Schizophrenia Research 246 (2022) 241–249

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

Developing and internally validating a prognostic model (P Risk) to improve the prediction of psychosis in a primary care population using electronic health records: The MAPPED study

Sarah A. Sullivan [a,b,*], Daphne Kounali [a,b], Richard Morris [b,c], David Kessler [a], Willie Hamilton [d], Glyn Lewis [e], Philippa Lilford [a], Irwin Nazareth [e]

[a] Centre for Academic Mental Health, Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK
[b] National Institute for Health Research, Bristol Biomedical Research Centre, University Hospitals Bristol NHS Foundation Trust, University of Bristol, Bristol, UK
[c] Centre for Academic Primary Care, Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK
[d] University of Exeter Medical School, Exeter EX1 2HW, UK
[e] UCL Division of Psychiatry, Maple House, Tottenham Court Rd, London W1T 7NF, UK

ARTICLE INFO

Keywords:
First episode psychosis
Prediction
Electronic health records
Primary care

ABSTRACT

Background: An accurate risk prediction algorithm could improve psychosis outcomes by reducing duration of untreated psychosis.
Objective: To develop and validate a risk prediction model for psychosis, for use by family doctors, using linked electronic health records.
Methods: A prospective prediction study. Records from family practices were used between 1/1/2010 to 31/12/2017 of 300,000 patients who had consulted their family doctor for any nonpsychotic mental health problem. Records were selected from Clinical Practice Research Datalink Gold, a routine database of UK family doctor records linked to Hospital Episode Statistics, a routine database of UK secondary care records. Each patient had 5–8 years of follow up data. Study predictors were consultations, diagnoses and/or prescribed medications, during the study period or historically, for 13 nonpsychotic mental health problems and behaviours, age, gender, number of mental health consultations, social deprivation, geographical location, and ethnicity. The outcome was time to an ICD10 psychosis diagnosis.
Findings: 830 diagnoses of psychosis were made. Patients were from 216 family practices; mean age was 45.3 years and 43.5 % were male. Median follow-up was 6.5 years (IQR 5.6, 7.8). Overall 8-year psychosis incidence was 45.8 (95 % CI 42.8, 49.0)/100,000 person years at risk. A risk prediction model including age, sex, ethnicity, social deprivation, consultations for suicidal behaviour, depression/anxiety, substance abuse, history of consultations for suicidal behaviour, smoking history and prescribed medications for depression/anxiety/PTSD/OCD and total number of consultations had good discrimination (Harrell's C = 0.774). Identifying patients aged 17–100 years with predicted risk exceeding 1.0 % over 6 years had sensitivity of 71 % and specificity of 84 %.
Funding: NIHR, School for Primary Care Research, Biomedical Research Centre.

1. Introduction

Clinical and social outcomes of psychosis are often poor. Approximately 25 % relapse within the first 3 years of treatment and residual symptoms are common. Many risk factors for a poor outcome, e.g. low socio-economic status (Simonsen et al., Jul 2007), genetic factors, age, sex, family history of psychosis and adverse life events (Perkins et al., 2020), are difficult to modify. Consequently, there is interest in the duration of untreated psychosis (DUP; the time-period between the first psychotic symptom and receiving specialist treatment), as it is positively associated with poorer outcomes (Marshall et al., 2004), and potentially modifiable.

* Corresponding author at: Centre for Academic Mental Health, Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK.
E-mail addresses: sarah.sullivan@bristol.ac.uk (S.A. Sullivan), daphne.kounali@bristol.ac.uk (D. Kounali), Richard.morris@bristol.ac.uk (R. Morris), david.kessler@bristol.ac.uk (D. Kessler), W.Hamilton@exeter.ac.uk (W. Hamilton), glyn.lewis@ucl.ac.uk (G. Lewis), philippa.lilford@bristol.c.uk (P. Lilford), i.nazareth@ucl.ac.uk (I. Nazareth).

https://doi.org/10.1016/j.schres.2022.06.031
Received 5 October 2021; Received in revised form 17 May 2022; Accepted 25 June 2022
Available online 14 July 2022

Most people with psychosis in the UK enter specialist mental health care via their family doctor. A shorter DUP is associated with more family doctor visits prior to diagnosis (Skeate et al., Sep 2002); family doctors are therefore a vital part of the psychosis care pathway. It is consequently important that family doctors recognise early warning signs of psychosis to expedite referral to specialist services. Accuracy of psychosis diagnoses in family doctor electronic health records (EHRs) are valid (Nazareth et al., 1993); however, family doctors under-identify more insidious symptoms (Platz et al., Dec 2006), because prodromal symptoms are frequently non-specific. Furthermore, most family doctors see few incident cases of psychosis annually and have little experience in early detection. There is also evidence (Chew-Graham et al., 2007) of barriers in making referrals to specialist mental health services. This is largely due to a lack of clarity between the referrers and the community mental health teams, leading to conflict between GPs and the mental health teams. We believe that joint decision-making can be enhanced if people at risk are identified by a validated risk algorithm. There are currently no risk prediction algorithms to screen the population at scale. Clinical symptoms scores and cognitive assessments, show promise, but are costly and too specialised for family doctors with limited potential for automation. An automatable risk prediction tool with predictors commonly collected by family doctors is urgently needed. There is potential to alert family doctors to the early warning signs of psychosis, if it were possible to use data already stored in EHRs.

Little is known about the progression of prodromal symptoms of psychosis and the few existing studies are based on high risk (Lunsford-Avery et al., May 2015) patients receiving specialist mental health treatment.

We previously (Sullivan et al., 2018) investigated whether prespecified symptoms identified attenders who later developed psychosis, using family doctor consultation data collected prior to a diagnosis of psychosis. We found that specific prodromal behaviours (e.g. cannabis use) and symptoms (e.g. depression and mania) were strongly associated with future psychosis. The positive predictive value of these factors varied with age and gender. We identified both a pattern in consultation frequency for some of the prodromal behaviours/symptoms up to 5 years before diagnosis, and that people later diagnosed with psychosis were heavier users of family doctor services than those developing other conditions.

We hypothesised that these candidate predictors, recorded in EHRs, could be used to develop an accurate, psychosis risk prediction model for family doctors.

2. Material and methods

2.1. Data source

A clinical register-based cohort was selected from the Clinical Practice Research Datalink Gold (CPRD Gold), a computerised database of anonymised, longitudinal UK family doctor records maintained by the Medicines and Healthcare Products Regulatory Agency. CPRD patients are representative of the UK population regarding age, sex and ethnicity (Herrett et al., Jun 2015). Participating practices enter demographic, diagnostic, consultation, prescribing, and referral data using a library of over 100,000 codes. CPRD Gold is an anonymised copy of these records. Validation studies (Walley and Mantgani, 1997) report that quality and completeness of data is high and that recording of a diagnosis of psychotic disorders was accurate.

2.2. Source population

All patients with up to standard (UTS- a practice-based quality metric based on the continuity of recording and the number of recorded deaths (Herrett et al., Jun 2015)) 'research quality' data registered with family practices in the CPRD Gold family doctor dataset with at least 5 years of follow up data, and links to secondary care outpatient data and Indices of Multiple Deprivation (IMD).

2.3. Study population (Fig. 1)

All patients with at least one consultation (Appendix A) and/or a prescription (Appendix B) for any mental health problem within the study period (1/1/10 to 30/9/18) and at least 365 days of registration before the first mental health consultation. The 'useful' follow-up period for each patient started with the later of current registration or practice entry date into the CPRD and ended with the earlier of transfer out or last collection date (determined from the practice file). Cohort entry date was the date of the first consultation or prescribed medication for any nonpsychotic mental health disorder (index consultation). We requested data from 1/1/2010 to ensure that the new dataset did not contain any patients from our previous study.

2.4. Exclusion criteria (Fig. 1)

A CPRD Gold record of: (a) a diagnosis of a psychotic disorder (Appendix C), and/or a prescription of an antipsychotic medication (Appendix D) made (b) either before the index consultation or within 365 days of the index consultation or (c) a diagnosis of epilepsy, head injury, dementia, or learning disability ever (Appendix E).

2.5. Data linkage

2.5.1. Hospital episode statistics (HES)

Psychosis is diagnosed in secondary care and later recorded in family doctor notes. To ensure that outcome detection was complete we linked CPRD Gold at the individual patient level to the HES (basic Admitted Care) database, which records secondary healthcare events in England. A linked HES record was provided for each patient in the defined source population if they had: at least one day of records which coincided with data collection for the linked dataset, a set of identifiers in CPRD Gold and HES datasets and received NHS-funded treatment at an English hospital during data collection. After linkage we excluded all patients with a HES ICD10 diagnosis of psychosis (F20-F29) before the index consultation date but retained those with a coded diagnosis within a year after the index consultation date. This was done to avoid counting historical data entered within the first year of a person having transferred from one practice to another, or because these apparent incident cases might have simply been prevalent cases (Lewis et al., 2005).

2.5.2. Patient-level index of multiple deprivation (IMD)

The IMD (Abel et al., 2016) combines 7 components of deprivation (income, employment, education, health, crime barriers to housing and living environment) to create a weighted score. The IMD is widely used within the UK to classify relative deprivation in small geographic areas (i.e. the Lower Super Output Area – LSOA). The patient's current or most recent postcode was used to assign a LSOA of residence, which is linked to quintiles of the IMD2015 score.

2.6. Socio-demographic measures

Family doctor practice-level ethnicity (Appendix F) and deprivation (see 2.5.2) according to postcode and geographical location (9 English regions).

2.7. Time at risk

From the first consultation and/or prescription for a nonpsychotic mental health problem or referral to mental health services. The end-date was the earliest date on which HES and/or CPRD records confirmed a psychosis diagnosis, or the date on which the individual left the family doctor practice, died, or a practice ceased providing data for CPRD.

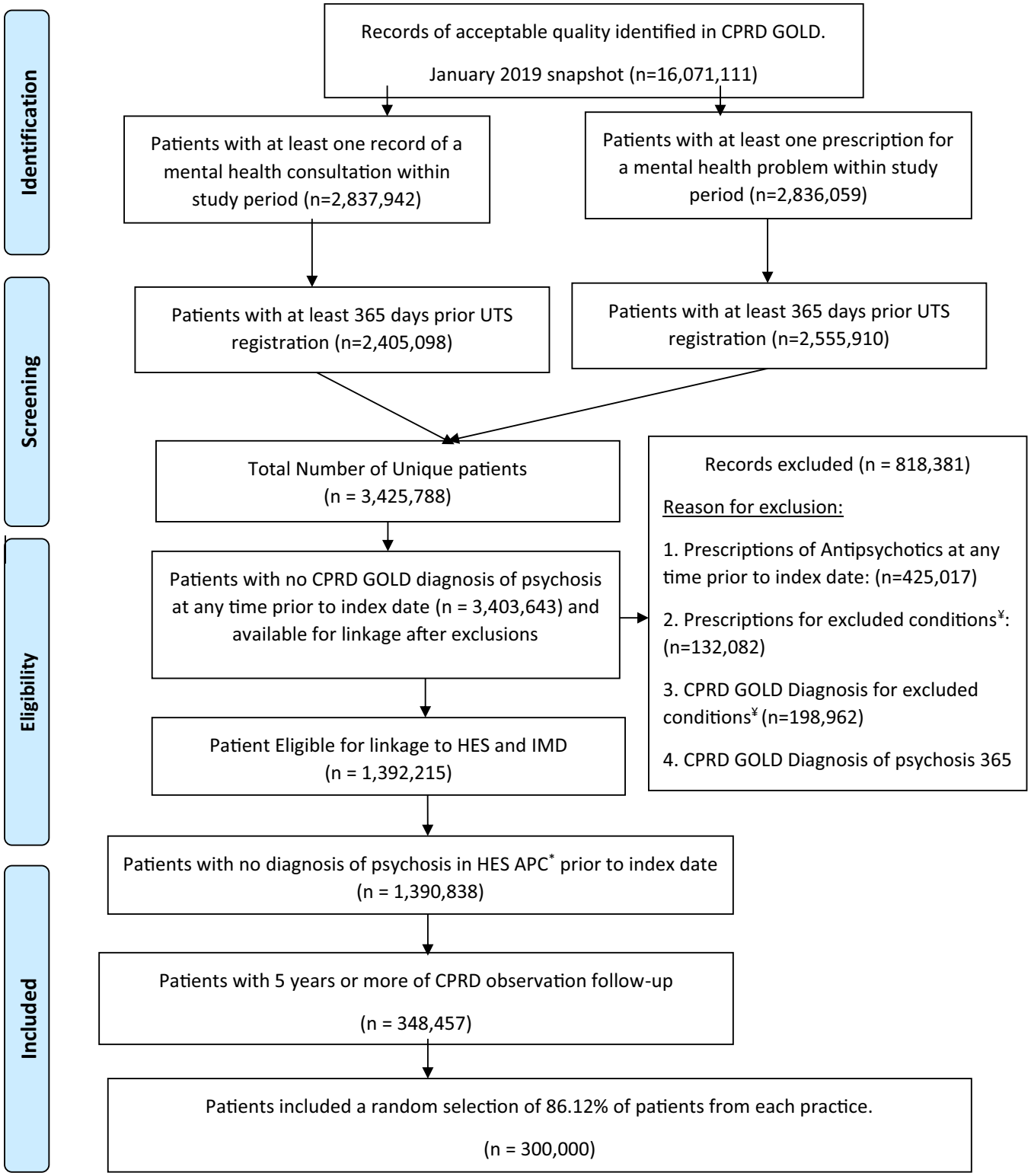

Fig. 1. Cohort Selection Flowchart.

2.8. Study measures

2.8.1. Primary outcome

Time to an ICD10 coded diagnosis of psychosis in HES and/or CPRD Gold (Appendix C).

2.8.2. Individual level predictors

A coded consultation/diagnosis of; depression/anxiety, suicidal behaviour (including self-harm, ideation, and attempts), smoking history (advice for and problems with cessation), problems with cannabis, attention deficit hyperactivity disorder (ADHD), obsessive compulsive disorder (OCD), sleep disturbance, mania, social isolation, blunted affect, role functioning problems, substance abuse, and bizarre

behaviour. Since further consultations could reveal new symptoms (or prescriptions) during follow-up, we allowed variables representing each symptom (or prescription) to change their value during follow-up, and hence were "time-dependent covariates". We also used historical (before cohort entry) coded consultations/diagnoses for all individual predictors (as listed above) and medication prescriptions for depression/anxiety, post-traumatic stress disorder (PTSD), OCD and substance abuse during follow up. Other predictors were age, gender, black ethnicity (Kirkbride et al., 2017), index of multiple deprivation. We also used the number of consultations for *predictors* during follow up and historically, and the total number of consultations for *any mental health problem* within the study period.

2.9. Sample size

Sample size calculations were based on conservative guidelines (Ogundimu et al., Aug 2016) using 20 events per model predictor. A recent study (Hardoon et al., 2013) investigating serious mental health disorder incidence (including psychosis) in a general population sample, found a rate of 46.4 per 100,000-person years at risk (PYAR). Since average follow-up time in CPRD was assumed to be 5 years, estimated incidence would be 0.232 %. A CPRD dataset of 300,000 people would therefore contain 695 psychosis diagnoses, providing sufficient precision for a model with up to 34 predictors.

2.10. Ethics

Approval was obtained from the CPRD's Independent Scientific Advisory Committee.

3. Theory/calculation

3.1. Model development and internal validation

We used statistical prediction methods rather than machine-learning so that the resulting prognostic index would be transparent to clinicians. Cox proportional hazards models (Anderson et al., Nov 1983) were used to evaluate the effects of multiple predictors. We used dates for time at risk when symptoms or medications were recorded at entry or during follow up, to act as time-dependent predictors. This was explicitly accounted for in the analyses by splitting individual records at the consultation timing. We also stratified follow-up time: <2 years, 2–6 years and >6 years.

We included all predictors except those which occurred for fewer than 15 patients of either sex at cohort entry in our "full model". Interactions were included between each predictor with follow up time where non-proportional hazards had been identified, and the interaction of age and sex, to reflect the well-known differing patterns of incidence with age in males and females. Model development involved identifying a "reduced model" through a backward step-down approach (with total residual Akaike Information Criterion as the stopping rule) to identify the most parsimonious set of predictors (Lawless and Singhal, 1978). We planned to assess the ability of the model to discriminate between individuals who did or did not later acquire a diagnosis of psychosis with the Harrell's C-statistic (Harrell, 2015). When simply re-applied to the same data from which the model was derived ("naïve estimate"), the C-statistic will be overoptimistic, so we used 300 bootstrap samples: these were chosen from the original sample (with replacement) (Efron and Tibshirani, 1994). For each bootstrap sample, a reduced model was identified and the C-statistic calculated both for the bootstrap sample and the original sample. The average degree of overoptimism was calculated and subtracted from the naïve estimate.

After our final model was identified, we calculated predicted risk over 6 years for patients whose follow up time exceeded 6 years, or who experienced the outcome within 6 years. Furthermore, we restricted this to patients aged 17–100 years at entry to the cohort since it is unlikely that the risk score would be applied by clinicians outside this age range. We calculated sensitivity, specificity, and likelihood ratios for risk thresholds of 0.5 %, 1 %, 1.5 % and 2 %. These low-risk thresholds were chosen because of the low incidence and prevalence of psychosis in primary care.

3.2. Missing data

Where predictors were not recorded, these were assumed to be absent on the consultation date. We aggregated individual data for ethnicity to the practice level, among patients with these data recorded. Data where the index of multiple deprivation was missing were omitted (0.2 %).

3.3. Data access and cleaning

The study sample was selected and cleaned by CPRD data managers.

3.4. Reporting guidelines

Findings are reported using RECORD (Benchimol et al., Oct 2015) and TRIPOD (Moons et al., 2015) guidelines.

3.5. Role of the funding source

The funders had no role in study design, data collection, analysis, or interpretation nor in the report writing or the decision to submit the paper for publication.

4. Results

4.1. Study population sociodemographic and clinical characteristics

Linkage was available for 1,392,312 patients with at least one record of a mental health consultation. Of these, 348,457 had at least 5 years of follow up data. CPRD charging policies precluded a population of >300,000 patients, so a random selection of 86 % of patients from each participating practice was made (Fig. 1).

830 coded diagnoses of a psychotic disorder were detected (Appendix G). Of these, 475 were recorded in CPRD Gold alone, 226 were recorded in HES alone, or were recorded in both ($n = 129$). The median follow-up period was 6.47 years (IQR 5.58–7.83, range 0.003–8.86 years).

Patients in the sample were from 216 family doctor practices. 50 % of the practices had between 708 and 1823 eligible patients registered; only 5 % of practices had fewer than 276 patients or >3186. 69 % of the sample had at least one clinical predictor (as listed in Section 2.8.2) recorded in their records during follow up.

The average age of the sample was 45.32 (IQR 34.16, 61.03) years, 43.5 % were male, 73.3 % were White and 1.4 % of Black ethnicity. 21 % had missing ethnicity and 0.2 %% had missing deprivation data. Of those with deprivation data, approximately 18 % were from the fifth most deprived area while 14 % were from London practices (Table A). The overall risk of developing psychosis in the sample population was 45.8 (95 % CI 42.8, 49.0)/100,000 PYAR.

Psychosis was more common in non-white males, living in the NE or NW of England or London and the two most deprived quintiles, and in those consulting for any clinical predictor (current and historical) apart from smoking history and those rarely recorded, (Table A).

The relationship between diagnosis and age by sex (Fig. 2 and Appendix H) shows a steady decrease in risk with age among males and an initial decrease with age among females until age 30, followed by a plateau and a slight increase at older ages. There was evidence (Fig. 3) of non-proportionality in relative hazard for males and females, and for different age groups, at different follow up periods. We also found (Fig. 3) that the relative risk for males, vs females, and older vs younger

Table A
Sociodemographic characteristics of study population in the development dataset.

Variable	Development dataset (n = 299,993)			
	No Diagnosis (n = 299,163)		Diagnosis (n = 830)	
	Mean (SD)	Median IQR [25 % 75 %]	Mean (SD)	Median IQR [25 % 75 %]
Age at index date (years)	47.12 (18.47)	47.22 [34.18, 61.03]	43.52 (18.71)	41.35 [29.26, 56.17]
	N	**%**	**N**	**%**
Sex				
Male	129,995	43.45 %	412	49.64 %
Female	169,168	56.55 %	418	50.36 %
Missing	7 excluded from 300,000			
Race/ethnicity[a]				
Black	4218	1.41 % (1.79)	17	2.04 % (2.26)
White	219,280	73.3 % (93.15)	675	81.33 % (89.76)
Asian	7107	2.38 % (3.02)	28	3.37 % (3.72)
Mixed	1741	0.58 % (0.74)	11	1.33 % (1.46)
Other	3059	1.02 % (1.30)	21	2.53 % (2.79)
Missing	63,758	21.31 %	78	9.40 %
Deprivation Index				
I least deprived	67,743	22.64 %	130	15.66 %
II	59,833	20.00 %	118	14.22 %
III	60,636	20.27 %	136	16.39 %
IV	57,895	19.35 %	192	23.13 %
V most deprived	52,616	17.59 %	250	30.12 %
Missing	440	0.15 %	4	0.48 %
Geographical Location of family doctor practice				
North-East	5732	1.92 %	29	3.49 %
North-West	47,618	15.92 %	164	19.76 %
Yorkshire & The Humber	6270	2.10 %	11	1.33 %
West Midlands	33,855	11.32 %	99	11.93 %
East of England	23,373	7.81 %	51	6.14 %
South-West	33,487	11.19 %	94	11.33 %
South Central	44,457	14.86 %	92	11.08 %
South-East Coast	63,543	21.24 %	140	16.87 %
London	40,828	13.65 %	150	18.07 %
Earliest Symptom				
Depression/anxiety	92,730	31.0 %	481	57.95 %
Suicidal Behaviour	7056	2.36 %	118	14.22 %
Smoking problems	138,251	46.21 %	409	49.28 %
Cannabis problems	850	0.28 %	15	1.81 %
ADHD	1648	0.55 %	16	1.93 %
OCD	1314	0.44 %	7	0.84 %
Sleep disturbance	30,743	10.28 %	130	15.66 %
Mania	989	0.33 %	15	1.81 %
Social Isolation	437	0.15 %	6	0.72 %
Blunted Affect	34	0.01 %	0	0.00 %
Bizarre Behaviour	49	0.02 %	1	0.12 %
Role Functioning problems	3413	1.14 %	21	2.53 %
Substance Abuse	18,467	6.17 %	133	16.02 %
	Mean (SD)	**Median IQR [25 % 75 %]**	**Mean (SD)**	**Median IQR [25 % 75 %]**
Health Service Usage				
Number of consultations with study symptoms within study period	4.72 (5.41)	3 [2 6]	5.78 (8.41)	4 [2 7]
Number of consultations with study symptoms before study start (historical visits)	2.67 (5.68)	1 [0 3]	4.43 (7.56)	2 [0 6]
Total number of consultations	20.89 (34.89)	5 [2 26]	25.71 (40.35)	11 [3 32]
For any mental health problem within study period (not just the study symptoms)				
	N	**%**	**N**	**%**
Prescribed Psychotropic Medications Current				
Mixed Depression/ Anxiety	106,701	35.67 %	551	66.39 %
Mixed Depression/ PTSD/OCD	101,730	34.00 %	531	63.98 %
Sleep disturbances	39,940	13.35 %	222	26.75 %
Substance Abuse	53,099	17.75 %	204	24.58 %
History of Consultations for index symptoms before entering the study				
Bizarre behaviour	25	0.01 %	1	0.12 %
Suicidal behaviour	13,895	4.64 %	120	14.46 %
Cannabis-associated	865	0.29 %	16	1.93 %
Depressive symptoms	97,105	32.46 %	424	51.08 %
Blunted effect	79	0.03 %	0	0.00 %
ADHD-like sympt	1330	0.44 %	13	1.57 %
OCD-like sympt.	1848	0.62 %	15	1.81 %
Social isolation	644	0.22 %	12	1.45 %
Role functioning probs	3274	1.09 %	31	3.73 %
Mania	1281	0.43 %	11	1.33 %
Sleep disturbance	31,107	10.40 %	124	14.94 %
Smoking related	91,208	30.49 %	274	33.01 %
Substance abuse	12,753	4.26 %	119	14.34 %

Key: SD – Standard Deviation, IQR – inter-quartile range, PTSD-post traumatic stress disorder, OCD – obsessive compulsive disorder, ADHD attention deficit hyperactivity disorder.

[a] Ethnicity categories collapsed (White = White British and White Irish, Black = Black Caribbean, Black African, Black Other, Asian = Indian, Pakistani, Bangladeshi, Other Asian, Chinese, Mixed = White and Black Caribbean, What and Black African, White and Asian, Any other mixed background, Other = Any other ethnic group.

patients, differed with time of follow up, thus we stratified follow up time into 3 periods: < 2 years, 2–6 years and >6 years.

4.2. Model development

The original "full model" (Appendix I) was compared with that of the "reduced model" (Appendices J and K). The full model showed evidence of better fit, but there was no change in discriminatory power between the two models, with the C-statistic remaining at 0.774 after bootstrapping.

The final reduced model prognostic index (PI) equation is shown in Appendix L in terms of its coefficients and standard errors, and as a final equation in Appendix K. In the reduced model increased risk was associated with current and historical symptoms and/or diagnoses of suicidal behaviour, cannabis-related problems and substance abuse and current symptoms and/or diagnoses of mania, prescriptions for

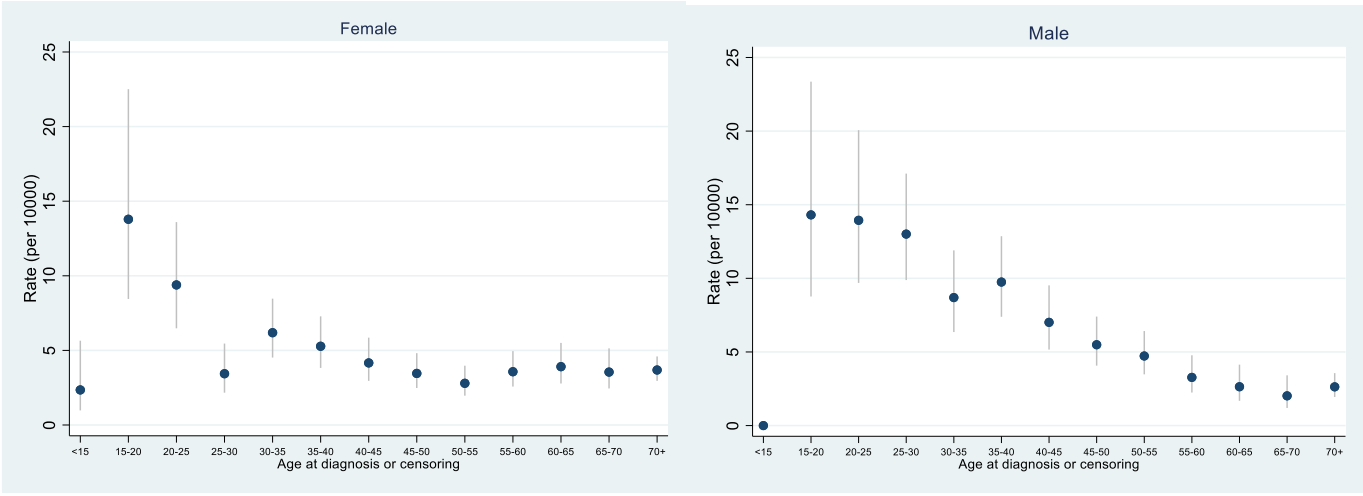

Fig. 2. Age specific incidence of psychosis, for females and males.

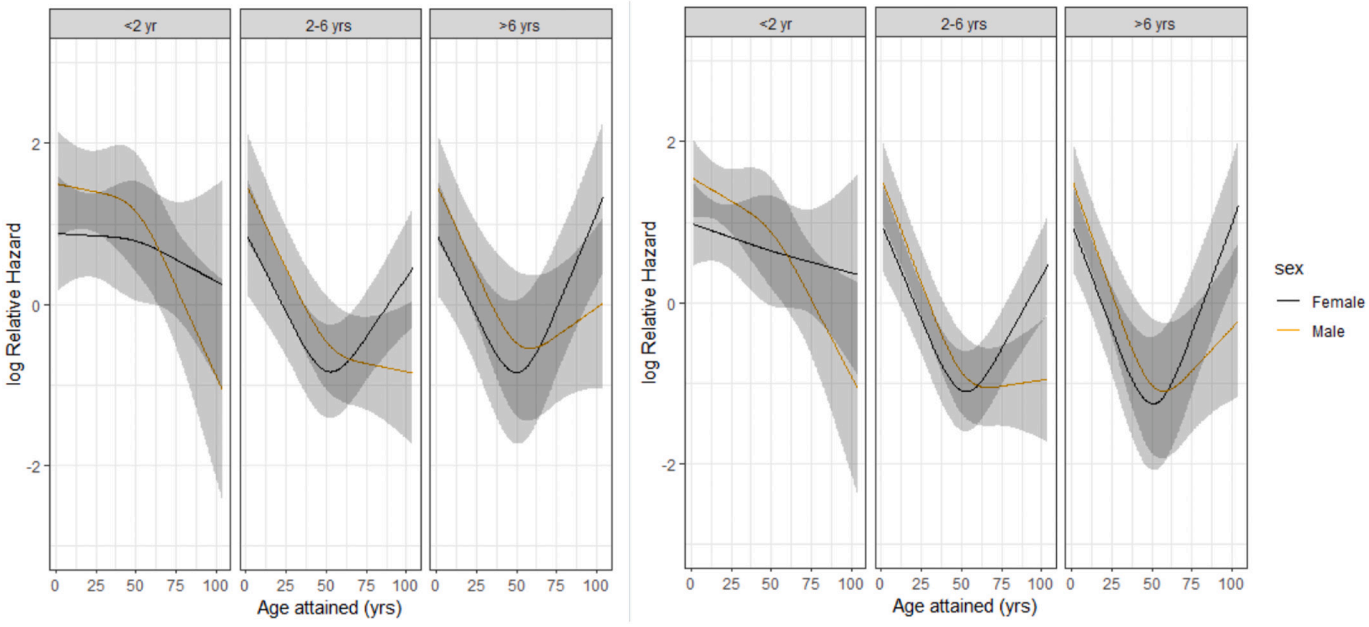

Fig. 3. Relative hazard by age group, for males and females separately, within different follow up periods.
left: Full Model (left panel) and reduced model (right panel).

depression or anxiety and sleep disturbance. Total number of previous consultations for any of the clinical predictors prior to the latest consultation was strongly related to risk (Appendix L). Those consulting with smoking history had a decreased risk, presumably in contrast to those who consulted with other mental health symptoms.

The final reduced model PI equation included demographic variables (age, sex, ethnicity, and deprivation) and current or historical consultations or diagnoses for suicidal behaviour, depression/anxiety, substance abuse and smoking and prescribed medications for depression, anxiety, PTSD, OCD and sleep disturbance and total number of consultations for the predictor symptoms. Table B shows hazard ratios for the factors and covariates, highlighting the strong association of some symptoms, notably consultations for suicidal behaviour (ideation, attempts and self-harm), as well as the increased risk associated with deprivation. After back transformation of two variables which had been transformed for use in the model, the hazard ratio for patients in practices where 1.22 % of patients identified as black was 1.22, compared with patients in practices where 0.62 % patients identified as black. Patients with 4 previous visits had 2.99 times the hazard of an event compared to those who had made no previous visits (see footnotes to Appendix L for full explanation). The strength of association between each component of the model and outcome are shown in Appendices L and M.

4.3. Model internal validation

The uncorrected C-statistic was 0.811 (0.812 when only patients aged 17–100 were included). Internal validation was through bootstrap resampling over 300 replications (White et al., 2011) (Appendix N). Model performance is shown in Table C, with correction for optimism using bootstrap techniques. The C-statistic was 0.774 after correction for bootstrapping over 8 years of follow up, although performance was slightly less good over 6 years (C = 0.742) and poor over 2 years (C = 0.544).

Table B
Reduced model: effects associated with individual predictor variables in the multivariable Cox proportional hazards regression analysis of risk of psychosis in the development dataset.

Predictor	Name of predictor in prediction algorithm	Units Change (lower to upper quartile for continuous variables)	Multivariable model Hazard Ratio (95 % CI)[a]
Age (years)	Age	38.0–64.4 years	0.96 (0.69, 1.33)
Sex	Male		
Male			1.17 (0.82, 1.68)
Female			1
Race/ethnicity			
% Black	tpblack	0.62 % - 1.82 %	1.22 (1.14, 1.30)
Deprivation index	IMD	Reference: 1st quintile (least deprived)	
2nd quintile			1.00 (0.78, 1.28)
3rd quintile			1.08 (0.85, 1.37)
4th quintile			1.40 (1.12, 1.76)
5th quintile			1.67 (1.34, 2.08)
Earliest symptom			
Suicidal behaviour	gpred2	0–1	2.48 (2.00, 3.07)
Depression/ anxiety	gpred4	0–1	1.38 (1.14, 1.65)
Substance abuse	gpred13	0–1	1.62 (1.32, 2.01)
Total number of previous consultations	trvisit	0–4 visits	2.99 (2.32, 3.85)
Prescribed psychotropic medications			
For depression/ anxiety/PTSD/ OCD	gpred4b	0–1	1.50 (1.15, 1.96)
For sleep disturbance	gpred11a	0–1	1.41 (1.20, 1.66)
History of consultations for predictor symptomS			
Suicidal behaviour	ghpred2	0–1	1.47 (1.19, 1.82)
Smoking problems	ghpred12	0–1	0.65 (0.55, 0.77)
Substance abuse	ghpred13	0–1	1.58 (1.26, 1.98)

Key: CI – confidence interval, PTSD post-traumatic stress disorder, OCD obsessive compulsive disorder.

[a] Hazard ratios for the each of the 4 variables which interact with other variables (sex, age, follow up time and prescription for depression, anxiety, PTSD or OCD) only apply at specific values of the other 3 variables. For example, the hazard ratio given for age is only true for a female patient, 2–6 years after entry to the cohort, with no prescription for depression/anxiety. Also, the hazard ratio given for males vs females is only true for a patient aged 51.04038 years, 2–6 years after entry to the cohort, with no prescription for depression/anxiety.

4.4. Sensitivity and specificity

Table D shows performance of the PI for different probability thresholds over 6 years of follow up. A patient with probability exceeding 1.0 % of developing psychosis carries a sensitivity of 71 %, specificity of 84 % and likelihood ratio of 4.49. Sensitivity and specificity varied across differing risk thresholds for probabilities of psychosis. Specificity increased with increasing risk thresholds and sensitivity decreased along with the likelihood ratio. The uncorrected C-statistic.

5. Discussion

We developed and internally validated a psychosis risk prediction model with a C-statistic of 0.774, indicating an almost 80 % probability that the risk score would be higher for someone who would develop psychosis than for someone who would not. This level of discrimination is considered good (Steyerberg et al., Jan 2010).

All predictors were positively associated with the outcome, except age and smoking history (which were inversely associated).

The findings suggest it is possible to accurately predict individual risk of a relatively rare mental health disorder using EHRs alone, without biomarkers. Although there is evidence that some biomarkers and neuroimaging tests can be useful for predicting risk, we have only examined predictors that are available in a family doctor clinical setting. A further important limitation for family doctors is the very high cost of accessing polygenic scores and neuroimaging tests.

P Risk is more accurate at detecting outcomes over periods longer than 2 years. This can be explained partly by our exclusion of all with a coded diagnosis of psychosis within a year of cohort entry to ensure incident cases and because by selecting a primary care population we are attempting to predict risk at a much earlier stage than has been the case with previous secondary care studies. Acute onset psychosis is rarer than that with a gradual onset (Mason et al., 2004) and may have different risk predictors. However, the accuracy of P Risk for detecting incident psychosis at 6 years, or even later, presents a valuable opportunity for very early intervention. Using retrospective information from family doctor records offers a low-cost option for identifying those at-risk, prompting family doctors to hold regular reviews –so-called 'safety-netting'.

It is also worth noting that accuracy, in terms of sensitivity and specificity varies across different risk thresholds as expected, suggesting that further work remains to be done to find appropriate levels for primary care clinicians to act on.

5.1. Comparison to previous literature

The incidence rate of psychosis was 1.6 times higher than that in a UK general population sample (29.4 per 100,000 PYAR (London, n.d.)) but similar to that (Hardoon et al., 2013) in patients consulting a family doctor for any physical or mental health symptom, despite our efforts to enrich our sample by selecting participants who consulted for nonpsychotic mental health problems.

To our knowledge this is the first family doctor psychosis prediction tool. Tools already exist for secondary care, and general population screening. Most secondary care tools were developed on populations at high risk of psychosis, making them unsuitable for family doctors due to the lack of predictor information.

Our results are comparable with those of a secondary care psychosis risk prediction calculator (Fusar-Poli et al., 2017) (Harrell's C statistic = 0.8). This study reported that substance use disorders were not useful predictors, in contrast to our findings. This may represent differences in the study populations, or in recording substance abuse in primary and secondary care. P Risk performed at least as well as other prediction models (Studerus et al., 2017) designed to predict transition to psychosis from those at high risk.

5.2. Strengths and limitations

Strengths are a very large database with real-world secondary care diagnostic outcomes and a long follow up. Our results are generalisable because our sample included densely populated urban, and more sparsely populated rural areas.

A major advantage of P Risk is that the information is already stored in EHRs, thus potentially allowing automation, and cost and time-

Table C
Internal Validation over 300 bootstrap samples.

Follow up time	Performance Measures	A. Equation from original sample, applied to same sample	B. Equation from bootstrap sample, applied to bootstrap sample	C. Equation from bootstrap sample, applied to original sample	D. Optimism = B-C	Index corrected = A-D	Harrell C optimism corrected = 0.5* (Dxy + 1)
2 years	Dxy	0.0923	0.0941	0.0887	0.0055	0.0869	0.544
	Slope	1	1	0.977	0.0293	0.9707	
6 years	Dxy	0.4970	0.5036	0.4909	0.0126	0.4843	0.742
	Slope	1	1	0.9695	0.0305	0.9695	
8 years	Dxy	0.5606	0.5671	0.5535	0.0136	0.5471	0.774
	Slope	1	1	0.9724	0.0276	0.9724	
Whole follow up	Dxy	0.5606	0.5654	0.5534	0.0120	0.5486	0.774
	Slope	1	1	0.9744	0.0256	0.9744	

Key: Dxy -.

Table D
Specificity and sensitivity of prognostic index based on predicted probabilities of developing psychosis over a 6-year period, among patients aged 17–100 years at entry to cohort.

Predicted probability of psychosis[a]	Specificity (n = 184,895)	Sensitivity (n = 629)	Likelihood ratio
≥0.5 %	65.2	87.4	2.51
≥1.0 %	84.2	71.1	4.49
≥1.5 %	91.2	58.0	6.61
≥2.0 %	94.5	49.0	8.86

(629 events occurred before 6 years, 184,895 survived 6 years with no psychosis).

[a] 4 thresholds not mutually exclusive.

effectiveness. New data collection is expensive, requires clinician training and sometimes specialist input.

There are limitations to our study. Family doctors differ in recording styles. To mitigate against this we included coded medications because prescribing data are more accurately recorded than symptom data. We found wide geographical variation in psychosis incidence, which may not reflect true regional difference, but rather service provision differences. Not all our outcomes were recorded in secondary care. This may be because we only linked to one HES database and the diagnoses may have been made in a different secondary care service. Also, approximately 20 % of psychosis patients are not in contact with secondary care (Reilly et al., 2012).

Our cohort was older than ideal for detecting new cases of psychosis. Patients entered the cohort when they consulted their family doctor, and the older mean age reflects the fact that middle-aged people consult family doctors more often than younger people. Our study design requirement for at least 5 years of follow up data may have induced selection bias, by reducing the number of young people in our sample who are more geographically mobile. It may also have induced a bias against the selection of people of all ages who frequently change address. Instability of accommodation may be associated with serious mental illness. Age did not have much impact on the psychosis warning signs in this sample, but this may be the result of the generally older mean age of patients who consult family doctors.

To ensure we were measuring incidence as an outcome, we deliberately excluded people who received a diagnosis of psychosis less than a year after cohort entry.

The smoking history predictor is difficult to interpret because it included consultations for smoking cessation as well as difficulties with giving up smoking.

We have included sensitivity and specificity statistics to make our findings clinically meaningful however there are disadvantages to these analyses. Firstly, the calculations used the same dataset of patients used to develop the P Risk model. Secondly, some of the predictors used in the risk score emerged during follow up, rather than at the start.

5.3. Clinical implications

Diagnostic delay is a problem in psychosis and there is convincing evidence that duration of untreated psychosis, caused by this delay, is related to outcomes (Marshall et al., 2004). Here we present findings of a prediction algorithm that has the potential to reduce delays in the early stages of the psychosis care pathway, although the key to this is in the mechanism of implementation, which will be the basis of a future study. P Risk may also help family doctors stratify risk of psychosis by identifying a group at higher risk who would benefit from watchful waiting rather than an immediate referral, as well as a very high-risk group that should have a rapid referral.

Sensitivity and specificity of those identified to be at a 0.5 % predicted probability of developing psychosis was reasonably good at 74 % and 82 % respectively. However, due to the low incidence of psychosis the use of this cut-off in practice could result in the identification of many false positive cases. Nevertheless, the application of this algorithm by family doctors would flag up cases who could be observed over the next few years to allow for early intervention in cases where prodromal symptoms of psychosis are observed. In this respect, P Risk may be applied to clinical guidelines around psychosis, such as the Access and Waiting Time (AWT) Standard (England, 2016), which specifies that all Early Intervention for Psychosis Teams (EITs) in England must provide assessment for an At-Risk Mental State. This has proved difficult to provide in an environment of restricted resources. However, if those identified by P Risk are monitored until they develop early prodromal symptoms of psychosis and are referred by a family doctor at that point for an EIT assessment, more focussed services can be directed at those at greater risk. This would be very difficult to achieve using clinical judgement alone.

This could help secondary care services tailor treatment i.e. sequential testing (Schmidt et al., 2017), close in clinical monitoring, and psychological treatments targeting prevention such as Cognitive Behavioural Therapy for Psychosis (Clark et al., 2016).

6. Conclusions

The findings provide evidence that it is possible to develop an accurate psychosis risk prediction model using EHRs and demographic predictors. This is the first psychosis prediction algorithm for use by family doctor practices. Family doctors are the first port of call and hence form a central part of the psychosis care pathway. The algorithm used data already entered in the patient record and does not require the collection of any new data. This is likely to be cost effective, although this needs to be evaluated in a future study.

7. Future research

P Risk is currently being externally validated in a separate EHR dataset. Subsequently, an implementation and feasibility studies has

recently been funded to investigate the operationalisation of P Risk on clinical software systems.

The communication of risk to patients should be carefully undertaken. This is an important area that deserves future research.

Data sharing

Raw data should be requested from CPRD. All programming code and the data specification can be requested from the corresponding author.

Role of the funding source

The funders have had no influence on conducting the research, the statistical analyses or on the publication of this paper. This research was funded by the National Institute of Health Research National School for Primary Care Research Grant Ref No 416 and by the NIHR Biomedical Research Centre.

Declaration of competing interest

None of the authors have any declarations of interest to declare.

Acknowledgement

This research was funded by the National Institute for Health Research (NIHR), School for Primary Care Research and the NIHR Biomedical Research Centre. Funding covered time of some of the research team and data costs. All authors have had full access to study data and accept responsibility to submit the paper for publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.schres.2022.06.031.

References

Abel, G.A., Barclay, M.E., Payne, R.A., 2016. Adjusted indices of multiple deprivation to enable comparisons within and between constituent countries of the UK including an illustration using mortality rates. Bmj Open 6 (11), e012750. Nov 15.
Anderson, J.R., Cain, K.C., Gelber, R.D., Nov 1983. Analysis of survival by tumor response. J. Clin. Oncol. 1 (11), 710–719.
Benchimol, E.I., Smeeth, L., Guttmann, A., et al., Oct 2015. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. Plos Med 12 (10).
Chew-Graham, C., Slade, M., Montana, C., Stewart, M., Gask, L., 2007. A qualitative study of referral to community mental health teams in the UK: exploring the rhetoric and the reality. BMC Health Serv. Res. 7.
Clark, S.R., Baune, B.T., Schubert, K.O., et al., 2016. Prediction of transition from ultra-high risk to first-episode psychosis using a probabilistic model combining history, clinical assessment and fatty-acid biomarkers. Transl Psychiatry 6 (9), e897. Sep 20.
Efron, B., Tibshirani, R., 1994. An introduction to the bootstrap (Monographs on statistics and applied Probability). Chapman and Hall.
England, N., 2016. Implementing the Early Intervention in Psychosis Access and Waiting Time Standard.
Fusar-Poli, P., Rutigliano, G., Stahl, D., Davies, C., Bonoldi, I., Reilly, T., McGuire, P., 2017. Development and validation of a clinically based risk calculator for the transdiagnostic prediction of psychosis. JAMA Psychiatry 74 (5), 493–500.
Hardoon, S., Hayes, J.F., Blackburn, R., Petersen, I., Walters, K., Nazareth, I., DPJ, Osborn, 2013. Recording of severe mental illness in United Kingdom primary care, 2000-2010. Plos One 8 (12). Dec 12.
Harrell, F.E., 2015. Regression Modelling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis, 2nd ed. springer, New York.
Herrett, E., Gallagher, A.M., Bhaskaran, K., Forbes, H., Mathur, R., van Staa, T., Smeeth, L., Jun 2015. Data resource profile: clinical practice research datalink (CPRD). Int J Epidemiol 44 (3), 827–836.
Kirkbride, J.B., Hameed, Y., Ioannidis, K., et al., 2017. Ethic minority status, age at immigration and psychosis risk in rural environments: evidence from the SEPEA study. Schizophr. Bull. 43 (6), 1251–1261.
Lawless, J.F., Singhal, K., 1978. Efficient screening of nonnormal regression models. Biometrics 34, 318–327.
Lewis, J.D., Bilker, W.B., Weinstein, R.B., Strom, B.L., 2005. The relationship between time since registration and measured incidence rates in the general practice research database. Pharmacoepidemiol. Drug Saf. 14, 443–451.
London, U.C.. Psychiatric Mapping Translated into Innovations for Care (PsyMaptic). Available at. https://www.psymaptic.org/welcome/.
Lunsford-Avery, J.R., MK, LeBourgeois, Gupta, T., Mittal, V.A., May 2015. Actigraphic-measured sleep disturbance predicts increased positive symptoms in adolescents at ultra high-risk for psychosis: a longitudinal study. Schizophrenia Research 164 (1–3), 15–20.
Marshall, M., Lewis, S., Lockwood, A., Drake, R., Croudace, T., Jones, P.B., 2004. Systematic review of the association between duration of untreated psychosis and outcome in cohorts of first episode patients. Schizophr. Res. 70 (1).
Mason, O., Startup, M., Halpin, S., Schall, U., Conrad, A., Carr, V., 2004. Risk factors for transition to first episode psychosis among individuals with 'at-risk mental states'. Schizophr. Res. 71 (2–3), 227–237.
Moons, K.G.M., Altman, D.G., Reitsma, J.B., et al., 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann. Intern. Med. 162 (1), W1–W73.
Nazareth, I., King, M., Haines, A., Rangel, L., Myers, S., 1993. Accuracy of diagnosis of psychosis on general-practice computer-system. Br. Med. J. 307 (6895), 32–34.
Ogundimu, E.O., Altman, D.G., Collins, G.S., Aug 2016. Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol 76, 175–182.
Perkins, D.O., Olde Loohuis, L., Barbee, J., et al., 2020. Polygenic risk score contribution to psychosis prediction in a target population of persons at clinical high risk. Am. J. Psychiatry 177 (2), 155–163.
Platz, C., Umbricht, D.S., Cattapan-Ludewig, K., Dvorsky, D., Arbach, D., Brenner, H.D., Simon, A.E., Dec 2006. Help-seeking pathways in early psychosis. Soc Psych Psych Epid 41 (12), 967–974.
Reilly, S., Planner, C., Hann, M., Reeves, D., Nazareth, I., Lester, H., 2012. The role of primary care in service provision for people with severe mental illness in the United Kingdom. PLoS One 7 (5), e36468.
Schmidt, A., Cappucciati, M., Radua, J., et al., Mar 2017. Improving prognostic accuracy in subjects at clinical high risk for psychosis: systematic review of predictive models and meta-analytical sequential testing simulation. Schizophrenia Bull 43 (2), 375–388.
Simonsen, E., Friis, S., Haahr, U., et al., Jul 2007. Clinical epidemiologic first-episode psychosis: 1-year outcome and predictors. Acta Psychiat Scand 116 (1), 54–61.
Skeate, A., Jackson, C., Birchwood, M., Jones, C., Sep 2002. Duration of untreated psychosis and pathways to care in first-episode psychosis - Investigation of help-seeking behaviour in primary care. Brit J Psychiat 181, S73–S77.
Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J., Kattan, M.W., Jan 2010. Assessing the performance of prediction models a framework for traditional and novel measures. Epidemiology 21 (1), 128–138.
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.
Sullivan, S.A., Hamilton, W., Tilling, K., Redaniel, T., Moran, P., Lewis, G., 2018. Early signs and symptoms of psychosis within primary-care: a nested case-control study using electronic primary-care records. JAMA Netw. Open 1 (7), 26.
Walley, T., Mantgani, A., 1997. The UK general practice research database. Lancet 44 (3), 827–836.
White, I.R., Royston, P., Wood, A.M., 2011. Multiple imputation using chained equations: issues and guidance for practice. Stat. Med. 30 (4), 377–399.