Supplementary Material

Table S1: Descriptive analyses of demographic and clinical variables

	All sample	**Non depressed**	**Incident**	**Remitted**	**Persistent**
	(*n=13922*)	(*n=12837*)	(*n=499*)	(*n=426*)	(*n=160*)
Age (mean±SD)	51.83 (±8.98)	51.94 (±9.02)	50.36 (±8.28)	50.61 (± 8.30)	50.86 (± 8.41)
Sex (female)	7597 (54.6)	6805 (53.0)	352 (70.5)	308 (72.3)	132 (82.5)
Ethnicity (Caucasian)	7220 (52.4)	6729 (53.0)	226 (46.2)	201 (47.4)	64 (40.0)
University degree (yes)	7443 (53.5)	6992 (44.3)	221 (44.3)	180 (42.3)	50 (31.2)
Familial Monthly income (mean±SD)	1748.00 (1437.9)	1782.21 (1456.22)	1443.60 (1143.58)	1297.85 (1110.44)	1154.24 (1074.91)
Marital status (Living with partner vs. Other)	9239 (66.4)	8617 (57.1)	303 (60.7)	237 (55.6)	82 (51.2)

Table S2: Absolute and relative number of missing instances

Variable	Absolute number of missing instances	Relative number of missing instances (%)
Sex	0	0
Age	0	0
University degree	0	0
Familial income	0	0
Benzodiazepines	0	0
Antidepressants	0	0
Marital status	1	0.01
Non-smoker	1	0.01
Self-report health	5	0.03
Obesity	6	0.05
Negative life events	8	0.05
SAD	17	0.11
Panic disorder	17	0.11
OCD	19	0.13
Heavy drinking	27	0.18
GAD	157	1.04
Ethnicity	184	1.22

Table S3: Performance measures for different cut-offs of class boundaries.

Cut-offs	**PPV**	**NPV**	**Sensitivity**	**Specificity**	**Balanced Accuracy**
Depression versus non depression					
0.25	0.09	0.98	0.94	0.27	0.61
0.50	0.19	0.97	0.67	0.78	0.73
0.75	0.36	0.95	0.38	0.95	0.66
Incident depression versus non depression					
0.25	0.03	0.98	0.99	0.14	0.50
0.50	0.07	0.98	0.61	0.75	0.68
0.75	0.23	0.97	0.12	0.99	0.55
Chronic depression versus non depression					
0.25	0.03	1.00	0.98	0.48	0.73
0.50	0.07	1.00	0.81	0.84	0.82
0.75	0.23	0.99	0.53	0.97	0.75

Table S4: Supplementary analyses for the three predictive models

Analysis	Method
M1 - Model described in the main manuscript	Elastic net, down-sampling, 10-fold CV, 1000 repetitions
M2 - Model excluding subjects with missing data in the generalized anxiety disorder variable	Elastic-net, down-sampling, 10-fold CV, 1000 repetitions
M3 - Model with no class imbalance correction	Elastic-net, no sampling technique, 10-fold CV, 1000 repetitions
M4 - Random Forest model	Random forest, down-sampling, 10-fold CV, 250 repetitions
M5 - Models with random splits train/test	Elastic-net, 100 random splits 75:25 (train:test), 10-fold CV, 250 repetitions
M6 - Sensitivity analysis model excluding the social anxiety disorder variable	Elastic-net, down-sampling, 10-fold CV, 250 repetitions
M7 - Sensitivity analysis model excluding the generalized anxiety disorder variable	Elastic-net, down-sampling, 10-fold CV, 250 repetitions
M8 - Sensitivity analysis model excluding the obsessive-compulsive disorder variable	Elastic-net, down-sampling, 10-fold CV-250 repetitions
M9 - Least absolute shrinkage and selection operator model	LASSO, down-sampling, 10-fold CV, 250 repetitions

Table S5: Performance measure for alternative models differentiating depressed versus non-depressed participants

Model	Sens	Spec	BA	PPV	NPV	AUC (CI)
M1	0.67	0.78	0.73	0.19	0.97	0.79 (0.76 – 0.82)
M2	0.67	0.80	0.73	0.24	0.96	0.80 (0.77 - 0.83)
M3	0.12	0.99	0.55	0.64	0.92	0.81 (0.79 - 0.84)
M4	0.68	0.83	0.75	0.24	0.97	0.84 (0.82 - 0.85)
M5	0.66 (0.65 - 0.66)*	0.79 (0.79 - 0.80)*	0.73 (0.72 - 0.73)*	0.21 (0.21 - 0.22)*	0.96 (0.96 - 0.97)*	0.79 (0.79 - 0.80)*
M6	0.65	0.79	0.72	0.21	0.96	0.79 (0.78 - 0.81)
M7	0.66	0.72	0.69	0.17	0.96	0.77 (0.75 - 0.78)
M8	0.67	0.76	0.72	0.19	0.96	0.78 (0.77 - 0.80)
M9	0.66	0.79	0.73	0.20	0.97	0.78 (0.77 - 0.81)

*The variable heavy drinker was discarded in this model

Table S6: Performance measure for alternative models differentiating participants with incident depression from participants who did not develop depression

Model	Sens	Spec	BA	PPV	NPV	AUC (CI)
M1	0.61	0.75	0.68	0.07	0.98	0.71 (0.66 – 0.77)
M2	0.66	0.68	0.67	0.07	0.98	0.75 (0.70 - 0.80)
M3	0.00	1.00	0.50	NaN	0.96	0.72 (0.68 - 0.77)
M4	0.72	0.74	0.73	0.10	0.98	0.81 (0.78 - 0.83)
M5	0.58 (0.57 - 0.59)	0.72 (0.72 - 0.72)	0.65 (0.65 - 0.66)	0.07 (0.07 - 0.08)	0.98 (0.98 - 0.98)	0.71 (0.71 - 0.72)
M6	0.56	0.72	0.64	0.07	0.98	0.72 (0.67 - 0.76)
M7	0.58	0.67	0.63	0.06	0.98	0.69 (0.65 - 0.74)
M8	0.60	0.69	0.64	0.07	0.98	0.70 (0.65 - 0.74)
M9*	0.62	0.71	0.66	0.08	0.98	0.73 (0.70 - 0.75)

*The variable ethnicity was discarded in this model

Table S7: Performance measure for alternative models differentiating participants without depression from those with chronic depression

Model	Sens	Spec	BA	PPV	NPV	AUC (CI)
M1	0.81	0.84	0.82	0.07	1.00	0.90 (0.86 - 0.95)
M2	0.77	0.84	0.80	0.06	1.00	0.91 (0.87 - 0.95)
M3	0.04	1.00	0.52	0.66	0.99	0.94 (0.90 - 0.97)
M4	0.95	0.81	0.88	0.05	1.00	0.94 (0.92 - 0.96)
M5	0.78 (0.77 - 0.79)	0.85 (0.85 - 0.86)	0.82 (0.81 - 0.82)	0.06 (0.06 - 0.07)	1.00 (1.00 - 1.00)	0.90 (0.90 - 0.91)
M6	0.84	0.84	0.84	0.06	1.00	0.92 (0.89 - 0.94)
M7	0.75	0.84	0.79	0.05	1.00	0.90 (0.88 - 0.92)
M8	0.84	0.81	0.82	0.04	1.00	0.90 (0.87 - 0.92)
M9*	0.80	0.84	0.82	0.06	1.00	0.90 (0.87 - 0.93)

*The variable panic disorder was discarded in this model

Table S8: Final values for α and λ used in the main models

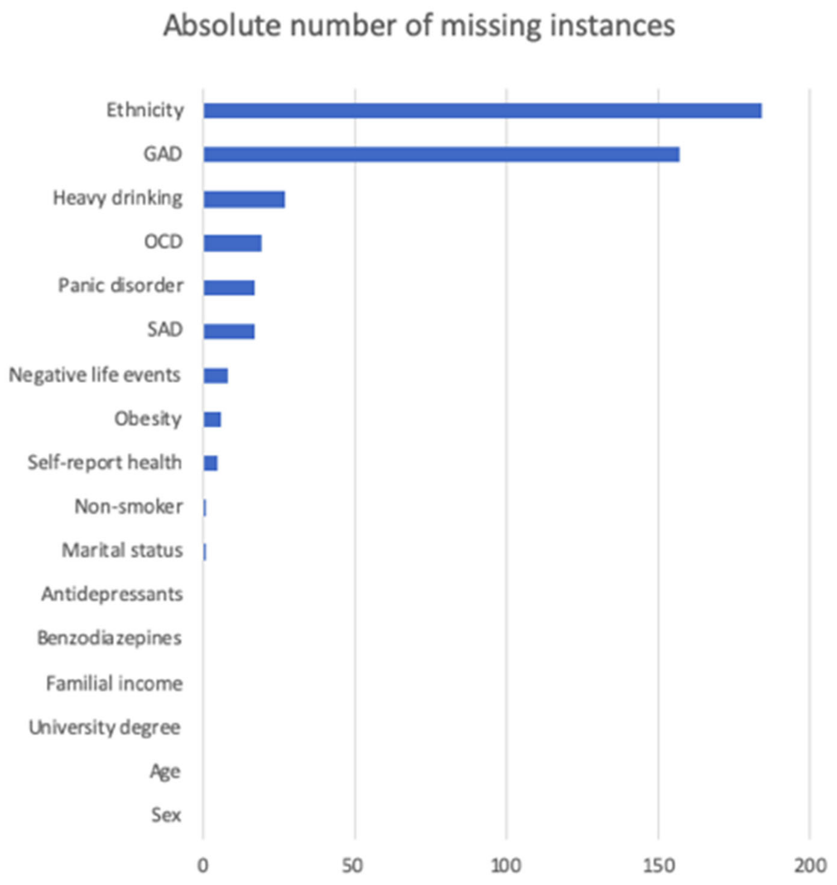
Model	α	λ
Depression vs. non-depression	0.1	0.02
Incident depression vs. non-depression	0.1	0.091
Chronic depression versus non-depression	0.1	0.099

Table S9: Elastic net regression penalized beta coefficients for the main models

Features	Model A	Model B	Model C
Intercept	-8.36	-3.99	-3.74
Age	-0.19	-0.13	-0.09
Familial monthly income	-0.20	-0.06	-0.17
Sex	0.63	0.49	0.83
Ethnicity	0.17	-0.11	-0.01
University degree	-0.13	0.00	-0.66
Married	-0.07	-0.06	-0.29
General health	-0.87	-0.35	-0.73
Obesity	0.34	0.02	0.08
Non-smoker	0.00	0.00	-0.33
Social phobia	2.27	0.25	0.51
Panic disorder	0.86	0.33	0.15
GAD	1.17	0.70	1.01
OCD	1.63	0.93	1.27
Heavy drinking	0.13	0.39	-0.33
Benzodiazepine	0.66	0.39	0.65
Use of antidepressants	0.63	0.48	0.43
Negative life events	0.36	0.21	0.67

Models differentiating (A) participants with depression from non-depressed participants; (B) participants with incident depression from participants who did not develop depression; (C) participants without depression from those with chronic depression.

Figure S1: Missing data distribution in absolute instances missed and percentage of missing data per variable.



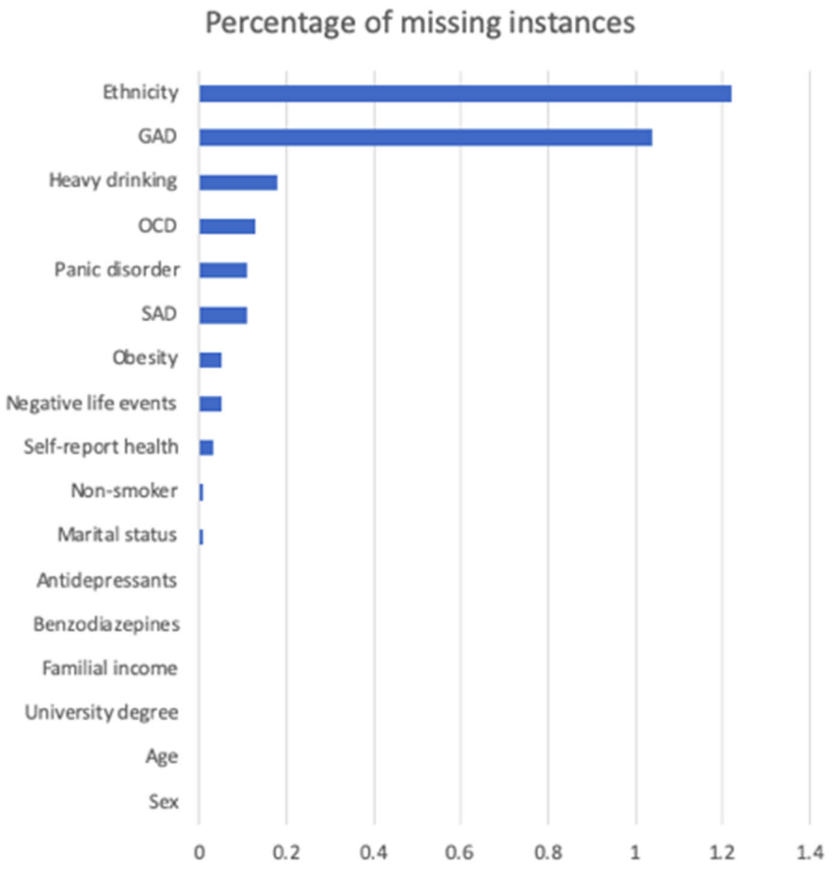

Percentage of missing instances
Ethnicity
GAD
Heavy drinking
OCD
Panic disorder
SAD
Obesity
Negative life events
Self-report health
Non-smoker
Marital status
Antidepressants
Benzodiazepines
Familial income
University degree
Age
Sex
0
0.2
0.4
0.6
0.8
1
1.2
1.4

Figure S2: PPVs and NPVs values for different cut-offs of class boundaries.

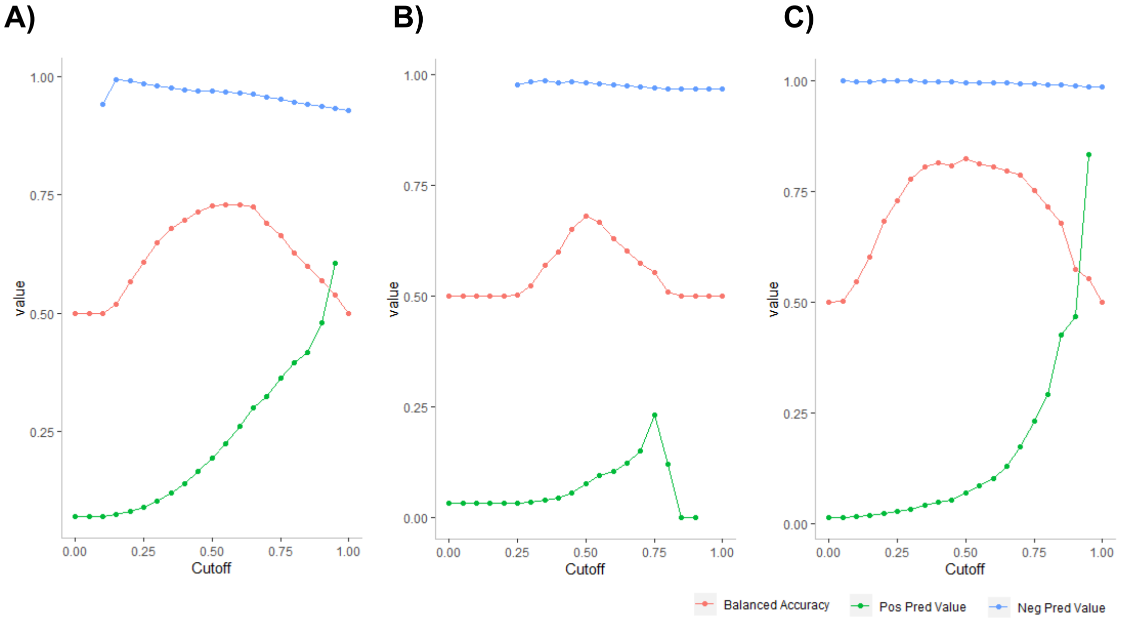


Figure S3: features selected in the model excluding missing data instances in the GAD variable with relative relevance weights

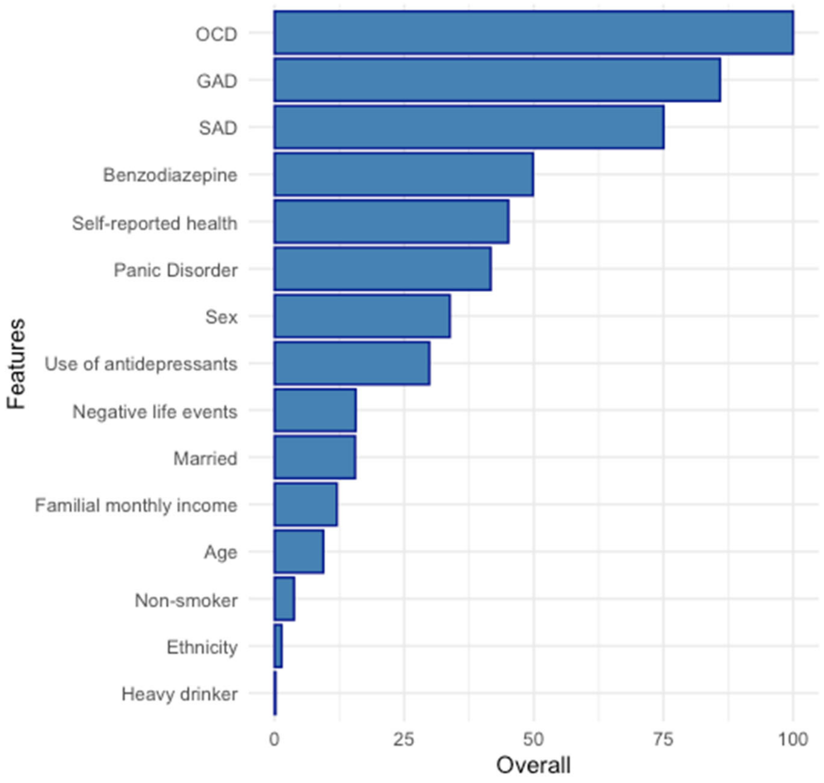
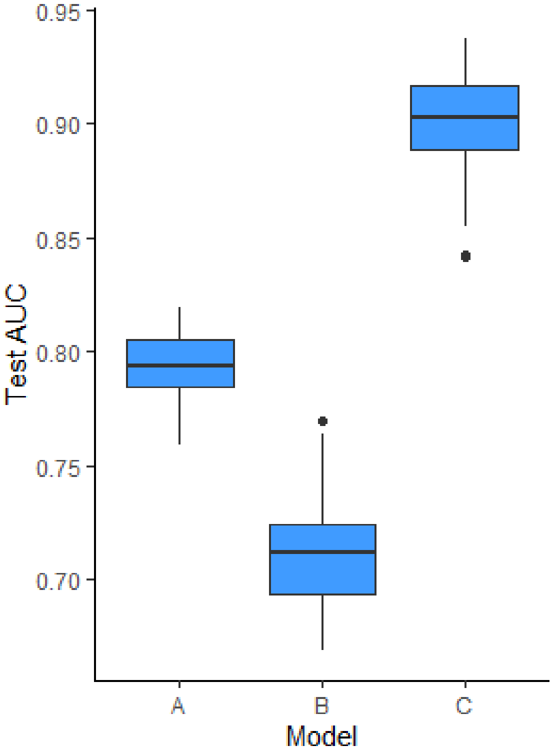


Figure S4: Boxplot for AUC test values for Model A (depressed versus non-depressed participants, table S5, model M5), Model B (participants with incident depression from participants who did not develop depression, table S6, model M5), and Model C (participants without depression from those with chronic depression, table S7, model M5).


Test AUC
A
B
C
Model