

The development and validation of a prognostic model to predict relapse in adults with remitted depression in primary care: secondary analysis of pooled individual participant data from multiple studies

Andrew S Moriarty^{*1 2 3}, Lewis W Paton^{1 2}, Kym IE Snell^{4,5}, Lucinda Archer^{4,5}, Richard D Riley^{4,5}, Joshua EJ Buckman^{6 7}, Carolyn Chew-Graham³, Simon Gilbody^{1 2}, Shehzad Ali^{2 8}, Stephen Pilling⁶, Nick Meader⁹, Bob Phillips^{1 10}, Peter Coventry², Jaime Delgadillo¹¹, David A Richards^{12 13}, Chris Salisbury¹⁴, Dean McMillan^{1 2}

1. Hull York Medical School, University of York
2. Department of Health Sciences, University of York
3. School of Medicine, Keele University, Keele, UK
4. Institute of Applied Health Research, University of Birmingham, UK
5. National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, UK
6. CORE Data Lab, Research Department of Clinical, Educational & Health Psychology, University College London, London, UK.
7. iCope - Camden & Islington NHS Talking Therapies for anxiety and depression services - Camden & Islington NHS Foundation Trust, London, UK
8. Department of Epidemiology and Biostatistics, Schulich School of Medicine & Dentistry, Western University, London, ON, Canada
9. Newcastle University, UK
10. Centre for Reviews and Dissemination, University of York, UK
11. Department of Psychology, University of Sheffield, UK
12. Faculty of Health and Life Sciences, University of Exeter, UK
13. Department of Health and Caring Sciences, Western Norway University of Applied Sciences, Inndalsveien 28, 5063 Bergen, Norway
14. Centre for Academic Primary Care, University of Bristol, UK

*Corresponding author: andrew.moriarty@york.ac.uk

What is already known on this topic

Relapse contributes to the morbidity and burden associated with depression and, while there is robust research confirming predictors of relapse, individualised risk prediction is a challenge.

What this study adds

We found that it is not possible to accurately predict individualised risk of relapse using prognostic factors that are routinely collected and available in primary care. We found evidence to suggest that relationship status (not being in a relationship) is associated with increased risk of relapse and warrants confirmatory prognostic factor research.

How this study might affect research, practice or policy

Future prognosis research in this area should focus on exploring the feasibility of routinely measuring and documenting additional prognostic factors in primary care (for example, adverse childhood events, relationship status and social support) and including these in prognostic models. Until we can more accurately identify individuals at increased risk of relapse, commonly used acute-phase treatments could be optimised to better prepare for and mitigate the risk of relapse and there is a need for brief, scalable relapse prevention interventions that could be provided more widely.

Key words: Relapse; Recurrence; Depression; Prognosis

ABSTRACT

Background

Relapse of depression is common and contributes to the overall associated morbidity and burden. We lack evidence-based tools to estimate an individual's risk of relapse after treatment in primary care, which may help us more effectively target relapse prevention.

Objective

Develop and validate a prognostic model to predict risk of relapse of depression in primary care.

Methods

Multilevel logistic regression models were developed, using individual participant data from seven primary care-based studies (n=1244), to predict relapse of depression. The model was internally validated using bootstrapping and generalisability was explored using internal-external cross-validation.

Findings

Residual depressive symptoms [Odds ratio (OR): 1.13 (95% CI: 1.07-1.20), $p < 0.001$] and baseline depression severity [OR: 1.07 (1.04-1.11), $p < 0.001$] were associated with relapse. The validated model had low discrimination [C-statistic 0.60 (0.55-0.65)] and miscalibration concerns [calibration slope 0.81 (0.31-1.31)]. On secondary analysis, being in a relationship was associated with reduced risk of relapse [OR: 0.43 (0.28-0.67), $p < 0.001$]; this remained statistically significant after correction for multiple significance testing.

Conclusions

We could not predict risk of depression relapse with sufficient accuracy in primary care data, using routinely recorded measures. Relationship status warrants further research to explore its role as a prognostic factor for relapse.

Clinical implications

Until we can accurately stratify patients according to risk of relapse, a universal approach to relapse prevention may be most beneficial, either during acute phase treatment or post-remission. Where possible, this could be guided by the presence or absence of known prognostic factors (e.g. residual depressive symptoms) and targeted towards these.

1. BACKGROUND

Depression is the leading cause of disability worldwide[1]; the vast majority of adults seeking treatment for depression are managed in primary care[2]. Relapse is common, with around half of people experiencing a relapse within one year of reaching remission[3]. This high relapse rate contributes to the overall morbidity and burden associated with depression[4].

An ability to predict an individual patient's risk of relapse after an episode of depression might assist clinicians in targeting relapse prevention interventions toward those at greatest risk. Well-established prognostic factors associated with increased risk of relapse are: residual depressive symptoms; previous depressive episodes; childhood maltreatment; comorbid anxiety; neuroticism; younger age of first onset, and rumination[5]. While the presence or absence of these prognostic factors can help refine estimates of overall prognosis to particular sub-groups, they do not effectively aid risk-stratification at the individual level. Sub-grouping methods have been used to predict average risk of relapse for groups of people with different combinations (or profiles) of prognostic factors[6]. However, individualised outcome prediction is best shaped using multiple prognostic factors in combination, in the form of multivariable prognostic models[7].

Our systematic review of prognostic models identified twelve studies of relapse prediction models[8]. The majority were at high overall risk of bias (the most significant limitations being inadequate sample size, inappropriate handling of missing data, and calibration or discrimination not reported). The developed models either demonstrated insufficient predictive performance on reported validation by the study authors, or they could not be feasibly implemented in a primary care setting due to the large number and type of included predictors. We concluded that we currently lack evidence-based tools to assist clinicians with risk prediction of depressive relapse in any clinical setting, and that new models are required to give accurate risk predictions in primary care settings.

2. OBJECTIVE

To develop and validate a prognostic model, for use in clinical primary care settings, to predict risk of relapse in adults with remitted depression[9].

3. METHODS

The methods align with PROGnosis RESearch Strategy (PROGRESS) recommendations[7] and the study is reported according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD-Cluster) guidance[10]. A Patient Advisory Group (PAG) contributed to this study, including selecting predictors, definition of outcome, target patient population and clinical application. The study was registered prospectively (ClinicalTrials.gov: NCT04666662). Further methodological details are available in our protocol paper[9].

3.1. Source of data, participants and setting

We formed the “PREDICTR” dataset from combined individual participant data (IPD) from UK primary care-based studies[9], identified through a literature search and review of the NIHR trials registry. Authors were asked to share data if studies included adult patients (18 years and over) with depression, measured using the Patient Health Questionnaire (PHQ-9) at a minimum of three time-points (to identify depression, remission, relapse/no relapse). We excluded studies in patient groups with significant psychiatric comorbidity and feasibility studies. The PREDICTR dataset is derived from all arms (control and intervention) of six RCTs of primary care-based interventions for depression (CADET, CASPER Plus, COBRA, Healthlines Depression, REEACT, and REEACT-2) and one observational cohort study (WYLOW) (Supplemental Table 1.1, Figure 1.1).

3.2. Start-point (remission)

Participants were in remission at the point of prediction. Participants must have had case-level depression at baseline (PHQ-9 score of 10 or more[11]), and (at 4 months after trial baseline): i) a post-treatment PHQ-9 score below the established cut-off of 10 (consistent with clinical recovery[3,12]) and ii) an improvement of ≥ 5 points on the PHQ-9 since depression diagnosis (which aligns with an established reliable change index to identify those with “reliable improvement”[13]).

3.3. End-point/outcome (relapse)

We coded participants as relapsed if they fulfilled the following criteria within 6-8 months post-remission: i) PHQ-9 score above the diagnostic cut-off (10 or more) and ii) ≥ 5 points greater than their symptom score at the time of remission. This is consistent with an established criteria for reliable and clinically significant deterioration[13].

3.4. Predictors

We identified predictors *a priori*, following a literature review and consensus within the multidisciplinary research team and the PAG. We included predictors that would currently be routinely available in primary care settings at the intended moment of prediction.

3.4.1. Predictors in primary analysis

The following variables have robust evidence for their role as relapse predictors[5,14] and were included in the model:

- Residual depressive symptoms (PHQ-9 score at remission (0-9); continuous variable);
- Previous episodes of depression (dichotomous predictor (0=no previous episodes, 1=one or more previous episodes);
- Comorbid anxiety (measured using the GAD-7[15] in six of the seven studies and the Clinic Interview Schedule-Revised (CIS-R)[16] in REEACT (see Supplemental Table 1.2). These measures were combined to create a composite score (z-score), modelled as a continuous predictor;
- Baseline severity of depressive symptoms (continuous predictor; PHQ-9 score at baseline (pre-treatment));
- RCT Intervention: to control for the presence of interventions within the RCTs, we coded the presence or absence of an effective intervention (based on the results of the RCT) as a dichotomous variable. This predictor was intended to control for the intervention as part of the model building process only; when making predictions in real-world primary care this predictor would always be set to zero (i.e., no experimental intervention present).

3.4.2. Exploratory predictors

These less well-evidenced predictors of relapse were included as part of an exploratory secondary analysis: age; gender; ethnicity; relationship status; multimorbidity (two or more long-term physical or mental health conditions, excluding depression and comorbid anxiety); employment status (unemployment being those of working age who do not have a job and are actively seeking one); and current antidepressant use[5,14,17].

3.5. Sample size

We used the *pmsampsize* package [18] to calculate the required minimum sample size of 722, with 145 events (see protocol for details[9]); our actual sample size (n=1244; 261 events) exceeded this.

3.6. Statistical analysis methods

3.6.1. Data pre-processing

Anonymised data were transferred and stored securely and harmonised in line with the pre-specified harmonisation procedure (Supplemental Tables 1.3 and 1.4).

3.6.2. Data integrity checks (risk of bias)

Data were summarised and checked against publications for key features such as number of participants (total and in each study arm), demographics, primary outcomes of study, relapse rates and missing data. Validity of data values were checked on data inspection and irregularities clarified through communication with the original authors. Risk of bias assessment was undertaken using the participants, predictors and outcome domains of PROBAST[19].

3.6.3. Missing data

Missing data were handled using multiple imputation with chained equations (MICE), under a missing at random assumption[20]. Missing values were imputed based on the values of other predictors and the outcome, using linear models for continuous predictors (residual symptoms, severity, comorbid anxiety) and logistic

models for binary predictors [number of previous episodes, RCT Intervention, outcome (relapse/no relapse)]. Imputation was undertaken for each study separately, preserving the clustering of participants within studies and any between-study heterogeneity in predictor effects and outcome prevalence. Each imputed dataset was then analysed separately using the same statistical methods, and the estimates were combined using Rubin's rules, to produce an overall estimate and measure of uncertainty of each regression coefficient and model performance measures[10]. We used thirty imputations, based on the maximum percentage of participants with one or more missing values across all individual studies[20].

3.6.4. Model development (primary analysis)

Multilevel multivariable logistic regression models were built to model the relationship of the predictors with the binary outcome (relapse/no relapse), forcing in all predictors. Model parameters were estimated via unpenalized maximum likelihood estimation, and then penalised post-estimation using a uniform shrinkage factor. The modelling preserved the clustering of participants within studies, with a random effect on the intercept, a random intervention effect, and allowing for between-study correlation in these effects. We explored non-linear relationships in the continuous variables using multivariable fractional polynomials (MFPs)[7]. Predictive performance statistics (C-statistic for discrimination; calibration slope and calibration-in-the-large) were calculated for the final developed model, first within each cluster in turn and then pooled using random effects meta-analysis to summarise the model's performance across clusters with estimates of the pooled average and 95% confidence intervals. Prediction intervals were constructed to estimate the model's likely performance in new but similar settings[10]. Calibration was also assessed visually by producing calibration plots with smooth calibration curves.

3.6.5. Model validation

The optimism of the developed model was measured using non-parametric bootstrapping. One hundred bootstrap samples (each stratified by study) were produced from the original dataset. Within each bootstrap sample, the same modelling procedures were used as for model development. The model estimated

using each bootstrap sample was then applied in both the same bootstrap sample ('apparent performance') and in the original (imputed) dataset ('test performance'). Each time, average performance measures were calculated by pooling within-study statistics using meta-analysis, as above.

Optimism was calculated as the difference between apparent and test performance; this process was repeated one hundred times and the average difference between the bootstrap (apparent) and test performance for each performance statistic provided the estimate of overall optimism for that statistic. Optimism-adjusted performance statistics (C-statistic, calibration slope and calibration-in-the-large) were subsequently derived. The uniform shrinkage factor (in this study, the optimism-adjusted calibration slope) was applied to all of the original estimated beta coefficients (to shrink them toward zero to address overfitting) to produce a penalised logistic regression model. Finally, the intercept was re-estimated (whilst constraining the penalised predictor effects at their shrunken value) to ensure overall calibration was maintained. This formed the final model. Generalisability of the model and between-study heterogeneity in model performance was assessed using internal-external cross-validation (IECV)[21].

3.6.6. Sensitivity analysis

To understand the impact of including a composite measure of comorbid anxiety calculated from both GAD-7 and CIS-R, a sensitivity analysis was performed measuring predictive performance statistics when omitting REEACT and using only GAD-7 as the measure of comorbid anxiety, rather than a z-score.

3.6.7. Secondary (exploratory) analyses

Univariable analyses were performed to evaluate the unadjusted association between each predictor variable and the outcome variable. Where univariable analysis found statistically significant associations (after accounting for multiple significance testing), the model was refit using all of the original included predictors plus the additional exploratory predictor, to explore the impact on model predictive performance (using only studies in which the exploratory predictor was available).

3.7. Ethics approval

The University of York's Health Sciences Research Governance Committee confirmed that this study was exempt from full ethical approval as it entailed the secondary analysis of anonymised data from studies that had already received ethical approval.

4. RESULTS

4.1. Summary of data

Table 1 presents the descriptive statistics for the IPD. Supplemental Table 2.1 summarises the missing data. Risk of bias and concerns around applicability were low (Supplemental Table 2.2).

Table 1: Descriptive statistics for IPD

Variable	Study							
	CADET	CASPER Plus	COBRA	Healthlines Depression	REEACT	REEACT-2	WYLOW	Combined PREDICTR IPD dataset
Total number included in analysis (remission)	158	101	169	110	221	159	326	1244
Number relapsed at 6-8 months (%)	32 (20)	28 (27.7)	19 (11.2)	24 (21.8)	34 (15.4)	17 (10.7)	107 (32.8)	261 (21)
Number with one or more previous episodes of depression (%)	112 (70.8)	82 (81.9)	104 (61.5)	89 (80.9)	154 (69.7)	159 (100)	124 (38)	824 (66.2)
Mean PHQ-9 at baseline (SD)	17.35 (4.20)	15.33 (3.39)	17.27 (4.05)	16 (3.70)	15.96 (3.92)	15.77 (3.54)	15.67 (4.17)	16.17 (3.98)
Mean PHQ-9 at remission (SD)	4.70 (2.69)	5.19 (2.50)	4.08 (2.56)	5.97 (2.30)	4.50 (2.61)	4.53 (2.62)	3.52 (2.31)	4.40 (2.60)
Mean GAD-7 (SD)	12.52 (4.83)	8.25 (4.85)	12.41 (4.96)	11.80 (4.44)	GAD-7 not used in REEACT (CIS-R anxiety measure**)	12.97 (4.44)	13.62 (4.63)	12.43 (4.92)
Mean age (SD)	43.2 (12.9)	71.8 (5.16)	45.3 (13.9)	Not applicable (age is categorical*)	40.3 (13.1)	43.3 (14.7)	41.7 (13.8)	45.1 (15.63)

Gender: Number Female (%)	113 (71.5)	65 (64.3)	100 (59.2)	80 (72.7)	148 (67)	102 (64.2)	190 (58.3)	798 (64.1)
Ethnicity*: Number White (%)	137 (86.7)	100 (99)	165 (97.6)	107 (97.2)	216 (97.7)	154 (97.5)	292 (94.2)	1171 (95.4)
Employment*: Number Employed (%)	133 (82.8)	Not collected	132 (78.1)	105 (95.4)	213 (96.3)	149 (93.7)	214 (65.6)	946 (82.9)
Relationship status*: Number In a relationship (%)	69 (43.7)	Not collected	106 (62.7)	Not collected	140 (75.7)	93 (58.9)	Not collected	408 (60.9)
Multimorbidity*: Number with multimorbidity (%)	88 (55.7)	86 (85.1)	96 (56.8)	Not collected	Not collected	Not collected	87 (29.7)	357 (49.5)
Current antidepressant use: Number on antidepressants at remission (%)	113 (71.5)	34 (41)	128 (75.7)	98 (90.7)	Not collected	55 (41.4)	130 (40.2)	558 (57.3)

*See Supplemental Table 2.3 for further detail on categorical predictors

**See Supplemental Table 2.4 for summary statistics for CIS-R anxiety subscale in REEACT

4.2. Univariable analysis

Table 2 presents the results from univariable multilevel models. Residual symptoms [OR: 1.13 (1.07-1.20)] and severity [OR: 1.07 (1.04-1.11)] were statistically significantly associated with relapse; number of previous episodes and comorbid anxiety were not.

Table 2: Univariable and multivariable associations between outcome and predictors (primary and secondary analysis)

Predictor	Univariable analysis*			Primary analysis** (multivariable) Seven studies (n=1244)		Secondary analysis** (multivariable) Four studies (n=707)****	
	Unadjusted odds ratio (95% CI)	p- value	Number of participants (N)	Regression beta coefficient (95% CI)***	p-value	Regression beta coefficient (95% CI)	p-value
Predictors in primary analysis							
Number of previous episodes	1.19 (0.84 to 1.72)	0.319	1117	0.13 (-0.24 to 0.50)	0.500	-0.15 (-0.66 to 0.37)	0.582
Residual symptoms	1.13 (1.07 to 1.20)	<0.001	1244	-2.11 (-3.07 to -1.14)	<0.001	0.07 (-0.01 to 0.15)	0.081
Baseline severity of depressive symptoms	1.07 (1.04 to 1.11)	<0.001	1244	0.09 (0.04 to 0.13)	<0.001	0.07 (0.01 to 0.13)	0.020
Comorbid anxiety (z-score)	1.04 (0.90 to 1.20)	0.589	1239	-0.13 (-0.30 to 0.03)	0.936	-0.12 (-0.37 to 0.14)	0.363
RCT intervention	0.99 (0.60 to 1.66)	0.981	1244	0.03 (-0.59 to 0.65)	0.936	-0.40 (-0.84 to 0.04)	0.076
Predictors in secondary analysis							
Relationship status (in a relationship)	0.43 (0.28 to 0.67)	<0.001	670			-0.79 (-1.23 to -0.34)	0.001
Comorbid anxiety (GAD-7)	1.00 (0.97 to 1.03)	0.943	1019				
Gender (female)	0.87 (0.62 to 1.10)	0.196	1244				

Ethnicity (white)		1.59 (0.86 to 2.93)	0.138	1227		
Age (continuous)		1.01 (1.00 to 1.02)	0.198	1133		
Age – categorical (years old)	<40	Reference category		1243		
	40-49	1.16 (0.80 to 1.68)	0.433			
	50-59	1.31 (0.88 to 1.95)	0.180			
	60-69	0.85 (0.51 to 1.43)	0.543			
	70+	1.93 (1.04 to 3.59)	0.037			
Employment (employed)		0.76 (0.52 to 1.11)	0.161	1141		
Multimorbidity		1.31 (0.90 to 1.90)	0.158	721		
Current antidepressant medication		0.97 (0.70 to 1.35)	0.853	974		
<p>* Univariable associations (unadjusted) between predictors and relapse within 6-8 months, after accounting for clustering by study</p> <p>**Beta coefficients before shrinkage (predictors adjusted for other variables in column)</p> <p>*** Intercept (baseline risk): -1.55 (95% CI: -2.12 to -1.00); Standard deviation of random effect on intercept: 0.43 (95% CI: 0.14 to 1.38); Standard deviation of random effect on slope (RCT Intervention): 0.49 (95% CI: 0.11 to 2.16); Correlation between random effects: -0.23 (-0.93 to 0.84)</p>						

****Secondary analysis included only studies with relationship status available (n=707)

4.3. Model development and apparent predictive performance

Table 2 presents the results of multivariable, multilevel logistic regression analysis for the primary analysis. The developed model, prior to shrinkage, had a pooled apparent performance of: C-statistic 0.62 (95% CI: 0.57-0.67), calibration slope of 0.95 (95% CI: 0.54-1.36), and calibration-in-the-large of 0.03 (95% CI: -0.49-0.54). See Supplemental Materials 3 for within-study performance statistics.

4.4. Internal validation, shrinkage and final equation

Optimism-adjusted performance statistics, after bootstrapping, were: C-statistic 0.60, calibration slope 0.81, and calibration-in-the-large 0.03. The final model (Table 3) was produced by multiplying the original beta regression coefficients (from Table 2) by 0.81 (the optimism-adjusted calibration slope), and re-estimating the intercept to ensure calibration-in-the-large.

Table 3: Summary of model's predictive performance for primary, sensitivity and secondary analyses

Measure of predictive performance*	Primary analysis			Secondary analysis (exploring relationship status as a predictor)	
	Development (apparent performance)	Internal validation (optimism-adjusted performance)	Internal-external cross validation	Without relationship status	With relationship status
				Apparent performance	Apparent performance
Number of participants (number of clusters)	1244 (7)	1244 (7)	1244 (7)	707 (4)	707 (4)
C-statistic (95% CI*)	0.62 (0.57 to 0.67)	0.60	0.60 (0.55 to 0.65)	0.60 (0.54 to 0.66)	0.63 (0.57 to 0.70)
Calibration slope (95% CI*)	0.95 (0.53 to 1.36)	0.81	0.81 (0.31 to 1.31)	0.94 (0.37 to 1.51)	0.96 (0.56 to 1.36)
Calibration-in-the-large (95% CI*)	0.03 (-0.49 to 0.54)	0.03	0.00 (-0.61 to 0.61)	0.01 (-0.25 to 0.27)	0.01 (-0.20 to 0.23)

*Pooled predictive performance across studies (within-study performance statistics all available in Supplemental Materials 3-6); 95% CI presented, where applicable (not for optimism-adjusted statistics)

Final shrunken model: Intercept (baseline risk): -1.49; Shrunken beta coefficients: Number of previous episodes: 0.11, Residual symptoms: -1.71; Severity: 0.07; Comorbid anxiety: -0.11; RCT Intervention: 0.02

To calculate the risk of relapse within 6-8 months, using this model, one would use the following formula: $\exp(\text{person's risk score}) / [1 + \exp(\text{person's risk score})]$

Where person's risk score (linear predictor) = $-1.49 + 0.11 (\text{Number of previous episodes}^1) - 1.71 (\text{Residual symptoms}^2) + 0.07 (\text{Severity}^3) - 0.11 (\text{comorbid anxiety}^4) + 0.02 (\text{RCT Intervention}^5)$

1. No previous episodes = 0; One or more previous episodes = 1
2. $X^{-0.5} - 0.43$ (where $X = (\text{residual_symptoms} + 1)$) – this is the adjustment for non-linear transformation and mean-centring
3. Severity – 16.17
4. Comorbid_anx_zscore + 0.118
5. This would be zero when applied in clinical practice, outside the context of an RCT

4.5. Internal-external cross-validation

Generalisability of the model was assessed using IECV (Supplemental Materials 4)[21]. Calibration plots were compared for each validation in each of the different studies (Figure 1). These demonstrate inadequate calibration in most studies and significant heterogeneity in predictive performance across clusters. For example, WYLOW study shows severe miscalibration, with estimated risks generally too low, whereas in the COBRA study estimated risks are generally too high. In some studies, calibration was generally excellent (e.g., Healthlines Depression).

4.6. Sensitivity analysis

We removed REEACT and repeated the modelling process on the remaining six studies, to assess the impact of using z-scores to model comorbid anxiety. This did not change the study conclusions (see Supplemental Material 5 for analysis).

4.7. Secondary analysis

On univariable multilevel logistic analysis, relationship status was a highly statistically significant predictor (after adjusting the significance level to account for multiple significance testing using the Bonferroni correction). To further explore relationship status as a predictor of relapse, we repeated the model development procedures used in the primary analysis for the studies that included relationship status (CADET, COBRA, REEACT, and REEACT-2). We conducted these analyses both with and without the relationship status variable to provide a direct comparison (see Supplemental Material 6). Relationship status remained a statistically significant relapse predictor after adjusting for other prognostic factors (previous episodes, residual symptoms, severity, and comorbid anxiety).

5. DISCUSSION

We developed a model for predicting depression relapse in adults with remitted depression in primary care. Generally, the model had suboptimal predictive performance, with heterogeneous calibration across clusters on IECV and C-statistic below that required for acceptable discrimination. We would not recommend implementation in its current form, though calibration was promising in a subset of

studies. Secondary analysis found a statistically significant association between relationship status and relapse.

5.1. Findings in the context of the literature

The performance of this model was similar to performance measures for previous relapse prediction models[8]. Residual symptoms were associated with relapse, which is consistent with existing literature[5,14]. Residual symptoms are also associated with a more chronic depression course and poorer psychosocial functioning[22] and, as such, are an established treatment target in depression. The pre-existing evidence for severity as a prognostic factor for relapse is more equivocal than for residual symptoms[5]. Residual symptoms are more likely in people with more severe initial depressive illness[23] and so the presence of residual symptoms may be a mediator of the relationship between baseline severity and relapse.

The lack of association between previous episodes and relapse in our study is not consistent with the consensus view[5,14]. A possible explanation is that previous episodes is most strongly associated with recurrence (which occurs over a longer time period than relapse) and therefore our 6–8-month follow-up was not sufficient to detect this association[5]. Comorbid anxiety is a recognised predictor of relapse[5]; in particular, higher anxiety levels at baseline have been found to predict a shorter time to relapse after treatment[24]. In this study, we modelled anxiety at baseline (when depressed); it may be that an isolated measure of anxiety symptom severity at a single time-point is a crude measure and less important than knowing an individual's history of comorbid anxiety.

Marital status (being single) is a risk factor for developing depression[17]. A recent study also identified being single or no longer married as being associated with a worse prognosis (more depressive symptoms) at 3-4 months (but not beyond 3-4 months)[25]. While marital status is not an established predictor of relapse[5,17], our systematic review of prognostic models identified low-quality evidence of an association between relapse and marital status[8]. A potential mechanism by which being in a relationship may be protective against relapse is through providing increased social support, although this is likely to be mediated by other factors (e.g., relationship quality). The lack of association between relapse and other exploratory

predictors (age, gender, ethnicity, employment status and multimorbidity) is consistent with the literature[5,14].

The prognostic factor research cited here is based on longitudinal studies, which examine the predictive value of specific variables based on sample-level trends. As our systematic review[8] and the current study demonstrate, combining these well-evidenced relapse predictors to produce individualised predictions of relapse risk remains challenging and calibration remains suboptimal for the purpose of personalised decision-making[7].

5.2. Strengths and limitations

This study was conducted according to best practice recommendations for methodology and reporting, with a sufficient sample size to produce precise risk estimates[26]. We pre-selected predictors with a robust, pre-existing evidence base, to mitigate the risk of overfitting associated with data-driven approaches for predictor selection.

While cohort studies and RCTs are recommended sources of data for prognostic model development[7], participants may differ from the general population in important ways and results should be interpreted with this in mind. For example, the majority of participants in this study were white, limiting our ability to explore ethnicity as a predictor. Furthermore, our IPD were drawn from a sub-sample of participants in the identified studies and therefore a potentially limited representation of the wider patient population. Some potentially useful predictors were not included [e.g., neuroticism, childhood maltreatment and rumination[5]] as they were not coded for in our cohorts and are not routinely measured in general practice settings. There was a risk of selection bias given the way studies were selected for inclusion.

There was heterogeneity in: IPD study populations (e.g., CASPER Plus included older adults); relapse prevalence; antidepressant use at baseline; settings and treatment dose. WYLOW followed-up patients after low-intensity CBT whereas interventions in other studies were more intensive and delivered over a greater number of sessions. While interventions were controlled for in the analysis, this heterogeneity could explain some of the observed miscalibration. Finally, outcomes

(remission and relapse) were defined according to PHQ-9 (less optimal than diagnostic interview) and over a time-period (6-8 months) necessitated by the IPD data collection points (although one that is aligned with established definitions of relapse[5]).

5.3. Implications for future research

The strong statistically significant association between relationship status and relapse in our study warrants further confirmatory prognostic factor research going forwards. Further research is needed to better understand whether other relapse predictors can be captured and recorded in an acceptable and valid way by primary care health professionals. For example, we know that routinely asking people about childhood maltreatment is not harmful[27], therefore this could feasibly form part of routine relapse risk assessment in primary care. There have been some efforts to develop clinically useful and valid brief instruments to measure rumination[28], which could be explored in a primary care setting. If robust evidence supported the clinical utility of measuring and documenting additional relapse predictors, health professionals might then adopt this as routine practice.

The existing prognostic factor research in this area is not conclusive[5]. There is likely value in further exploratory prognostic factor research to examine the role of other variables (for example, those associated with depression onset and poor prognosis) as relapse predictors. Improved risk prediction may theoretically be possible by incorporating a wider range of predictors (e.g., biomarkers, genetics and brain imaging), although such data is unlikely to be routinely or widely available in primary care settings. Researchers are encouraged to develop and implement a set of core predictors based on a standard set of measurements to be integrated in future studies. Better data linkage and systems integration across health and wellbeing services may also be beneficial. A prospective, naturalistic cohort study would allow for the inclusion of a wider range of participants and measurement of pre-defined predictor and outcome information. This would potentially allow for more useful predictive models to be developed than secondary analysis of pre-existing data, although would be more costly and time-consuming.

In this study, we modelled the outcome of relapse as a binary outcome; it would be informative as part of future work to model the outcome on its continuous scale. We focused on outcome occurrence by 6-8 months, but other time-points may be of interest. As our IECV (which used the mean intercept and predictor-outcome associations to estimate performance) demonstrated, there was heterogeneity in the external performance and generalisability of the model was not guaranteed. An alternative approach to IECV that could be considered in future is local recalibration or intercept selection, where similarities in the outcome frequency or baseline characteristics (for example, mean age or proportion female) of a new population of interest is used to guide the intercept when applying the model in a different context[29]. If the predictive performance of relapse prediction models can be improved in the future through recalibration or updating, clinical usefulness (using net benefit analysis) must be considered prior to implementation.

6. CLINICAL IMPLICATIONS

Existing relapse risk prediction models are currently insufficiently accurate, and unlikely to be suitable to guide the provision of relapse prevention in primary care. There are different approaches to prevention: universal approaches, which target whole populations; selective approaches, which target higher-risk groups; and indicated approaches, directed at individuals. In the absence of sufficiently accurate relapse risk prediction tools, we argue that a universal approach to relapse prevention of depression in primary care is currently warranted. This is likely to require a systems approach to mitigating the risk and improving the management of relapse for all patients. This could mean targeting treatment at known prognostic factors (e.g., focussing on reducing residual symptoms) or providing interventions during the acute phase of depression treatment that target mechanisms of relapse. Clinicians should ensure they consider relapse risk, discuss this with patients and prioritise relapse prevention planning where appropriate[30]. Longer term, brief, inexpensive and scalable relapse prevention interventions are likely to be required for use in primary care.

FIGURE LEGENDS

Figure 1: Calibration plots for internal-external cross-validation within each study

ACKNOWLEDGEMENTS

Thank you to the patient advisory group without whose contributions this study would not have been possible: Greg Ball, Joanne Castleton, Penney Mayall, Gillian Payne, Sue Penn, and Emma Williams. Thank you to Trevor Sheldon, Paul Tiffin and Joanne Reeve for their role as thesis advisory panel members.

COMPETING INTERESTS

None.

FUNDING

This report is independent research supported by the National Institute for Health and Care Research (NIHR Doctoral Research Fellowship, Dr Andrew Moriarty, DRF-2018-11-ST2-044). KIES, RDR and LA are supported by funding from the NIHR Birmingham Biomedical Research Centre (BRC, IS-BRC-1215-20009). CC-G is part funded by NIHR Applied Research Collaboration (ARC) West Midlands (NIHR200165). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

AUTHOR CONTRIBUTIONS

ASM, CC-G, SG, SA and DM conceptualised the study and acquired funding. ASM led on data curation, study administration and oversight, and wrote the original draft. ASM, LWP, KIES, LA, RDR, JEJB and NM contributed to methodology and data analysis. SG, JD, DAR and CS were responsible for collecting the data and agreed to share these for the purpose of this study. SP, BP and PC provided input on content and methodology. All authors contributed to the pre-registered protocol and study plan and all authors contributed to and approved the final manuscript. ASM is the guarantor.

REFERENCES

- [1] World Health Organisation. Weekly Epidemiological Report. vol. 30. 2017. <https://doi.org/10.1371/jour>.
- [2] Ferencik EK, Ramanuj P, Pincus HA. Depression in primary care: part 1 - screening and diagnosis. BMJ 2019;365:l794. <https://doi.org/10.1136/bmj.l794>.

- [3] Ali S, Rhodes L, Moreea O, McMillan D, Gilbody S, Leach C, et al. How durable is the effect of low intensity CBT for depression and anxiety? Remission and relapse in a longitudinal cohort study. *Behaviour Research and Therapy* 2017;94:1–8. <https://doi.org/10.1016/j.brat.2017.04.006>.
- [4] Gauthier G, Mucha L, Shi S, Guerin A. Economic burden of relapse/recurrence in patients with major depressive disorder. *J Drug Assess* 2019;8:97–103. <https://doi.org/10.1080/21556660.2019.1612410>.
- [5] Buckman JEJ, Underwood A, Clarke K, Saunders R, Hollon SD, Fearon P, et al. Risk factors for relapse and recurrence of depression in adults and how they operate: A four-phase systematic review and meta-synthesis. *Clin Psychol Rev* 2018;64:13–38. <https://doi.org/10.1016/j.cpr.2018.07.005>.
- [6] Saunders R, Cohen ZD, Ambler G, Derubeis RJ, Wiles N, Kessler D, et al. A Patient Stratification Approach to Identifying the Likelihood of Continued Chronic Depression and Relapse Following Treatment for Depression. *J Pers Med* 2021;11. <https://doi.org/10.3390/jpm11121295>.
- [7] Riley RD, van der Windt D, P C, Moons K. *Prognosis Research in Healthcare: Concepts, Methods, and Impact*. First edit. Oxford University Press.; 2019.
- [8] Moriarty AS, Meader N, Snell KIE, Riley RD, Paton LW, Dawson S, et al. Predicting relapse or recurrence of depression: Systematic review of prognostic models. *British Journal of Psychiatry* 2022. <https://doi.org/10.1192/bjp.2021.218>.
- [9] Moriarty AS, Paton LW, Snell KIE, Riley RD, Buckman JEJ, Gilbody S, et al. The development and validation of a prognostic model to PREDICT Relapse of depression in adult patients in primary care: protocol for the PREDICTR study. *Diagn Progn Res* 2021;5. <https://doi.org/10.1186/s41512-021-00101-x>.
- [10] Debray TPA, Collins GS, Riley RD, Snell KIE, Van Calster B, Reitsma JB, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-Cluster): explanation and elaboration. *BMJ* 2023:e071058. <https://doi.org/10.1136/bmj-2022-071058>.
- [11] Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: Validity of a brief depression severity measure. *J Gen Intern Med* 2001;16:606–13. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>.
- [12] Clark DM, Layard R, Smithies R, Richards DA, Suckling R, Wright B. Improving access to psychological therapy: Initial evaluation of two UK demonstration sites. *Behaviour Research and Therapy* 2009;47:910–20. <https://doi.org/10.1016/j.brat.2009.07.010>.
- [13] McMillan D, Gilbody S, Richards D. Defining successful treatment outcome in depression using the PHQ-9: A comparison of methods. *J Affect Disord* 2010;127:122–9. <https://doi.org/10.1016/j.jad.2010.04.030>.
- [14] Wojnarowski C, Firth N, Finegan M, Delgadillo J. Predictors of depression relapse and recurrence after cognitive behavioural therapy: a systematic review and meta-analysis. *Behavioural and Cognitive Psychotherapy* 2019;47:514–29. <https://doi.org/10.1017/S1352465819000080>.
- [15] Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Arch Intern Med* 2006;166:1092–7. <https://doi.org/10.1001/archinte.166.10.1092>.
- [16] Lewis G, Pelosi AJ, Araya R, Dunn G. Measuring psychiatric disorder in the community: A standardized assessment for use by lay interviewers. *Psychol Med* 1992;22:465–86. <https://doi.org/10.1017/S0033291700030415>.

- [17] Burcusa SL, Iacono WG. Risk for recurrence in depression. *Clin Psychol Rev* 2007;27:959–85. <https://doi.org/10.1016/j.cpr.2007.02.005>.
- [18] Ensor J. PMSAMPSIZE: Stata module to calculate the minimum sample size required for developing a multivariable prediction model 2018.
- [19] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med* 2019;170:51. <https://doi.org/10.7326/M18-1376>.
- [20] White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011;30:377–99. <https://doi.org/10.1002/sim.4067>.
- [21] Takada T, Nijman S, Denaxas S, Snell KIE, Uijl A, Nguyen TL, et al. Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets. *J Clin Epidemiol* 2021;137:83–91. <https://doi.org/10.1016/j.jclinepi.2021.03.025>.
- [22] Kennedy N, Foy K. The Impact of Residual Symptoms on Outcome of Major Depression. *Curr Psychiatry Rep* 2005;7:441–6.
- [23] Paykel E, Ramana R, Cooper Z, Hayhurst H, Kerr J, Barocka A. Residual symptoms after partial remission: An important outcome in depression. *Psychol Med* 1995;25:1171–80.
- [24] Forand NR, Derubeis RJ. Pretreatment anxiety predicts patterns of change in cognitive behavioral therapy and medications for depression. *J Consult Clin Psychol* 2013;81:774–82. <https://doi.org/10.1037/a0032985>.
- [25] Buckman JEJ, Saunders R, Stott J, Arundell LL, O'Driscoll C, Davies MR, et al. Role of age, gender and marital status in prognosis for adults with depression: An individual patient data meta-analysis. *Epidemiol Psychiatr Sci* 2021;30:e42. <https://doi.org/10.1017/S2045796021000342>.
- [26] Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;m441. <https://doi.org/10.1136/bmj.m441>.
- [27] Becker-Blease KA, Freyd JJ. Research participants telling the truth about their lives: The ethics of asking and not asking about abuse. *American Psychologist* 2006;61:218–26. <https://doi.org/10.1037/0003-066X.61.3.218>.
- [28] Marchetti I, Mor N, Chiorri C, Koster EHW. The brief state rumination inventory (BSRI): Validation and psychometric evaluation. *Cognit Ther Res* 2018;42:447–60. <https://doi.org/10.1007/s10608-018-9901-1>.
- [29] Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013;32:3158–80. <https://doi.org/10.1002/sim.5732>.
- [30] Bockting CL, Hollon SD, Jarrett RB, Kuyken W, Dobson K. A lifetime approach to major depressive disorder: The contributions of psychological interventions in preventing relapse and recurrence. *Clin Psychol Rev* 2015;41:16–26. <https://doi.org/10.1016/j.cpr.2015.02.003>.