BJPsych | The British Journal of Psychiatry (2024)
224, 55–65. doi: 10.1192/bjp.2023.141

Original Article

Algorithmic fairness in precision psychiatry: analysis of prediction models in individuals at clinical high risk for psychosis

Derya Şahin, Lana Kambeitz-Ilankovic, Stephen Wood, Dominic Dwyer, Rachel Upthegrove, Raimo Salokangas, Stefan Borgwardt, Paolo Brambilla, Eva Meisenzahl, Stephan Ruhrmann, Frauke Schultze-Lutter, Rebekka Lencer, Alessandro Bertolino, Christos Pantelis, Nikolaos Koutsouleris and Joseph Kambeitz, for the PRONIA Study Group

Background

Computational models offer promising potential for personalised treatment of psychiatric diseases. For their clinical deployment, fairness must be evaluated alongside accuracy. Fairness requires predictive models to not unfairly disadvantage specific demographic groups. Failure to assess model fairness prior to use risks perpetuating healthcare inequalities. Despite its importance, empirical investigation of fairness in predictive models for psychiatry remains scarce.

Aims

To evaluate fairness in prediction models for development of psychosis and functional outcome.

Method

Using data from the PRONIA study, we examined fairness in 13 published models for prediction of transition to psychosis (*n* = 11) and functional outcome (*n* = 2) in people at clinical high risk for psychosis or with recent-onset depression. Using accuracy equality, predictive parity, false-positive error rate balance and false-negative error rate balance, we evaluated relevant fairness aspects for the demographic attributes 'gender' and 'educational attainment' and compared them with the fairness of clinicians' judgements.

Results

Our findings indicate systematic bias towards assigning less favourable outcomes to individuals with lower educational attainment in both prediction models and clinicians' judgements, resulting in higher false-positive rates in 7 of 11 models for transition to psychosis. Interestingly, the bias patterns observed in algorithmic predictions were not significantly more pronounced than those in clinicians' predictions.

Conclusions

Educational bias was present in algorithmic and clinicians' predictions, assuming more favourable outcomes for individuals with higher educational level (years of education). This bias might lead to increased stigma and psychosocial burden in patients with lower educational attainment and suboptimal psychosis prevention in those with higher educational attainment.

Keywords:

Ethics; psychotic disorders/schizophrenia; schizophrenia; risk assessment; stigma and discrimination.

Copyright and usage

Precision psychiatry seeks to provide the right treatment to the right patient at the right time, using algorithms to predict disease trajectory or treatment response. Although promising for improving mental healthcare and intervention efficacy, reliance on predicted outcomes over current presentations raises potential hazards. As an example, justice is one of the four major principles of bioethics (the other three being beneficence, non-maleficence and autonomy) and it refers to a fair and equitable treatment of patients and distribution of resources.[1] Algorithmic fairness (also referred to as fairness in this paper) is a principle linked to justice in bioethics and it can be operationalised by a number of different metrics.[2] A prediction algorithm can be considered fair for a demographic attribute if the predictions do not systematically disadvantage any subgroups regarding the respective demographic attribute (including but not limited to gender, education, age, race, ethnicity and socioeconomic status). Evidence so far shows that biases in medical algorithms can lead to underdiagnosis and disparate allocation of health resources in different subpopulations.[3,4] To prevent the emergence, perpetuation or reinforcement of health disparities through the use of prediction models, fairness considerations should be incorporated into their development and implementation.

Furthermore, most medical decisions are made by clinicians alone, who can also be subject to biases.[5,6] When assessing fairness of prediction models, measuring biases in clinicians' judgements can serve as an important benchmark against which model fairness can be more pragmatically evaluated. For the ethical and fair implementation of precision psychiatry approaches, a comprehensive understanding of clinical decision-making including all agents involved in the process is crucial.

Psychosis risk states and prediction models

Psychotic disorders remain one of the most substantial contributors to global burden of disease.[7] Unfortunately, existing treatment options do not sufficiently alleviate clinical symptoms once psychosis is fully developed.[8] Thus, research efforts have focused on an indicated preventive approach to identify and treat people at clinical high risk (CHR) for psychosis and to avoid a deleterious disease course.[9,10] Robust evidence suggests that individuals at CHR have a probability of approximately 25% of developing psychosis within 2 to 3 years.[10] Moreover, they often already suffer from or develop psychiatric disorders other than psychosis,[11] experience cognitive impairment[12] or retain CHR symptoms.[11,13] Thus, individuals at CHR are a heterogeneous population regarding their disease trajectory.[14]

A multitude of models for prediction of psychosis in CHR populations have been developed in recent years.[15] Moreover, recent studies have explored the potential of machine learning to predict further important clinical end-points for people on the psychosis spectrum, such as the development of poor psychosocial

https://doi.org/10.1192/bjp.2023.141 Published online by Cambridge University Press

functioning.[16,17] To use such models in clinical practice, it is of utmost importance to ensure both high robustness and generalisability – a prerequisite that has only recently started to be addressed.[18,19] An equally important prerequisite for the deployment of prediction algorithms in clinical settings is their compliance with ethical principles. Although previous work has focused on the use of algorithms in clinical practice on a theoretical level, addressing issues such as trust,[20] privacy,[21] transparency[21,22] and fairness,[23] the field so far lacks empirical investigations into the ethics of prediction algorithms.

In this study, we made a detailed investigation of published algorithms which were generated to predict clinical outcomes related to psychosis in individuals at CHR for psychosis or with recent-onset depression – of which 11 were developed to predict transition to psychosis and two to predict psychosocial outcome. The predictions of these algorithms were investigated with respect to multiple fairness criteria and compared with predictions of clinicians.

Method

We evaluated the fairness of psychosis prediction models published in peer-reviewed journals on data from the PRONIA cohort (https://cordis.europa.eu/project/id/602152). PRONIA (Prognostic Tools for Early Psychosis Management) is an EU-funded, naturalistic, multisite study conducted in research sites across Europe and Australia. Over 1700 participants (patients at CHR for psychosis, with recent-onset psychosis or with recent-onset depression and healthy controls) underwent a comprehensive clinical assessment, neuropsychological testing, provided blood markers and underwent a multimodal neuroimaging protocol. All participants were followed-up for 18 months to assess multiple outcome parameters. Further demographic and clinical characteristics of the sample have been previously published.[17,24] Written informed consent was obtained from all participants. The study was registered with the German Clinical Trials Register (DRKS00005042) and approved by the local research ethics committees in each location.

We investigated algorithms for the prediction of two clinically highly relevant outcomes: transition to psychosis and development of poor psychosocial functioning. In a first step, previously published clinical prediction models for the transition to psychosis were tested in the PRONIA data-set as an independent test sample and predictions were evaluated with respect to fairness criteria explained below. In this analysis, we focused on a subset of six predictive models with an area under the curve (AUC) of 65% or higher based on a previous analysis[18] and also included a model that was validated independently.[19,25] In addition, we investigated fairness in prediction models for transition to psychosis based on further data modalities (clinical/neurocognitive, neuroimaging and genetic data as well as their combination) developed in the PRONIA study using data relating to CHR for psychosis and recent-onset depression.[24] In a second step, we analysed predictive models of social and role functioning outcomes in individuals at CHR for psychosis or with recent-onset depression, as published by Koutsouleris et al.[17] Social and role functioning were measured using the Global Functioning: Social Scale and the Global Functioning: Role Scale;[26] for the analyses in our study, scores >7 points were taken as indicators of good functioning, whereas ≤7 points indicated poor functioning (7 points indicating mild impairment in social or role functioning, 8 points indicating good social or role functioning: for further details see the Supplementary material, available online at https://dx.doi.org/10.1192/bjp.2023.141).[26,27] In all analyses, we used the performance of clinical raters as a pragmatic benchmark against which the relative benefits and hazards of model bias can be assessed.

Fairness was investigated with respect to two relevant sensitive attributes: gender and educational attainment. Education was binarised to higher and lower educational level using the median of years of education in the sample as a cut-off. For each prediction model, we calculated the accuracy, balanced accuracy, true positive rate, true negative rate, positive predictive value and negative predictive value for all sensitive attributes. Owing to the small number of participants of non-European ethnicity in the PRONIA sample, analysing fairness for the sensitive attributes of race and ethnicity was not possible.

We chose four fairness metrics that are relevant in the context of outcome prediction in people at CHR for psychosis. For each sensitive attribute, we tested whether all subgroups exhibit: (a) equal balanced accuracy (accuracy equality), (b) equal positive predictive value (predictive parity), (c) equal false-positive rate (false-positive error rate (FPER) balance) and (d) equal false-negative rate (false-negative error rate (FNER) balance).[2] Owing to class imbalance in outcome variables (transition to psychosis, poor functional outcome), we used balanced accuracy as a measure of accuracy equality. The performance measures (accuracy, true-positive rate (TPR), false-positive rate (FPR), false-negative rate (FNR), positive predictive value (PPV)) of one group were taken as the reference values of the respective demographic attributes. The fairness metrics were calculated as the ratio of each attribute's group metric to the reference group metric. Therefore, a value of 1 indicates absolute fairness, and the more a value deviates from 1, the more pronounced the disparities. Values between 0.8 and 1.25, indicating that a subgroup's metric values were at least 80% of the subgroup with the highest metric values, were considered as fulfilling the fairness criteria according to the so-called four-fifths rule.[28]

We used permutation testing to assess whether prediction models showed statistically significant deviations from fair predictions.[29] Fairness metrics were computed in 10 000 samples with randomly permuted sensitive attributes to create a null distribution. Separate permutation tests were performed for each fairness criterion. Throughout the study, a Bonferroni-corrected threshold of $\alpha < 0.05$ was considered significant.

All statistical analyses were performed in R (R-Studio Version 1.3.1093, R Version 4.0.4, for Windows).

Results

Table 1 summarises key characteristics of the PRONIA sample used for the analyses. Table 2 lists all performance and fairness metrics for all models and sensitive attributes. There were no significant differences in psychosocial outcome or rates of transition to psychosis between male and female participants. Transition rates in participants with higher and lower educational level at baseline did not significantly differ either. However, participants with lower educational level showed a significantly higher rate of poor outcome in psychosocial functioning (CHR sample: poor role functioning in 54% of participants with higher educational level versus 75% in those with lower educational level ($\chi^2 = 7.737$, $P = 0.003$); poor social functioning in 45% of participants with higher educational level versus 69% in those with lower educational level ($\chi^2 = 10.316$, $P < 0.001$)).

The investigated algorithms were often not within the predefined permissible fairness range for the sensitive attributes of gender (27 out of 64 criteria) and education (34 out of 64 criteria). For the four fairness criteria that were analysed according to the two sensitive attributes on 13 algorithms and clinicians' predictions, statistically significant fairness violations emerged for FPER in only three models and FNER in only one model. These significant deviations all emerged for the sensitive attribute education and coherently consisted in assigning more favourable outcomes to

https://doi.org/10.1192/bjp.2023.141 Published online by Cambridge University Press

https://doi.org/10.1192/bjp.2023.141 Published online by Cambridge University Press

Table 1 Key characteristics of the study sample

			Sensitive attribute								
			Gender					Education			
	n	Overall	Female	Male	t or χ^2	*p*	*n*	Higher	Lower	t/χ^2	*p*
CHR only sample											
Participants, *n*		224	119	105				70	150		
Age, years: mean (s.d.)	224	23.73 (5.31)	23.21 (5.37)	24.31 (5.20)	−1.555[a]	0.121[a]	220	26.55 (5.11)	22.35 (4.72)	5.816[a]	<0.001[a]*
Education years, mean (s.d.)	220	13.60 (2.74)	13.68 (2.74)	13.50 (2.75)	0.470[a]	0.639[a]	220	16.69 (1.81)	12.16 (1.72)	17.580[a]	<0.001[a]*
Transition to psychosis, *n* (%)	198	24 (12%)	9 (9%)	15 (16%)	1.692[b]	0.129[b]	194	7 (11%)	17 (13%)	0.037[b]	0.670[b]
High role function, *n* (%)[c]	202	65 (32%)	38 (36%)	27 (28%)	0.858[b]	0.281[b]	199	30 (46%)	34 (25%)	7.737[b]	0.003[b]*
High social function, *n* (%)[c]	202	77 (38%)	43 (40%)	34 (36%)	0.247[b]	0.521[b]	199	36 (55%)	41 (31%)	10.316[b]	<0.001[b]*
SIPS Positive Symptoms, mean (s.d.)	223	1.65 (0.90)	1.67 (0.87)	1.62 (0.94)	0.427[a]	0.670[a]	220	1.43 (0.89)	1.76 (0.90)	−2.497[a]	0.014[a]*
SIPS Negative Symptoms, mean (s.d.)	221	1.73 (1.14)	1.64 (1.09)	1.83 (1.19)	−1.250[a]	0.213[a]	220	1.50 (1.11)	1.84 (1.14)	−2.071[a]	0.040[a]
SIPS Disorganised Symptoms, mean (s.d.)	221	0.85 (0.73)	0.81 (0.57)	0.90 (0.87)	−0.941[a]	0.348[a]	220	0.75 (0.67)	0.90 (0.76)	−1.529[a]	0.128[a]
SIPS General Psychopathology, mean (s.d.)	221	1.98 (0.96)	2.03 (0.88)	1.92 (1.06)	0.827[a]	0.409[a]	220	1.86 (0.99)	2.03 (0.95)	−1.245[a]	0.215[a]
Gender	224						220			0.100[b]	0.752[b]
Female, *n* (%)	119							38 (54%)	78 (52%)		
Male, *n* (%)	105							32 (46%)	72 (48%)		
CHR and recent-onset depression combined sample											
Participants, *n*		393	206	187				154	234		
Age, years: mean (s.d.)	393	24.53 (5.75)	24.33 (5.89)	24.75 (5.58)	−0.724[a]	0.470[a]	388	27.14 (5.25)	22.81 (5.34)	7.911[a]	<0.001[a]*
Education years, mean (s.d.)	388	14.12 (2.89)	14.15 (2.82)	14.08 (2.96)	0.222[a]	0.824[a]	388	16.94 (1.99)	12.26 (1.58)	24.540[a]	<0.001[a]*
Transition to psychosis, *n* (%)	366	27 (7%)	10 (5%)	17 (10%)	1.981[a]	0.108[b]	361	8 (5.4%)	19 (8.9%)	1.092[b]	0.212[b]
High role function, *n* (%)[c]	205	65 (32%)	38 (35%)	27 (28%)	0.958[b]	0.259[b]	202	30 (46%)	34 (25%)	8.312[b]	0.002[b]*
High social function, *n* (%)[c]	205	77 (38%)	43 (40%)	34 (35%)	0.312[b]	0.482[b]	202	36 (55%)	41 (30%)	11.057[b]	<0.001[b]*
SIPS Positive Symptoms, mean (s.d.)	390	1.13 (0.95)	1.14 (0.95)	1.12 (0.96)	0.228[a]	0.820[a]	386	0.86 (0.84)	1.31 (0.99)	−4.726[a]	<0.001[a]*
SIPS Negative Symptoms, mean (s.d.)	387	1.65 (1.07)	1.50 (1.01)	1.81 (1.12)	−2.810[a]	0.005[a]*	386	1.45 (1.00)	1.78 (1.11)	−2.959[a]	0.003[a]*
SIPS General Psychopathology, mean (s.d.)	387	0.73 (0.69)	0.67 (0.53)	0.79 (0.82)	0.480[a]	0.080[a]	386	0.59 (0.56)	0.83 (0.75)	−3.607[a]	<0.001[a]*
SIPS Disorganised Symptoms, mean (s.d.)	387	1.94 (0.96)	1.97 (0.89)	1.92 (1.03)	−1.757[a]	0.632[a]	386	1.87 (0.95)	1.99 (0.97)	−1.257[a]	0.210[a]
Gender	393						388			0.001[b]	0.971[b]
Female, *n* (%)	206							80 (52)	122 (52)		
Male, *n* (%)	187							74 (48)	112 (48)		

CHR, clinical high risk for psychosis; SIPS, Structured Interview for Psychosis-Risk Syndromes.
a. Welch two-sample *t*-test.
b. Pearson's χ^2-test.
c. Global Functioning: Social Scale and Global Functioning: Role Scale; cut-off: 7.
* Significance level $p < 0.05$.

https://doi.org/10.1192/bjp.2023.141 Published online by Cambridge University Press

Table 2 Performance matrices and fairness indices of psychosis transition and functional outcome prediction algorithms

Model	TP	TN	FP	FN	ACC	BAC	TPR	TNR	PPV	NPV	Attribute	Accuracy equality	Predictive parity	FPER balance	FNER balance
Transition to psychosis															
Clinicians' ratings (*n* = 334) (CHR & recent-onset depression)	16	261	49	8	82.5%	75.4%	66.7%	84.2%	24.6%	97.0%					
Female (*n* = 175)	6	145	21	3	86.0%	77.0%	66.7%	87.3%	22.2%	98.0%	Gender	1.030 (0.876)	0.844 (0.736)	0.651 (0.197)	1.000 (0.970)
Male (*n* = 159)	10	116	28	5	78.6%	73.6%	66.7%	80.6%	26.3%	95.9%					
High education (*n* = 129)	5	108	14	2	87.1%	80.0%	71.4%	88.5%	26.3%	98.2%	Education	0.911 (0.645)	0.909 (0.822)	1.630 (0.038)	1.240 (0.704)
Low education (*n* = 204)	11	152	35	6	79.7%	73.0%	64.7%	81.3%	23.9%	96.2%					
Clinicians' ratings (*n* = 198) (CHR only)	16	126	42	6	74.4%	73.9%	72.7%	75.0%	27.6%	95.5%					
Female (*n* = 175)	6	71	17	2	79.8%	77.8%	75.0%	80.7%	26.1%	97.3%	Gender	1.090 (0.638)	0.913 (0.842)	0.618 (0.897)	0.875 (0.148)
Male (*n* = 159)	10	55	25	4	69.0%	70.1%	71.4%	68.8%	28.6%	93.2%					
High education (*n* = 129)	5	46	11	2	79.7%	76.1%	71.4%	80.7%	31.2%	95.8%	Education	0.977 (0.905)	0.838 (0.718)	1.460 (0.108)	0.933 (0.959)
Low education (*n* = 204)	11	79	31	4	71.7%	72.6%	73.3%	71.8%	26.2%	95.2%					
Model based on clinical data (*n* = 331)	21	200	105	5	66.6%	73.2%	80.8%	65.6%	15.3%	97.8%					
Female (*n* = 171)	9	110	51	1	69.4%	79.1%	90.0%	68.3%	13.8%	99.2%	Gender	1.170 (0.247)	0.809 (0.655)	0.846 (0.350)	0.400 (0.426)
Male (*n* = 163)	12	92	55	4	63.7%	68.8%	75.0%	62.6%	17.2%	96.0%					
High education (*n* = 139)	7	94	37	1	72.6%	79.6%	87.5%	71.8%	15.1%	99.0%	Education	0.877 (0.399)	1.050 (0.907)	1.390 (0.023)	1.780 (0.341)
Low education (*n* = 189)	14	104	67	4	62.2%	69.3%	77.8%	60.8%	15.8%	96.7%					
Model based on MRI (*n* = 326)	21	183	118	4	62.3%	72.4%	84.0%	60.8%	13.7%	98.1%					
Female (*n* = 169)	9	93	66	1	60.1%	74.2%	90.0%	58.5%	10.9%	99.0%	Gender	1.070 (0.616)	0.628 (0.342)	1.120 (0.388)	0.500 (0.630)
Male (*n* = 157)	12	90	52	3	64.7%	71.7%	80.0%	63.4%	17.7%	97.0%					
High education (*n* = 138)	7	80	50	1	62.9%	74.5%	87.5%	61.5%	11.5%	98.8%	Education	0.955 (0.770)	1.390 (0.357)	1.010 (0.945)	1.410 (0.713)
Low education (*n* = 182)	14	101	64	3	62.8%	71.8%	82.4%	61.2%	16.0%	97.5%					
Model based on PRS (*n* = 298)	23	152	121	2	58.1%	73.8%	92.0%	55.7%	13.4%	98.9%					
Female (*n* = 150)	10	69	71	0	51.9%	74.6%	100.0%	49.3%	10.1%	100.0%	Gender	1.060 (0.572)	0.544 (0.234)	1.360 (0.016)	NA (0.515)
Male (*n* = 148)	13	83	50	2	64.4%	74.5%	86.7%	62.4%	18.6%	97.9%					
High education (*n* = 129)	7	62	58	1	53.6%	69.6%	87.5%	51.7%	9.5%	98.6%	Education	1.090 (0.437)	1.810 (0.076)	0.859 (0.319)	0.471 (0.599)
Low education (*n* = 164)	16	86	61	1	61.2%	76.3%	94.1%	58.5%	17.1%	99.1%					
Stacked model based on clinical, MRI and PRS data (*n* = 334)	19	270	38	7	86.6%	80.4%	73.1%	87.7%	31.3%	97.7%					
Female (*n* = 171)	8	141	20	2	87.1%	83.8%	80.0%	87.6%	26.6%	98.7%	Gender	1.100 (0.550)	0.717 (0.463)	1.030 (0.901)	0.640 (0.561)
Male (*n* = 163)	11	129	18	5	85.9%	78.2%	68.7%	87.7%	36.8%	96.4%					
High education (*n* = 139)	7	121	10	1	92.1%	89.9%	87.5%	92.4%	39.7%	99.2%	Education	0.811 (0.265)	0.700 (0.458)	2.150 (0.013)	2.670 (0.117)
Low education (*n* = 189)	12	143	28	6	82.1%	75.1%	66.7%	83.6%	27.8%	96.4%					
NAPLS (*n* = 165)	18	71	74	2	53.7%	69.5%	90.0%	49.0%	19.4%	97.3%					
Female (*n* = 83)	5	37	39	2	50.5%	60.1%	71.4%	48.7%	11.0%	95.0%	Gender	0.768 (0.064)	0.403 (0.177)	1.010 (0.973)	NA (0.254)
Male (*n* = 82)	13	34	35	0	57.0%	74.6%	100.0%	49.3%	27.3%	100.0%					
High education (*n* = 56)	3	35	17	1	67.8%	71.2%	75.0%	67.3%	15.2%	97.2%	Education	1.040 (0.749)	1.400 (0.364)	1.850* (<0.001)	0.250 (0.533)
Low education (*n* = 107)	15	36	55	1	47.3%	66.7%	93.8%	39.6%	21.3%	97.3%					
Hengartner (*n* = 167)	20	65	81	1	50.7%	69.9%	95.2%	44.5%	19.6%	98.5%					
Female (*n* = 85)	7	32	45	1	45.7%	64.5%	87.5%	41.6%	13.1%	97.1%	Gender	0.874 (0.129)	0.488 (0.205)	1.120 (0.452)	NA (Inf) (0.520)
Male (*n* = 82)	13	33	36	0	55.8%	73.9%	100.0%	47.8%	26.8%	100.0%					
High education (*n* = 57)	4	27	25	1	54.4%	66.0%	80.0%	51.9%	13.8%	96.4%	Education	1.140 (0.131)	1.650 (0.148)	1.230 (0.124)	NA (Inf) (0.982)
Low education (*n* = 107)	16	37	54	0	49.2%	70.3%	100.0%	40.7%	22.7%	100.0%					
Lencz (*n* = 149)	12	103	29	5	76.9%	74.3%	70.6%	78.0%	27.5%	95.7%					
Female (*n* = 76)	4	57	13	2	80.2%	74.0%	66.7%	81.4%	21.4%	97.0%	Gender	0.977 (0.909)	0.670 (0.531)	0.730 (0.410)	1.220 (0.797)
Male (*n* = 73)	8	46	16	3	73.4%	73.5%	72.7%	74.2%	32.2%	94.2%					
High education (*n* = 50)	2	40	6	2	84.1%	68.5%	50.0%	87.0%	23.3%	95.6%	Education	1.210 (0.346)	1.250 (0.647)	2.120 (0.034)	0.462 (0.478)
Low education (*n* = 96)	10	60	23	3	72.3%	74.6%	76.9%	72.3%	29.1%	95.5%					

(*Continued*)

Table 2 (*Continued*)

Model	TP	TN	FP	FN	ACC	BAC	TPR	TNR	PPV	NPV	Attribute	Accuracy equality	Predictive parity	FPER balance	FNER balance
Malda (*n* = 172)	13	108	41	10	70.4%	64.5%	56.5%	72.5%	24.0%	91.6%					
Female (*n* = 87)	6	58	21	2	73.6%	74.2%	75.0%	73.4%	22.0%	96.7%	Gender	1.360 (0.148)	0.850 (0.768)	0.930 (0.805)	0.469 (0.318)
Male (*n* = 85)	7	50	20	8	67.1%	59.0%	46.7%	71.4%	25.9%	86.2%					
High education (*n* = 57)	4	40	12	1	77.2%	78.5%	80.0%	76.9%	25.0%	97.6%	Education	0.714 (0.234)	0.947 (0.914)	1.340 (0.200)	2.500 (0.061)
Low education (*n* = 112)	9	65	29	9	66.1%	59.6%	50.0%	69.1%	23.7%	87.8%					
Metzler (*n* = 164)	18	69	74	3	52.8%	67.0%	85.7%	48.3%	19.1%	96.0%					
Female (*n* = 83)	6	38	37	2	52.9%	62.8%	75.0%	50.7%	13.3%	95.3%	Gender	0.872 (0.446)	0.542 (0.310)	0.907 (0.568)	3.250 (0.186)
Male (*n* = 81)	12	31	37	1	52.7%	68.9%	92.3%	45.6%	24.5%	96.9%					
High education (*n* = 56)	4	32	19	1	64.3%	71.4%	80.0%	62.7%	17.1%	97.0%	Education	0.962 (0.812)	1.180 (0.700)	1.630* (0.003)	0.625 (0.771)
Low education (*n* = 105)	14	35	54	2	46.2%	63.4%	87.5%	39.3%	20.1%	94.7%					
Michel (*n* = 163)	17	60	83	3	46.9%	63.5%	85.0%	42.0%	16.7%	95.3%					
Female (*n* = 81)	6	34	40	1	49.1%	65.8%	85.7%	45.9%	12.4%	97.3%	Gender	1.050 (0.775)	0.602 (0.350)	0.867 (0.351)	0.929 (0.984)
Male (*n* = 82)	11	26	43	2	44.9%	61.1%	84.6%	37.7%	20.6%	92.8%					
High education (*n* = 56)	3	23	29	1	46.4%	59.6%	75.0%	44.2%	9.5%	95.8%	Education	1.110 (0.532)	2.150 (0.075)	1.070 (0.651)	0.500 (0.653)
Low education (*n* = 105)	14	36	53	2	47.2%	64.0%	87.5%	40.4%	20.4%	94.9%					
Walder (*n* = 168)	13	113	32	10	75.1%	67.2%	56.5%	77.9%	28.1%	92.2%					
Female (*n* = 83)	6	53	22	2	71.0%	72.8%	75.0%	70.7%	20.1%	96.6%	Gender	1.230 (0.310)	0.487 (0.270)	2.110 (0.016)	0.469 (0.276)
Male (*n* = 85)	7	60	10	8	78.8%	66.2%	46.7%	85.7%	41.2%	88.2%					
High education (*n* = 55)	3	44	6	2	85.5%	74.0%	60.0%	88.0%	32.5%	95.8%	Education	0.876 (0.623)	0.831 (0.746)	2.410* (0.004)	1.110 (0.844)
Low education (*n* = 108)	10	64	26	8	68.5%	63.3%	55.6%	71.1%	27.1%	89.2%					
Functional outcome															
Machine learning: role functioning (*n* = 205)	38	111	29	27	72.3%	68.4%	57.8%	79.0%	56.1%	80.1%					
Female (*n* = 108)	23	51	19	15	68.5%	66.7%	60.5%	72.9%	54.8%	77.3%	Gender	0.985 (0.921)	0.913 (0.659)	1.900 (0.022)	0.888 (0.706)
Male (*n* = 97)	15	60	10	12	77.3%	70.6%	55.6%	85.7%	60.0%	83.3%					
High education (*n* = 91)	22	40	13	16	68.1%	66.7%	57.9%	75.5%	62.9%	71.4%	Education	1.030 (0.836)	0.770 (0.328)	0.767 (0.510)	1.000 (0.993)
Low education (*n* = 111)	15	69	16	11	75.7%	69.4%	57.7%	81.2%	48.4%	86.3%					
Clinicians' rating: role functioning (*n* = 202)	58	71	68	5	64.0%	71.8%	92.1%	51.4%	46.4%	93.4%					
Female (*n* = 105)	31	34	35	5	61.1%	67.7%	86.1%	49.3%	47.0%	87.2%	Gender	0.876 (0.063)	1.040 (0.825)	1.080 (0.668)	NA (0.040)
Male (*n* = 97)	27	37	33	0	66.0%	76.4%	100.0%	52.9%	45.0%	100.0%					
High education (*n* = 91)	36	25	28	2	67.0%	71.0%	94.7%	47.2%	56.3%	92.6%	Education	0.972 (0.699)	0.641 (0.072)	0.868 (0.448)	2.280 (0.248)
Low education (*n* = 110)	22	46	39	3	61.6%	71.1%	88.0%	54.1%	36.1%	93.9%					
Machine learning: social functioning (*n* = 205)	65	95	33	12	78.7%	79.8%	84.4%	75.2%	67.7%	88.7%					
Female (*n* = 108)	38	46	19	5	77.8%	79.6%	88.4%	70.8%	66.7%	90.2%	Gender	1.050 (0.547)	1.010 (0.886)	1.320 (0.280)	0.565 (0.428)
Male (*n* = 97)	27	49	14	7	78.4%	78.6%	79.4%	77.8%	65.9%	87.5%					
High education (*n* = 91)	43	34	12	2	84.6%	84.7%	95.6%	73.9%	78.2%	94.4%	Education	0.805 (0.021)	0.686 (0.055)	0.922 (0.801)	7.030* (0.006)
Low education (*n* = 111)	22	60	19	10	73.9%	72.3%	68.8%	75.9%	53.7%	85.7%					
Clinicians' rating: social functioning (*n* = 202)	67	67	59	9	66.5%	70.9%	88.2%	53.6%	53.6%	88.2%					
Female (*n* = 105)	36	33	30	6	64.8%	69.0%	85.7%	52.4%	54.5%	84.6%	Gender	0.947 (0.532)	1.060 (0.757)	1.030 (0.878)	1.620 (0.310)
Male (*n* = 97)	31	34	29	3	67.0%	72.6%	91.2%	54.0%	51.7%	91.9%					
High education (*n* = 91)	42	24	22	3	72.5%	72.8%	93.3%	52.2%	65.6%	88.9%	Education	0.903 (0.233)	0.625 (0.044)	0.953 (0.795)	2.900 (0.059)
Low education (*n* = 110)	25	43	36	6	61.6%	67.5%	80.6%	54.4%	41.0%	87.8%					

TP, true positive; TN, true negative; FP, false positive; FN, false negative; ACC, accuracy; BAC, balanced accuracy (BAC); TPR, true-positive rate; TNR, true-negative rate; PPV, positive predictive value; NPV, negative predictive value; FPER, false-positive error rate; FNER, false-negative error rate; CHR, clinical high risk for psychosis; PRS, polygenic risk score; MRI, magnetic resonance imaging; NAPLS, North American Prodrome Longitudinal Study risk calculator.
* Bonferroni corrected *P* < 0.05. All fairness metrics are stated as metric value (*P*-value).

participants with higher educational level; we did not find any statistically significant fairness violations for gender.

All performance and fairness metrics are presented in Table 2. Figure 1 depicts fairness metrics for the sensitive attribute gender, Fig. 2 for the sensitive attribute education.

Fairness of models for the prediction of transition to psychosis

In a first analysis, we investigated six previously published prediction models for the transition to psychosis that showed the highest accuracy when applied to the PRONIA data-set as an independent test sample[18] and the North American Prodrome Longitudinal Study (NAPLS) risk calculator, which has been validated for the prediction of transition to psychosis in the PRONIA data-set.[19,25] Features used in prediction models can be found in Supplementary Table 1. Five out of seven examined models fulfilled the criteria accuracy equality for both gender and education, with accuracy equalities in the permissible range and no significant deviations of accuracy in any of the sensitive attributes. The four-fifths rule was violated by the prediction model of Malda et al (2019),[18] which had a higher accuracy for females and participants with higher educational level, and by the NAPLS risk calculator, which had a higher accuracy for males. On predictive parity, we observed a systematic deviation in all seven prediction algorithms and clinicians' predictions towards higher positive predictive values for males, although none reached statistical significance. All prediction models, including clinicians' predictions, showed a higher false-positive rate for participants with lower educational level, with three models showing statistically significant deviations (model by Metzler:[18] FPER = 1.630, $P = 0.003$; model by Walder:[18] FPER = 2.410, $P = 0.004$; NAPLS risk calculator:[25] FPER = 1.850, $P < 0.001$).

Overall, there was no systematic difference of bias between clinicians and algorithms.

In a second analysis, we investigated multimodal models for the prediction of transition to psychosis based on clinical, neuroimaging and genetic data and a stacked model based on the listed data modalities, which were previously developed in the PRONIA data-set. All four models fulfilled accuracy equality for both sensitive attributes, with a slight systematic deviation for higher balanced accuracies for females, whereas the positive predictive value was higher for males in all models and in clinicians' predictions. In

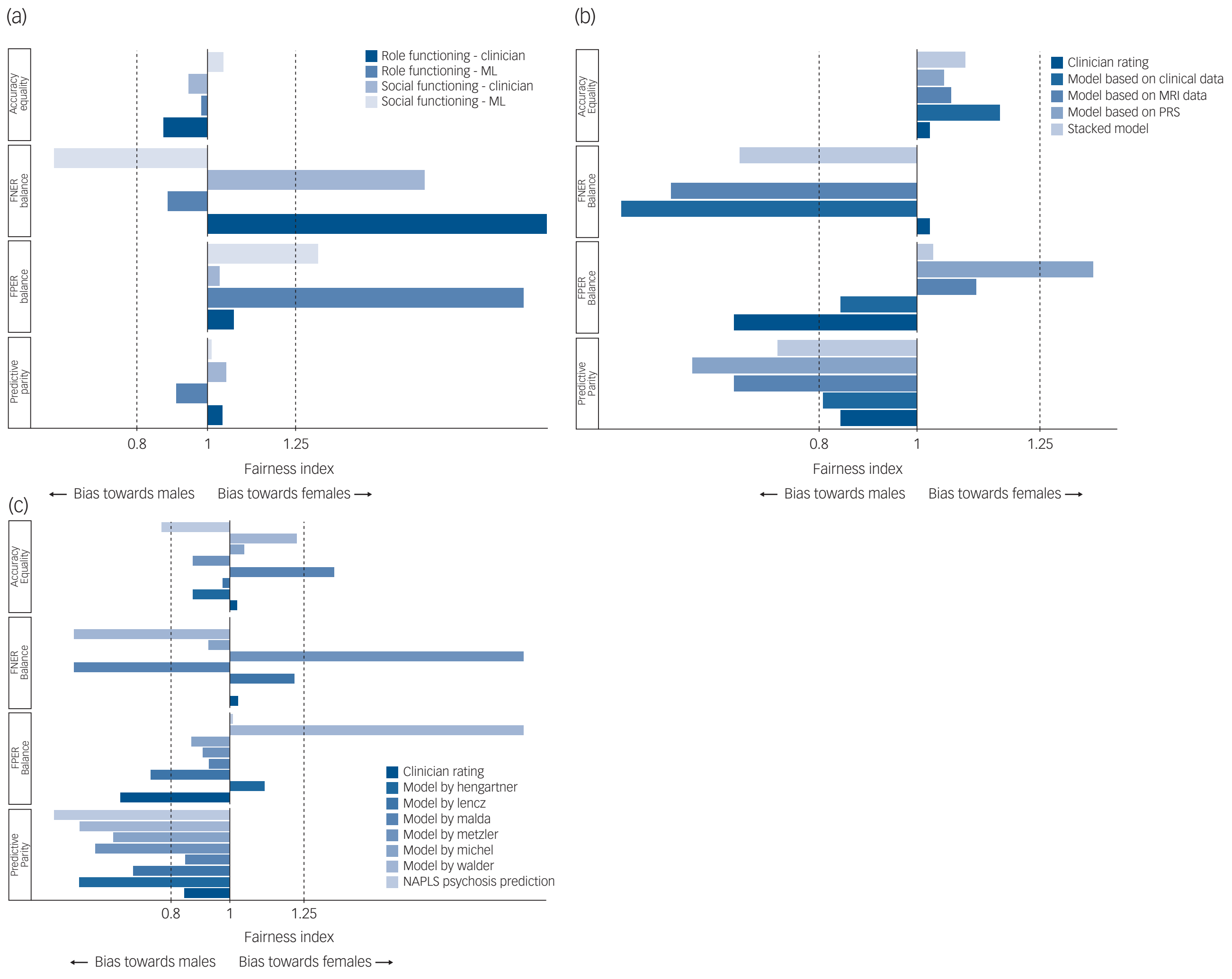

Fig. 1 Fairness of prediction models validated on PRONIA data for the sensitive attribute 'gender', with males as the reference group. (a) The fairness of predictions of functional outcome. (b) and (c) The fairness of predictions of transition to psychosis. The continuous line at $x = 1$ shows absolute fairness and the dashed lines at $x = 0.8$ and $x = 1.25$ cover the permissible fairness range according to the four-fifths rule. Values higher than 2 were replaced with $x = 2$ in the figures. The false-negative error rate (FNER) balance could not be calculated for the model by Hengartner, the North American Prodrome Longitudinal Study (NAPLS) risk calculator and polygenic risk score (PRS) model because there were 0 false negatives in the reference group. ML, machine learning; MRI, magnetic resonance imaging.

https://doi.org/10.1192/bjp.2023.141 Published online by Cambridge University Press

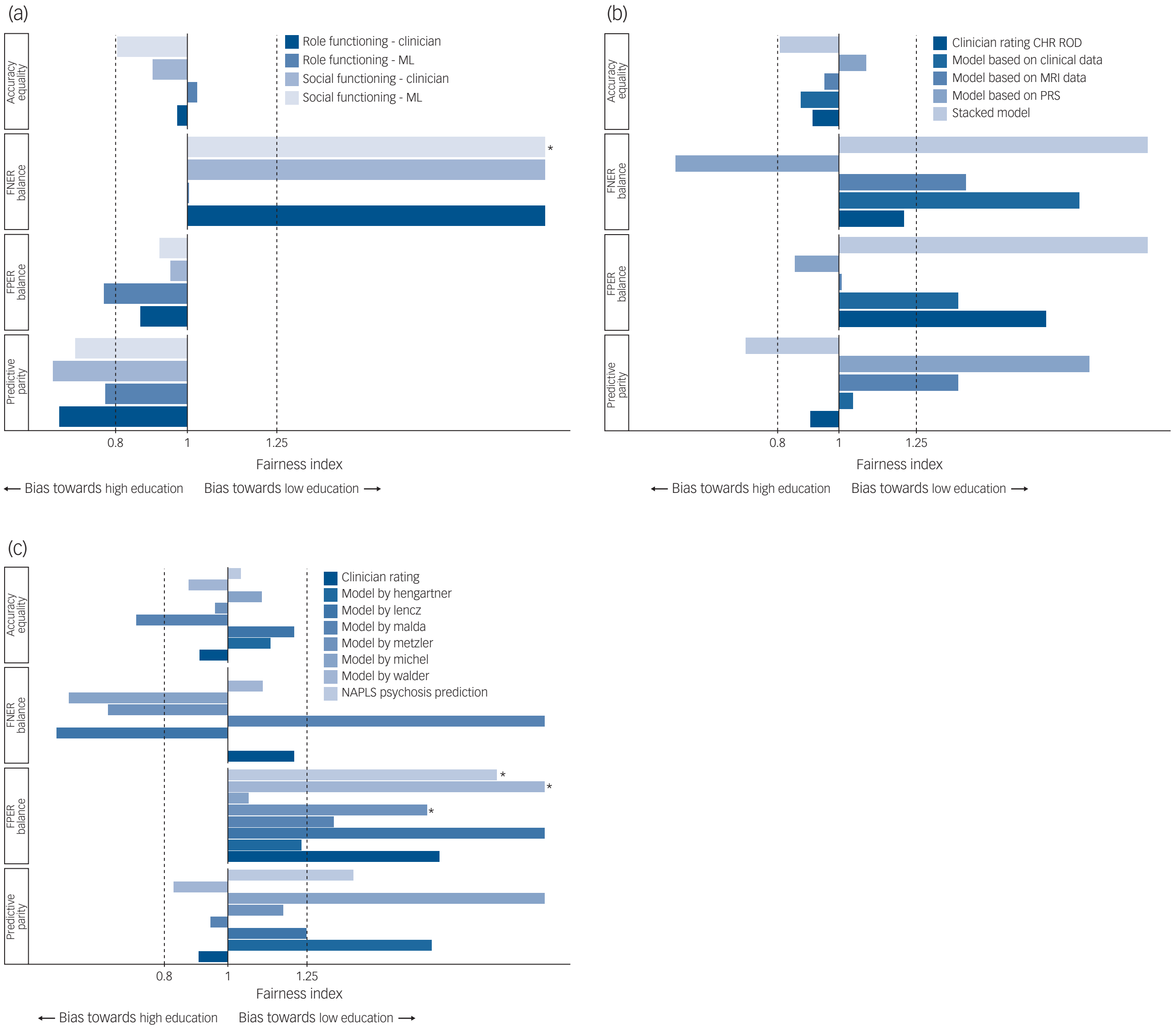

Fig. 2 Fairness of prediction models validated on PRONIA data for the sensitive attribute ‘education’, binarised high/low, with participants with a higher educational level as the reference group.
(a) The fairness of predictions of functional outcome. (b) and (c) The fairness of predictions of transition to psychosis. The continuous line at $x = 1$ shows absolute fairness and the dotted lines at $x = 0.8$ and at $x = 1.25$ cover the permissible fairness range according to the four-fifths rule. Values higher than 2 were replaced with $x = 2$ in the figures. The false-negative error rate (FNER) balance could not be calculated for the model by Hengartner and the North American Prodrome Longitudinal Study (NAPLS) risk calculator as there were 0 false negatives in the reference group. *Bonferroni corrected $P < 0.05$. ML, machine learning; CHR, clinical high risk for psychosis; ROD, recent-onset depression; MRI, magnetic resonance imaging.

comparison, clinicians’ ratings fulfilled accuracy equality, FNER balance and predictive parity, while violating FPER balance, with higher false-positive rates in males.

For education, all machine learning models and the clinicians’ ratings fulfilled accuracy equality. Positive predictive rate was lower for participants with higher educational level in clinical/neuropsychological, neuroimaging- and genetics-based models, whereas it was higher for those with higher educational level in the stacked model.

Fairness of models for the prediction of psychosocial functioning

In a fourth analysis we investigated fairness criteria in a model for the prediction of psychosocial outcomes which was previously developed in the PRONIA data-set.[17] There was no violation of accuracy equality for any of the tested sensitive attributes. Clinicians assigned more favourable outcomes to males, whereas machine learning-based predictions fulfilled three of four fairness criteria, violating only predictive equality. False-positive rates were higher in for females (role functioning: FPR = 27.1% for females versus 14.3% for males; social functioning: FPR = 29.2% for females versus 22.2% for males).

For education, neither the clinicians’ ratings nor the machine learning models fulfilled the four-fifths rule for all fairness criteria. Both clinicians’ and machine learning’s predictions had a higher positive predictive rate for participants with higher educational level for both role and social functioning. False-negative rates were higher for participants with lower educational level, indicating that the rate of those who were incorrectly assigned as poor outcome was higher in people with lower educational level (social functioning machine learning model: FNER = 7.030, $P = 0.006$). Algorithmic

https://doi.org/10.1192/bjp.2023.141 Published online by Cambridge University Press

predictions were not more systematically biased than clinicians' predictions.

Discussion

In the present study, we empirically investigated algorithmic fairness on a range of models for the prediction of outcome in people at clinical high risk for psychosis. A substantial number of investigated algorithms were not within the predefined permissible fairness range for the sensitive attributes of gender and years of education.

Educational bias in models

Overall, apart from algorithms based on neuroimaging data and polygenic risk score (PRS), there was a general tendency to predict more favourable outcomes for participants with higher educational level at baseline. In the majority of examined prediction models, participants with lower educational level were more often falsely predicted to have a transition to psychosis or a poor functional outcome than those with more education: in 6 out of 11 prediction models for transition to psychosis, both accuracy and balanced accuracy were higher for participants with higher educational level, whereas with the exception of the PRS-based prediction model, all prediction models for transition to psychosis had a higher FPR in participants with lower educational level. Clinicians similarly had a higher rate of false positives for participants with lower educational level. Given that the analysed models were masked to years of education, one explanation for these findings could be that they included features directly or indirectly associated with education years. Supporting this hypothesis, the three significantly unfair models included education-related features: the model by Metzler et al included verbal IQ,[18] the model by Walder et al included scholastic adjustment in childhood and functioning[18] and the NAPLS risk calculator included scores from two neuropsychological tests (the Hopkins Verbal Learning Test – Revised total raw score, testing verbal learning and memory, and the Brief Assessment of Cognition in Schizophrenia symbol coding raw score, testing processing speed).[25] Furthermore, the models based on polygenic risk score and neuroimaging data – two data modalities that are likely to be less associated with education years than neuropsychological data – were the only models without a bias towards participants with higher educational level: higher positive predictive values in participants with lower educational level were revealed in both the PRS and magnetic resonance imaging (MRI) models, and the PRS model was the only model with higher false-negative and false-positive rates as well as lower accuracy for participants with higher educational level. Another hypothetical explanation for the discrepancies in prognostic performance due to educational level is that medical research does not sufficiently represent people with lower educational level and the consequent lack of sufficient data leads to worse algorithmic performance.

Similar to algorithmic educational bias, clinicians also assigned more favourable outcomes to participants with higher educational level. Clinicians' educational bias is a phenomenon that has not been explored in this context before. There are studies showing bias in clinicians' behaviour in medicine, with the focus so far being on factors such as race, ethnicity, gender, age and weight (bias against obese patients).[30] Specifically, in psychiatry, research on bias-related misdiagnosis shows that previous experiences and certain prototypes might bias psychiatrists' clinical decision-making while simultaneously being a source of expertise.[31] In psychiatry, bias-related misdiagnoses have been examined for race,[31] sexual orientation[32] and ethnicity,[6] although education-related bias has not been explored to date. A bias associating more favourable outcomes to patients with higher educational level could be caused by confirmation bias of clinicians: because educational attainment is one of the most relevant environmental risk factors for psychosis,[33] clinicians might be more likely to associate higher educational level with lower risks of transition to psychosis or poor psychosocial outcome. To the best of our knowledge, our study is the first to show a tendency towards educational bias in psychiatry.

Gender-related bias

In all prediction models for transition to psychosis, positive predictive values were consistently higher for males, and in 9 out of 11 models for transition to psychosis, the deviation from absolute fairness in positive predictive value transgressed the predefined permissible range of fairness. The clinicians' predictions also had a higher positive predictive value for males, remaining in the permissible fairness range. Furthermore, clinicians more often falsely predicted poor functional outcome for females, whereas machine learning models more often falsely predicted good functional outcome for females. The male and female patients in the PRONIA data-set did not significantly differ in their key characteristics, including age, years of education and risk symptoms on the Structured Interview for Psychosis-Risk Syndromes (SIPS).

Clinician versus algorithmic bias

Our results suggest that algorithmic predictions were not systematically more biased than clinicians' predictions. Of 11 algorithms that predicted transition to psychosis, only 1 (Walder) was less fair than clinicians' predictions on all fairness criteria for the sensitive attribute gender, and only one (the stacked model) was less fair on all fairness criteria for the sensitive attribute education. Although clinicians were less fair regarding false-positive predictions for the sensitive attribute gender, they made fairer predictions from an overall accuracy point of view. Similarly, although clinicians were fairer regarding false-negative predictions for the sensitive attribute education, 6 out of 11 prediction models were fairer on overall accuracy. Thus, our data did not show any sign of generalised systematic bias of algorithms in comparison with clinicians' predictions.

Algorithmic fairness is a novel field of research in medicine and, so far, investigations of fairness of algorithmic predictions without benchmarking them to standard procedures found biased decisions disadvantageous to specific populations characterised by race,[4,34] gender[32] and age.[35] However, they did not quantify the amount of clinician bias: thus, it was not possible to conclude whether the prediction models led to more or less fairness compared with standard procedures. By comparing algorithmic and clinicians' predictions, we allow a framework for weighing the additional harm that can be caused by employing algorithms in predictive medicine. Based on our findings that the examined algorithms did not show stronger bias compared with clinicians' predictions, their clinical use might not pose an ethical conflict from a fairness perspective. We underline here that comparing an algorithm with the current standard (i.e. clinicians' judgements) might serve as a pragmatic benchmark for assessing the ethical permissibility of an algorithm in terms of fairness.

Fairness metrics and their relevance

Our study shows that the degree of fairness differs depending on the fairness metric. The algorithms we analysed all showed comparable and little bias in accuracy equality; however, their degree of fairness vacillated more in FPER and FNER balance. The priority that should be given to a fairness metric would depend on the predictive

https://doi.org/10.1192/bjp.2023.141 Published online by Cambridge University Press

task at hand. In our study, we focused on the prediction of transition to psychosis, in which case, a non-recognition (false-negative prediction) would have different consequences than overdiagnosis (false-positive prediction). Non-recognition would possibly deprive the mispredicted individual of necessary preventive measures that could delay the onset, speed up the diagnosis and ameliorate the course of disease. In this context, a higher educational level could present a double-edged sword, as it is associated with higher health literacy,[36] better health outcomes[37] and might be a protective factor against psychiatric disorders;[38] but at the same time, a bias associating more favourable outcomes to patients with higher educational level could lead to a higher risk of non-recognition of CHR states and to lack/delay of necessary preventive health measures. On the other hand, higher rates of overdiagnosis in patients with lower educational level could reinforce stigma against these patients, could harm them by overtreatment with medications with substantial side-effects (e.g. antipsychotics) but might simultaneously lead to better preventive care through more frequent follow-up visits and early recognition of psychosis symptoms.

Strengths

To the best of our knowledge, this study is the first to examine bias in psychosis prediction and the first to compare predictive bias between clinicians and algorithms in medicine. By testing the fairness of various predictive algorithms based on different data modalities, we provide evidence for a pattern of educational bias in psychosis prediction. None of the models we analysed included education years and only one model contained gender as a model feature (model by Malda et al); therefore, the investigated models mostly did not directly include the sensitive attributes they were analysed for, allowing an assessment of inherent biases incorporated in the models. Moreover, by analysing models that were developed using other data-sets than PRONIA, we could also assess bias independent from training data-sets, allowing a more realistic assessment of model fairness. The educational bias we detected was consistent in all models. By analysing models based on different data modalities, such as genetic, imaging and clinical data, we assessed whether model bias depends on data modalities used in model development and showed that models based on genetic and imaging data are less susceptible to educational bias. In addition, our study is the first in the literature to find educational bias in psychosis prediction.

Limitations

Our study has several limitations. First, the ethnic and racial homogeneity of our sample – consisting mostly of White European patients – limits the generalisability of our findings. Fairness analyses need to be replicated in samples with higher heterogeneity to control for ethnic and racial bias, especially since previous work hints at racial and ethnic bias in psychiatric diseases.[6,31] Second, our work compares bias in algorithms with bias in clinicians' predictions but cannot address the question of whether these biases are compounded in algorithm-supported decision-making processes in which predictions are not made by a clinician or an algorithm alone but algorithms are used to support clinicians' decisions. Third, the follow-up period of the patients presented in this study was limited to 18 months. Longer follow-ups can change the performance metrics of the presented models for transition to psychosis, as the number of patients developing psychosis will probably be higher over a longer follow-up period. Fourth, although we found evidence for an educational bias, educational status in our study was based on years of education and not on the type and level of education. However, even in a sample with a relatively high overall educational status, we found a clear tendency of predicting more favourable outcomes for those with more years of education. Fifth, in the absence of a consensus, we applied the four-fifths rule as an orientation for a permissible fairness range. However, we note that this range is arbitrary and can be adapted according to the fairness question at hand. Sixth, we analysed the algorithms according to the categories and cut-offs of the original studies, since changing categories or shifting cut-offs would require training the algorithms anew, which would result in new algorithms. Thus, although our research focused on assessing the fairness of existing algorithms and addressed the deployment phase of algorithms, future research should also investigate the development phase and control for effects of changing labels/categories as well as shifting cut-offs. Finally, although this study unravels a pattern of educational bias, it remains unclear whether the observed educational bias would translate to a disadvantage in mental healthcare for patients with higher educational level at clinical high risk for psychosis. Psychosis prediction algorithms are not widely used in clinical psychiatry and in order to investigate the consequences of such bias, data from clinical settings in which predictive algorithms are employed are necessary. Data from clinical settings would also allow researchers to determine whether there are compounding effects of clinicians' interaction with a biased algorithm.

Implications

Even though educational status was not directly included as a predictive feature in the algorithms we evaluated, the consistent pattern of educational bias found in our study highlights the fact that dissecting biases in algorithms requires comprehensive analyses and may not always be straightforward since fairness violations might be present for demographic attributes that are not included in the algorithmic decision-making process.

As predictive algorithms gain importance in medical practice, it is of paramount importance to ensure their compliance with ethical principles. The accuracy of clinical decisions based on predictions can only be tested at a later point, by which time the disparate allocation of resources based on algorithmic decision-making might already have resulted in disparities in healthcare, putting certain demographic groups at higher risk or depriving them of necessary preventive measures. Thus, fairness as a core principle of bioethics should be incorporated in the development of precision medicine approaches to avoid the emergence, perpetuation and reinforcement of health disparities.

Derya Şahin, Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital, University of Cologne, Cologne, Germany; **Lana Kambeitz-Ilankovic**, Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital, University of Cologne, Cologne, Germany; and Department of Psychology, Faculty of Psychology and Educational Sciences, Ludwig-Maximilian University, Munich, Germany; **Stephen Wood**, Centre for Youth Mental Health, University of Melbourne, Melbourne, Victoria, Australia; and Orygen, the National Centre of Excellence for Youth Mental Health, Melbourne, Victoria, Australia; **Dominic Dwyer**, Department of Psychology, Faculty of Psychology and Educational Sciences, Ludwig-Maximilian University, Munich, Germany; and Orygen, the National Centre of Excellence for Youth Mental Health, Melbourne, Victoria, Australia; **Rachel Upthegrove**, Institute for Mental Health and Centre for Brain Health, University of Birmingham, Birmingham, UK; and Early Intervention Service, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK; **Raimo Salokangas**, Department of Psychiatry, University of Turku, Turku, Finland; **Stefan Borgwardt**, Department of Psychiatry (University Psychiatric Clinics, UPK), University of Basel, Basel, Switzerland; and Department of Psychiatry and Psychotherapy, University of Lübeck, Lübeck, Germany; **Paolo Brambilla**, Department of Neurosciences and Mental Health, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy; and Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy; **Eva Meisenzahl**, Department of Psychiatry and Psychotherapy, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany; **Stephan Ruhrmann**, Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital, University of Cologne, Cologne, Germany; **Frauke Schultze-Lutter**, Department of Psychiatry and Psychotherapy, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany; Department of Psychology, Faculty of Psychology, Airlangga University, Surabaya, Indonesia; and University Hospital of Child and Adolescent Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland; **Rebekka Lencer**, Department of Psychiatry and Psychotherapy, University of Lübeck, Lübeck, Germany; and Institute for

https://doi.org/10.1192/bjp.2023.141 Published online by Cambridge University Press

Translational Psychiatry, University of Münster, Münster, Germany; **Alessandro Bertolino**, Department of Basic Medical Science, Neuroscience and Sense Organs, University of Bari Aldo Moro, Bari, Italy; **Christos Pantelis**, Melbourne Neuropsychiatry Centre, University of Melbourne & Melbourne Health, Melbourne, Victoria, Australia; **Nikolaos Koutsouleris**, Department of Psychology, Faculty of Psychology and Educational Sciences, Ludwig-Maximilian University, Munich, Germany; Max-Planck Institute of Psychiatry, Munich, Germany; and Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK; **Joseph Kambeitz**, Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital, University of Cologne, Cologne, Germany; **for the PRONIA Study Group**

Correspondence: Derya Şahin. Email: deryasahin@protonmail.ch

First received 24 Mar 2023, final revision 2 Aug 2023, accepted 28 Sep 2023

Supplementary material

Supplementary material is available online at https://doi.org/10.1192/bjp.2023.141.

Data availability

Supplementary findings supporting this study are available on request from the corresponding author. The data are not publicly available owing to Institutional Review Board restrictions, since the participants did not consent to their data being publicly available. Code used for the analysis will be made available under https://github.com/deryasahinmd.

Author contributions

D.Ş.: conceptualisation, methodology, investigation, software, analysis, writing (original draft), visualisation, funding acquisition; L.K.-I., S.W., D.D., R.U., R.S., S.B., P.B., E.M., S.R., F.S.-L., R.L., A.B., C.P., N.K.: data acquisition, investigation, resources, writing (review and editing); J.K.: data acquisition, investigation, resources, conceptualization, methodology, software, analysis, supervision, writing (review and editing).

Funding

This work was funded by a Gerok grant awarded to D.Ş. by the Medical Faculty of University of Cologne under German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) support. PRONIA is a Collaborative Project funded by the European Union under the 7th Framework Programme under grant agreement no. 602152.

Declaration of interest

L.K.-I., F.S.-L. and R.U. are members of *BJPsych* editorial board and did not take part in the review or decision-making process of this paper.

References

1 Beauchamp TL, Childress JF. *Principles of Biomedical Ethics* (7th edn). Oxford University Press, 2013.

2 Verma S, Rubin J. Fairness definitions explained. *FairWare '18: Proceedings of the International Workshop on Software Fairness (Gothenburg, 29 May 2018)*. Association for Computing Machinery, 2018 (https://doi.org/10.1145/3194770.3194776).

3 Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med* 2021; **27**: 2176–82.

4 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366**: 447–53.

5 Schwartz RC, Blankenship DM. Racial disparities in psychotic disorder diagnosis: a review of empirical literature. *World J Psychiatry* 2014; **4**: 133–40.

6 Anglin DM, Malaspina D. Ethnicity effects on clinical diagnoses compared to best-estimate research diagnoses in patients with psychosis: a retrospective medical chart review. *J Clin Psychiatry* 2008; **69**: 941–5.

7 Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 2020; **396**: 1204–22.

8 Catalan A, Richter A, Salazar de Pablo G, Vaquerizo-Serrano J, Mancebo G, Pedruzo B, et al. Proportion and predictors of remission and recovery in first-episode psychosis: systematic review and meta-analysis. *Eur Psychiatry* 2021; **64**: e69.

9 Oliver D, Reilly TJ, Baccaredda Boy O, Petros N, Davies C, Borgwardt S, et al. What causes the onset of psychosis in individuals at clinical high risk? A meta-analysis of risk and protective factors. *Schizophr Bull* 2020; **46**: 110–20.

10 Salazar de Pablo G, Radua J, Pereira J, Bonoldi I, Arienti V, Besana F, et al. Probability of transition to psychosis in individuals at clinical high risk: an updated meta-analysis. *JAMA Psychiatry* 2021; **78**: 970–8.

11 Beck K, Andreou C, Studerus E, Heitz U, Ittig S, Leanza L, et al. Clinical and functional long-term outcome of patients at clinical high risk (CHR) for psychosis without transition to psychosis: a systematic review. *Schizophr Res* 2019; **210**: 39–47.

12 Catalan A, Salazar de Pablo G, Aymerich C, Damiani S, Sordi V, Radua J, et al. Neurocognitive functioning in individuals at clinical high risk for psychosis: a systematic review and meta-analysis. *JAMA Psychiatry* 2021; **78**: 859–67.

13 Addington J, Cornblatt BA, Cadenhead KS, Cannon TD, McGlashan TH, Perkins DO, et al. At clinical high risk for psychosis: outcome for nonconverters. *Am J Psychiatry* 2011; **168**: 800–5.

14 Salazar de Pablo G, Besana F, Arienti V, Catalan A, Vaquerizo-Serrano J, Cabras A, et al. Longitudinal outcome of attenuated positive symptoms, negative symptoms, functioning and remission in people at clinical high risk for psychosis: a meta-analysis. *EClinicalMedicine* 2021; **36**: 100909.

15 Sanfelici R, Dwyer DB, Antonucci LA, Koutsouleris N. Individualized diagnostic and prognostic models for patients with psychosis risk syndromes: a meta-analytic view on the state of the Art. *Biol Psychiatry* 2020; **88**: 349–60.

16 Kambeitz-Ilankovic L, Vinogradov S, Wenzel J, Fisher M, Haas SS, Betz L, et al. Multivariate pattern analysis of brain structure predicts functional outcome after auditory-based cognitive training interventions. *NPJ Schizophr* 2021; **7**: 40.

17 Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, Rosen M, Ruef A, Dwyer DB, et al. Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: a multimodal, multisite machine learning analysis. *JAMA Psychiatry* 2018; **75**: 1156–72.

18 Rosen M, Betz LT, Schultze-Lutter F, Chisholm K, Haidl TK, Kambeitz-Ilankovic L, et al. Towards clinical application of prediction models for transition to psychosis: a systematic review and external validation study in the PRONIA sample. *Neurosci Biobehav Rev* 2021; **125**: 478–92.

19 Koutsouleris N, Worthington M, Dwyer DB, Kambeitz-Ilankovic L, Sanfelici R, Fusar-Poli P, et al. Toward generalizable and transdiagnostic tools for psychosis prediction: an independent validation and improvement of the NAPLS-2 risk calculator in the multisite PRONIA cohort. *Biol Psychiatry* 2021; **90**: 632–42.

20 Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: theory of trust for AI in healthcare. *Intell Based Med* 2020; **1–2**: 100001.

21 Jacobson NC, Bentley KH, Walton A, Wang SB, Fortgang RG, Millner AJ, et al. Ethical dilemmas posed by mobile health and machine learning in psychiatry research. *Bull World Health Organ* 2020; **98**: 270–6.

22 Starke G, Schmidt B, De Clercq E, Elger BS. Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry. *AI Ethics* 2023; **3**: 303–14.

23 Xu J, Xiao Y, Wang WH, Ning Y, Shenkman EA, Bian J, et al. Algorithmic fairness in computational medicine. *EBioMedicine* 2022; **84**: 104250.

24 Koutsouleris N, Dwyer DB, Degenhardt F, Maj C, Urquijo-Castro MF, Sanfelici R, et al. Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA Psychiatry* 2021; **78**: 195–209.

25 Cannon TD, Yu C, Addington J, Bearden CE, Cadenhead KS, Cornblatt BA, et al. An individualized risk calculator for research in prodromal psychosis. *Am J Psychiatry* 2016; **173**: 980–8.

26 Cornblatt BA, Auther AM, Niendam T, Smith CW, Zinberg J, Bearden CE, et al. Preliminary findings for two new measures of social and role functioning in the prodromal phase of schizophrenia. *Schizophr Bull* 2007; **33**: 688–702.

27 Carrión RE, McLaughlin D, Goldberg TE, Auther AM, Olsen RH, Olvet DM, et al. Prediction of functional outcome in individuals at clinical high risk for psychosis. *JAMA Psychiatry* 2013; **70**: 1133–42.

28 Bobko P, Roth PL. The four-fifths rule for assessing adverse impact: an arithmetic, intuitive, and logical analysis of the rule and implications for future research and practice. In *Research in Personnel and Human Resources Management* (vol 23) (ed JJ Martocchio): 177–98. Emerald Group Publishing Limited, 2004.

29 Ojala M, Garriga GC. Permutation tests for studying classifier performance. *J Mach Learn Res* 2010; **11**: 1833–63.

30 Chapman EN, Kaatz A, Carnes M. Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. *J Gen Intern Med* 2013; **28**: 1504–10.

31 Adeponle AB, Groleau D, Kirmayer LJ. Clinician reasoning in the use of cultural formulation to resolve uncertainty in the diagnosis of psychosis. *Cult Med Psychiatry* 2015; **39**: 16–42.

32 Barr SM, Roberts D, Thakkar KN. Psychosis in transgender and gender non-conforming individuals: a review of the literature and a call for more research. *Psychiatry Res* 2021; **306**: 114272.

33 Dickson H, Hedges EP, Ma SY, Cullen AE, MacCabe JH, Kempton MJ, et al. Academic achievement and schizophrenia: a systematic meta-analysis. *Psychol Med* 2020; **50**: 1949–65.

34 Guo LN, Lee MS, Kassamali B, Mita C, Nambudiri VE. Bias in, bias out: under-reporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection – a scoping review. *J Am Acad Dermatol* 2022; **87**: 157–9.

35 Abbasi-Sureshjani S, Raumanns R, Michels BEJ, Schouten G, Cheplygina V. Risk of training diagnostic algorithms on data with demographic bias. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings* (eds J Cardoso, H Van Nguyen, N Heller, PH Abreu, I Isgum, W Silva, et al.): 183–92. Springer International Publishing, 2020.

36 Stormacq C, Van den Broucke S, Wosinski J. Does health literacy mediate the relationship between socioeconomic status and health disparities? Integrative review. *Health Promot Int* 2019; **34**: e1–17.

37 Kunst AE, Bos V, Lahelma E, Bartley M, Lissau I, Regidor E, et al. Trends in socioeconomic inequalities in self-assessed health in 10 European countries. *Int J Epidemiol* 2005; **34**: 295–305.

38 Erickson J, El-Gabalawy R, Palitsky D, Patten S, Mackenzie CS, Stein MB, et al. Educational attainment as a protective factor for psychiatric disorders: findings from a nationally representative longitudinal study. *Depress Anxiety* 2016; **33**: 1013–22.

EXTRA CONTENT ONLINE

https://doi.org/10.1192/bjp.2023.141 Published online by Cambridge University Press