of medication adherence, which may have affected the number of FES patients meeting the criteria for TRS and its subtypes (McCutcheon et al., 2015). We were unable to carry out an external validation of our models due to the lack of comparable data. The percentage of missing values across variables might have affected the imputations and induced some bias

in the estimates of the effects in the model. Nonetheless, the proportion of missingness in the present study was comparable to many longitudinal datasets (Ajnakina et al., 2020; Morgan et al., 2014) and within the range for missForest to handle it efficiently (Stekhoven and Bühlmann, 2012). It may also be argued that missForest might was unable to handle well a very small dataset with a relatively large portion of missing values leading to the observed overfitting especially for RIDGE models, alluding to a possibility that MICE method might have been more appropriate. Although some overfitting might have come from utilising MissForest, MICE would have introduced the problem of having multiple models with a different set of variables and models' parameters selected for each multiple imputation, which would have not been ideal for clinical practice. Other methods entailing a combination of LASSO and MICE techniques have been proposed (Liu et al. 2016), which again would have not been clear and interpretable enough for clinicians.

Finally, it may be argued that time-to-even models might have been more appropriate for developing models for estimating individual risk of TRS subtypes in the following 5 years. However, defining patients with schizophrenia as TRS is significantly dependent on when clozapine has been prescribed by a treating clinician; the average time to clozapine prescription is 5 years, which is considerably longer than recommended by NICE guidelines (Howes et al., 2012). In fact, clozapine has been shown to be commenced in less than half of those with TRS over the course of the 5 years follow-up (Lally et al, 2016) meaning the use of high-dose antipsychotics or polypharmacy prior to clozapine is needed to identify those who are TRS (Howes et al., 2012). For these reasons, the time to TRS onset cannot be accurately estimated. Thus, to avoid introducing any biases, we chose models for binary outcomes rather than time to event.

4.2. Conclusion

Using factors that are known to be associated with poor schizophrenia outcomes and employing advanced statistical shrinkage methods, our results showed it was possible to

predict with sufficient accuracy who would meet criteria for early-treatment resistance and late-treatment resistance during the 5-year follow-up after the first contact with mental health services for schizophrenia using regularised regression models. However, sensitivity, especially for RIDGE model, was low implying that further work is necessary to explore way of improving these prediction models for such rare but important outcomes before they can be used for a more in-depth risk assessment, follow-up monitoring and individually tailored prevention strategies.

Funding

This work was supported by the National Institute for Health Research Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. O.A. is funded by the National Institute for Health Research (NIHR Post-Doctoral Fellowship - PDF-2018-11-ST2-020) for this project. R.M.M. receives salary support from the NIHR Maudsley BRC. MDF is funded by Clinician Scientist Medical Research Council fellowship (project reference MR/M008436/1). A.T. is supported by a National Institute of Health Research Maudsley BRC Preparatory Clinical Research Training Fellowship. H.L.F is supported by an MQ Fellows Award (MQ14F40). F.G. is part funded by the National Institute for Health Research Collaboration for Leadership in Applied Health Research & Care Funding scheme with support from the National Institute for Health Research Biomedical Research Centre at South London and Maudsley NHS Foundation Trust. F.G. also receives support from the Stanley Medical Research Institute and the Maudsley Charity. D.S. and D.A. were part funded part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

Acknowledgements

The patients included in the present study were recruited in collaboration with the GAP and PUMP study teams and the South London and Maudsley (SLaM) NHS Foundation Trust. We would like to thank the patients who gave up their time to take part in this study and all of the staff and students who worked tirelessly to collect the data.

Competing interests

R.M.M. has received honoraria from Janssen, Astra-Zeneca, Lilly, and BMS. A.S.D. has received honoraria from Janssen and Roche Pharmaceuticals. F.G. has received honoraria for advisory work and lectures from Lundbeck, Otsuka and Sunovion and has a family member with professional links to Lilly and GSK. The other authors (OA, DA, JL, MDF, HLF, AT, VM, CP, PD, AS) have no competing interests to declare.

5. References

Ajnakina et al., 2017. Patterns of illness and care over the 5 years following onset of psychosis in different ethnic groups; the GAP-5 study. Soc Psychiatry Psychiatr Epidemiol 5, 017-1417.

Ajnakina et al., 2018b. Validation of an algorithm-based definition of treatment resistance in patients with schizophrenia. Schizophr Res. 197, 294-297

Ajnakina O, Lally J, Di Forti M, et al. 2018a. Different types of childhood adversity and 5-year outcomes in a longitudinal cohort of first-episode psychosis patients. Psychiatry Res. 269:199-206.

Ajnakina, O., Cadar, D., Steptoe, A. 2020. Interplay between Socioeconomic Markers and Polygenic Predisposition on Timing of Dementia Diagnosis. J Am Geriatr Soc. 68(7):1529-1536

Altman, D. G. & Bland, J. M. 1994a. Diagnostic tests 2: Predictive values. BMJ 309, 102.

Altman, D. G. & Bland, J. M. 1994b. Diagnostic tests. 1: Sensitivity and specificity. BMJ 308, 1552.

Austin, P.C. & Steyerberg, E.W. 2017. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res 26, 796-808.

Bernardini et al., 2017. Risk Prediction Models in Psychiatry: Toward a New Frontier for the Prevention of Mental Illnesses. J Clin Psychiatry 78, 572-583.

Bifulco et al., 2005. The childhood experience of care and abuse questionnaire (CECA.Q): validation in a community series. Br J Clin Psychol 44, 563-81.

Califf et al., 1997. Selection of thrombolytic therapy for individual patients: development of a clinical model. GUSTO-I Investigators. Am Heart J 133, 630-9.

Christodoulou et al., 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 110, 12-22.

Collins et al., 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Bjog 122, 434-43.

Costantino JP, Gail MH, Pee D, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J Natl Cancer Inst.* 1999;91(18):1541-1548.

Cowley et al., 2019. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. Diagn Progn Res 3, 16.

de Jong et al., 2019. Sample size considerations and predictive performance of multinomial logistic prediction models. Stat Med 38, 1601-1619.

Derksen, S. & Keselman, H.J. 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. Br J Math Stat Psychol 42, 265-282.

Di Forti et al., 2013. Daily Use, Especially of High-Potency Cannabis, Drives the Earlier Onset of Psychosis in Cannabis Users. Schizophr Bull. 40(6),1509-17

Di Forti M, Morgan C, Dazzan P, et al. 2009. High-potency cannabis and the risk of psychosis. Br J Psychiatry. 195(6): 488-91.

Efron, B.T., Robert 1997. Improvements on Cross-Validation: The 632+ Bootstrap Method. Journal of the American Statistical Association 92, 548-560.

Fan, J. & Lv, J. 2010. A Selective Overview of Variable Selection in High Dimensional Feature Space. Stat Sin 20, 101-148.

Fisher HL, Jones PB, Fearon P, et al. 2010. The varying impact of type, timing and frequency of exposure to childhood adversity on its association with adult psychotic disorder. Psychol Med. 40(12): 1967-78.

Fisher, H. L., T. K. Craig, P. Fearon, K. Morgan, P. Dazzan, J. Lappin, G. Hutchinson, G. A. Doody, P. B. Jones, P. McGuffin, R. M. Murray, J. Leff, and C. Morgan. 2011. Reliability and comparability of psychosis patients' retrospective reports of childhood abuse, Schizophr Bull, 37: 546-53.

Fusar-Poli, P. & Meyer-Lindenberg, A. 2016. Forty years of structural imaging in psychosis: promises and truth. Acta Psychiatr. Scand. 134, 207-224.

Gillespie et al., 2017. Is treatment-resistant schizophrenia categorically distinct from treatment-responsive schizophrenia? a systematic review. BMC Psychiatry 17, 12.

Harrell, F.E., Jr., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Stat Med, 1996. **15**(4): 361-87.

Hastie T, Tibshirani R & J., F. 2009. Model assessment and selection. The elements of statistical learning: data mining, inference and prediction. Springer.

Hastie, T., Tibishirani, R and Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer New York.

Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *Bmj.* 2010;341:c6624.

Hoerl A, Kennard R. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12:55-67

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 Psych 201, 481-5.

Kay, S.R., Fiszbein, A. & Opler, L.A. 1987. The positive and negative syndrome scale PANSS for schizophrenia. Schizophr Bull 13, 261-76.

Kennedy et al., 2014. The social and economic burden of treatment-resistant schizophrenia: a systematic literature review. Int Clin Psychopharmacol 29, 63-76.

Kontopantelis, E., White, I.R., Sperrin, M. et al. 2017 Outcome-sensitive multiple imputation: a simulation study. BMC Med Res Methodol 17:2 doi 10.1186/s12874-016-0281-5

Kuhn, M. & Johnson, K. 2013. Applied predictive modeling. Springer: New York.

Lally J, Ajnakina O, Di Forti M, et al. 2016. Two distinct patterns of treatment resistance: clinical predictors of treatment resistance in first-episode schizophrenia spectrum psychoses. Psychol Med. 8: 1-10.

Leucht, S., Samara, M., Heres, S., Patel, M. X., Woods, S. W. & Davis, J. M., 2014. Dose equivalents for second-generation antipsychotics: the minimum effective dose method. Schizophr Bull 40, 314-26.

Liu, Y., Wang, Y., Feng, Y., Wall, M.M. 2017. Variable selection and prediction with incomplete high-dimensional data. Ann Appl Stat. 2016 Mar; 10(1): 418-450

Malla et al., 2006. Predictors of rate and time to remission in first-episode psychosis: a two-year outcome study. Psychol Med 36, 649-58.

Mallett et al., 2002. Social environment, ethnicity and schizophrenia. A case-control study. Soc Psychiatry Psychiatr Epidemiol 37, 329-35.

McCutcheon et al., 2015. Treatment resistant or resistant to treatment? Antipsychotic plasma levels in patients with poorly controlled psychotic symptoms. J Psychopharmacol 29, 892-7.

McGuffin, P., Farmer, A. & Harvey, I. 1991. A polydiagnostic application of operational criteria in studies of psychotic illness. Development and reliability of the OPCRIT system. Arch Gen Psychiatry. 488:764-70.

Meltzer et al., 1997. Age at onset and gender of schizophrenic patients in relation to neuroleptic resistance. Am J Psychiatry 154, 475-82.

Molent et al., 2019. Functional neuroimaging in treatment resistant schizophrenia: A systematic review. Neurosci Biobehav Rev 104, 178-190.

Moons et al., 2006. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 59, 1092-101.

Moons et al., E. 2012. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new biomarker. Heart 98, 683-90.

Morgan, C., Lappin, J., Heslin, M., Donoghue, K., Lomas, B., Reininghaus, U., Onyejiaka, A., Croudace, T., Jones, P.B., Murray, R.M., Fearon, P., Doody, G.A., Dazzan, P., 2014. Reappraising the long-term course and outcome of psychotic disorders: the AESOP-10 study. Psychol. Med. 44, 2713-2726.

Murray et al., 1992. A neurodevelopmental approach to the classification of schizophrenia. Schizophr Bull 18, 319-32.

Murray, M., David, A., Ajnakina, O. 2020. Prevention of Psychosis: Moving on from the At-Risk Mental State to Universal Primary Prevention. Psychol Med, *in press*

National Institute for Health and Clinical Excellence guideline, 2014. Psychosis and schizophrenia in adults: treatment and management (Clinical guideline 178). Royal College of Psychiatrists: London.

Oba, S., M. A. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii. 2003. A Bayesian missing value estimation method for gene expression profile data, Bioinformatics, 19: 2088-96.

Osborne et al., 2012. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27-46.

Palaniyappan L, Marques TR, Taylor H, et al. Globally Efficient Brain Organization and Treatment Response in Psychosis: A Connectomic Study of Gyrification. *Schizophr Bull* 2016;42(6):1446-1456.

Perkins, N.J. & Schisterman, E.F. 2006. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol 163, 670-5.

Riley et al., 2019. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 38, 1276-1296.

Salvador et al., 2017. Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. PLoS One 12, e0175683.

Sartorius et al., 1996. Long-term follow-up of schizophrenia in 16 countries. A description of the International Study of Schizophrenia conducted by the World Health Organization. Soc Psychiatry Psychiatr Epidemiol 31, 249-58.

Singh et al., 2005. Determining the chronology and components of psychosis onset: The Nottingham Onset Schedule NOS. Schizophr Res 80, 117-30.

Smart et al., 2019. Predictors of treatment resistant schizophrenia: a systematic review of prospective observational studies. Psychol Med, 1-10.

Smith, N., D. Lam, A. Bifulco, and S. Checkley. 2002. 'Childhood Experience of Care and Abuse Questionnaire (CECA.Q). Validation of a screening instrument for childhood adversity in clinical populations', Soc Psychiatry Psychiatr Epidemiol, 37: 572-9.

Stekhoven Daniel J. & Bühlmann Peter 2012. MissForest-non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112-118.

Sterne et al., 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 29.

Stewart, R., Soremekun, M., Perera, G., Broadbent, M., Callard, F., Denis, M., Hotopf, M., Thornicroft, G. & Lovestone, S. 2009. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre SLAM BRC case register: development and descriptive data. BMC Psychiatry 9, 9-51.

Steyerberg E. 2019. *Clinical Prediction Models. A practical approach to development, validation, and updating.* Second Edition ed: Springer Nature Switzerland

Steyerberg et al., 2000. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 19, 1059-79.

Steyerberg et al., 2001. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making 21, 45-56.

Steyerberg et al., 2010. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21, 128-38.

Steyerberg, E. 2009. Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating. Springer: New York.

Steyerberg, E.W., Eijkemans, M.J. & Habbema, J.D. 1999. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 52, 935-42.

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, 1163–1178

Trotta A, Di Forti M, Iyegbe C, et al. Familial risk and childhood adversity interplay in the onset of psychosis. BJPsych Open 2015; 1(1): 6-13.

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, 1-13.

WHO 1992. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. Geneva.

WHO 1994. Schedules for Clinical Assessment in Neuropsychiatry: Version 2: Manual. World Health Organization, Division of Mental Health.

Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. 1998. Prediction of coronary heart disease using risk factor categories. *Circulation.* 97(18):1837-1847.

World Health Organization 1992. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. World Health Organization.

World Health Organization 1994. Schedules for Clinical Assessment in Neuropsychiatry: Version 2 : Manual. World Health Organization, Division of Mental Health.

Table 1. Internally validated prediction accuracy of the prediction models developed to predict onset of E-TR in patients with first episode schizophrenia spectrum disorders during the 5-year follow-up

Performance measures	**LASSO**	**RIDGE**
AUC	0.74	0.77
Calibration slope β	1.204	1.264
Calibration-in-the-large α	0.188	0.028
PPV	0.44	0.48
NPV	0.86	0.84
Sensitivity	0.66	0.00
Specificity	0.71	1.00
N of cases in the outcome	56	56

E-TR, early treatment resistant schizophrenia spectrum disorders; SE, standard error; AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value

Table 2. Internally validated prediction accuracy of prediction models developed to predict onset of L-TR in patients with first episode schizophrenia spectrum disorders during the 5-year follow-up

Performance measures	**LASSO**	**RIDGE**
AUC	0.77	0.75
Calibration slope β	1.838	1.658
Calibration-in-the-large α	0.504	0.394
PPV	0.42	0.59
NPV	1.00	0.81
Sensitivity	0.62	0.00
Specificity	1.00	1.00
N of cases in the outcome	24	24

L-TR, late treatment resistant schizophrenia spectrum disorders; SE, standard Error; AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value