Journal of Anxiety Disorders 64 (2019) 40–44

Contents lists available at ScienceDirect

Journal of Anxiety Disorders

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

External validation of the International Risk Prediction Algorithm for the onset of generalized anxiety and/or panic syndromes (The Predict A) in the US general population

Yeshambel T. Nigatu[a,b], JianLi Wang[a,c,*]

[a] Institute of Mental Health Research, University of Ottawa, Ottawa, Canada
[b] The Ottawa Hospital Research Institute, Ottawa, Canada
[c] School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Canada

ARTICLE INFO

Keywords:
Generalized anxiety disorders
Panic disorder
Risk prediction
NESARC
PredictA

ABSTRACT

Introduction: Multivariable risk prediction algorithms are useful for making clinical decisions and health planning. While prediction algorithms for new onset of anxiety disorders in Europe and elsewhere have been developed, the performance of these algorithms in the Americas is not known. The objective of this study was to validate the PredictA algorithm for new onset of anxiety and/or panic disorders in the US general population.
Methods: Longitudinal study design was conducted with approximate 2-year follow-up data from a total of 24 626 individuals who participated in Wave 1 and 2 of the US National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) and who did not have generalized anxiety disorder (GAD) and panic disorder in the past year at Wave 1. The PredictA algorithm was directly applied to the selected participants.
Results: Among the participants, 5.4% developed GAD and/or panic disorder over two years. The PredictA algorithm had a discriminative power (*C*-statistics = 0.62, 95%CI: 0.61; 0.64), but poor calibration ($p < 0.001$) with the NESARC data. The observed and the mean predicted risk of GAD and/or panic disorders in the NESARC were 5.3% and 3.6%, respectively. Particularly, the observed and predicted risks of GAD and/or panic disorders in the highest decile of risk score in the NESARC participants were 13.3% and 10.4%, respectively.
Conclusion: The PredictA algorithm has acceptable discrimination, but the calibration with the NESARC data was poor. The PredictA algorithm is likely to underestimate the risk of GAD/panic disorders in the US population. Therefore, the use of PredictA in the US general population for predicting individual risk of GAD and/or panic disorders is not encouraged.

1. Introduction

Anxiety disorders are one of the most prevalent mental disorders in the general population, and are associated with immense health care costs (Bandelow & Michaelis, 2015). Anxiety disorders are estimated to be the ninth leading causes of disability globally (GBD, 2016). According to large population-based surveys, up to 33.7% of the population is affected by an anxiety disorder during their lifetime (Kessler et al., 2003; Martin, 2003). Substantial under-recognition and undertreatment of these disorders have been demonstrated (Cornelius, van der Klink, Brouwer, & Groothoff, 2014). There is no evidence that the prevalence of anxiety disorders have changed in the past years (Bandelow & Michaelis, 2015).

Preventing new onset or incident cases of anxiety disorders can reduce the overall disease burden of anxiety disorders on society. One of the challenges in the prevention of anxiety disorders is its multifactorial etiology. Evidence has shown that multiple factors play a role in influencing the development of generalized anxiety disorder (GAD) and other anxiety related disorders, including age, sex, educational level, marital status, employment status, ethnicity, living alone or with others, physical illness, lifetime depression, stress, financial strain, self-rated physical and mental health, alcohol use, childhood adversity, major life events, poor social support and experiences of discrimination on grounds of sex, age, ethnicity, appearance, disability, or sexual orientation (Chan et al., 2012; Cornelius et al., 2014; de Wit et al., 2010; Duivis, Vogelzangs, Kupper, de Jonge, & Penninx, 2013; Karsten, Nolen, Penninx, & Hartman, 2011Martin, 2003; Moreno-Peral et al., 2014; Prins et al., 2011).

For the purpose of early identification and early intervention, the PredictA algorithm was developed in primary care attendees in four

[*] Corresponding author at: University of Ottawa Institute of Mental Health Research, 1145 Carling, Avenue, Ottawa ON K1Z 7K4, Canada.
E-mail address: Jianli.Wang@theroyal.ca (J. Wang).

https://doi.org/10.1016/j.janxdis.2019.03.004
Received 12 November 2018; Received in revised form 1 February 2019; Accepted 19 March 2019
Available online 03 April 2019

European countries, who were between the ages of 18 and 75 and who did not have GAD and/or panic disorders in the past six months (King et al., 2011). The algorithm was developed to predict individuals' risks of GAD and/or panic disorders in the next two years. Because the predictive performance of a model based on the development data is often optimistic, it is important that the developed model is validated in different populations, in different geographic regions or in different time periods (Hasin, Goodwin, Stinson, & Grant, 2005; Ruan et al., 2008). This addresses the accuracy of a model in individuals from a different but plausibly related population. However, most reports evaluating prediction models focus on the issue of internal validity, leaving the important issue of external validity behind. The PredictA international algorithm had good performance in the development data (King et al., 2011). It was also validated with Chilean and Dutch data as part of the PredictA study (King et al., 2011). To our knowledge, the algorithm has not been independently validated in the general population. In the present study, the objective was to validate the PredictA algorithm in the US general population.

2. Methods

2.1. Study design and population

We used the data from the longitudinal cohort of the US National Epidemiological Survey on Alcohol and Related Conditions (NESARC). The NESARC was a nationally representative survey of the US general population funded by the National Institute on Alcohol Abuse and Alcoholism. Wave 1 of the NESARC was conducted between 2001 and 2002 in 43 093 respondents aged 18 years and older. Wave 2 of the NESARC was conducted between 2004 and 2005, about 3 years after Wave 1. Among the Wave 1 participants, 34 653 completed interviews at Wave 2. For current analysis, we included 24 626 participants who were aged 18–75 years and who did have generalized anxiety disorder (GAD) and/or panic disorder in the past year at Wave 1. A detailed description of the design and field procedures of the NESARC has described elsewhere (Grant et al., 2009; Hasin et al., 2005). The NESARC data were collected using face-to-face computer-assisted interviews by trained lay interviewers. As current study was a secondary data analysis of public use data, ethics review was waived by the Conjoint Health Research Ethics Review Board of University of Calgary.

2.2. Measures

2.2.1. Predictors

There are 10 predictors in the PredictA algorithm. The NESARC contains the following predictors similar with those in the PredictA study, which were measured using the same instruments or similar questions:

1) Time (follow-up of 6 and 24 months)
2) Age (years)
3) Sex (male/female)
4) For the predictor "Difficulties in paid and unpaid work", the NESARC did not include questions about work stress as measured by the Job Content Questionnaire in the PredictA. We used the answers to the questions: experiencing difficulties with boss or co-workers, and being fired or laid off in the past 12 months, as a proxy predictor. It was dichotomized as having or not having difficulties for paid or unpaid work.
5) Physical component score (PCS) measures physical quality of life in the past month, which was assessed by the Medical Outcomes study—Short Form (SF-12, version 2) (Jenkinson et al., 1997) in both the NESARC and the PredictA study.
6) Mental component score (MCS) measures past month mental quality of life. It was assessed by Medical Outcomes study—Short Form (SF-12, version 2) (Jenkinson et al., 1997) in both the NESARC and the PredictA. The PCS and MCS scores were standardized, ranging from 0 to 100.
7) History of depression in first-degree relatives was assessed as part of the AUDADIS(15). Same as the PredictA study, the NESARC participants were asked about whether their biological parents and siblings ever had depression (yes/no).
8) Any lifetime depression prior to 12 months at Wave 1 (yes/no) was assessed using AUDADIS based on the DSM-IV criteria (Ruan et al., 2008).
9) Country: As we validated the PredictA model in the US population, in our validation, we entered "0" for the coefficient of "country", assuming that the NESARC participants were similar with the UK sample.

2.2.2. Generalized anxiety disorder and/or panic disorders

Our main outcome of interest was new onset of GAD and/or panic syndrome in the next two years since baseline, assessed using the Alcohol Use Disorder and Associated Disabilities Interview Schedule (AUDADIS), based on the DSM-IV criteria (Ruan et al., 2008), a fully structured diagnostic interview that can be used by trained lay interviewers. Lifetime and past-year diagnoses were assessed at Wave 1. At Wave 2, diagnoses since Wave 1 were assessed.

2.3. Statistical analysis

The outcome variable of the prediction algorithm was new onset of GAD and/or panic disorders in the next two year since baseline, ascertained with Wave 2 data. We applied the PredictA algorithm in the NESARC data, using the exact same coefficients of the nine predictors in the PredictA model: time, age, sex, education, experiencing difficulties at work and laid off, physical and mental health, family history, and lifetime depression, country (Table 1).

We applied the prediction model directly to the selected NESARC participants with and without re-calibration. Re-calibration is a method of adjusting an existing model to predict risk in a new setting. It involves estimating only two new parameters that are expected to produce reasonable predictions beyond the dataset used for recalibration. The logit risk score (Z) was recalibrated to predict onset of GAD and panic disorders by fitting a logistic model with Z as the predictor variable, i.e. the slope (a) and intercept (b) were estimated for the model logit = $a + bZ$ (Steyerberg & Harrell, 2016).

Table 1
Predictors in the PredictA algorithm and the regression coefficients after shrinkage.

Prognostic factor	Levels in factor	Coefficients[a]
Constant		−0.560
Time	Months	0.005
Age	Each year	0.001
Sex	Male	
	Female	0.163
Difficulties in paid and unpaid work	No difficulties or often supported	
	Difficulties without support	0.380
Physical health	Each point on SF-12 subscale score	−0.029
Mental health	Each point on SF-12 subscale score	−0.033
First-degree relative with emotional problem	No	
	Yes	0.304
Lifetime depression	No	
	Yes	0.382
Country	UK	
	Spain	0.100
	Slovenia	−0.285
	Portugal	−0.108

[a] Regression coefficients after shrinkage (King et al., 2011).

We assessed the model performance by discrimination and calibration. Discrimination is the ability of a prediction model to separate those who experienced the outcome events from those who did not. We quantified discrimination by calculating the C statistic, which is identical to the area under a receiver operating characteristic (ROC) curve when the outcome is binary, also known as AUC. Calibration measures how closely the predicted outcomes agree with actual outcomes (or accuracy). For this we used the Hosmer–Lemeshow (H–L) χ^2 statistics. A χ^2 statistic was calculated to compare the differences between the mean predicted and the observed risks; large *p*-value (i.e., greater than 0.05) indicates good calibration.

We also assessed the calibration by grouping individuals into deciles of risk and visually comparing the observed and the predicted risk, so that the overall calibration and the areas with over or under prediction could be identified. We re-calibrated the algorithm to improve the agreement between the predicted and observed risks. All analyses were performed using Stata release 13 (Stata Corp. LP, USA).

3. Results

3.1. Sample characteristics

The baseline characteristics of the participants in the PredictA study and the NESARC are presented in Table 2. The participants in NESARC and Europe4 resembled each other, but slightly differed in gender and education. At baseline, the proportion of female was higher in the PredictA study than in NESARC (66.1% versus 55.3%). In the NESARC sample, 12-month prevalence of GAD and panic disorders at Wave 1 were 7.4% while the 6-month prevalence of GAD and panic disorders in the PredictA sample was 5.2%.

3.2. External validation of the PredictA study in the US population

In NESARC participants who had complete data on all predictors (n = 24 626), 5.6% developed GAD and/or panic disorders over two years, which is almost equivalent with the incidence of anxiety or panic syndrome in PredictA sample (5.5%).

Table 2
Demographic characteristics of the US and European population.

Characteristics	US population *N* (%)	European population[a] *N* (%)
Age, years		
18–29	5723 (20.4)	762 (15.2)
30–39	6296 (22.4)	840 (16.7)
40–49	6080 (21.6)	895 (17.8)
50–59	4749 (16.9)	983 (19.6)
60–69	3453 (12.3)	1030 (20.5)
70–76	1815 (6.5)	507 (10.1)
Sex		
Female	18559 (55.3)	3336 (66.3)
Male	12557 (44.7)	1692 (33.7)
Married or living with partner		
No	12395 (44.1)	1455 (28.9)
Yes	15721 (55.9)	3567 (71.1)
Education		
Higher education	16910 (56.3)	1431 (28.6)
Secondary	8550 (28.5)	1493 (29.8)
Primary/no education	4557 (15.2)	1656 (33.1)
Trade/other	–	429 (0.9)
Employment		
Employed/fulltime student	21254 (77.1)	2651 (52.8)
Retired	4013 (14.6)	1278 (25.5)
Unemployed/others	1055 (3.8)	1090 (21.7)
Unable to work	1240 (4.5)	–
Born in country of residence		
Yes	25 053 (83.7)	4580 (92.2)
No	4881 (16.3)	385 (7.8)
Ethnicity		
White European	22752 (75.8)	4928 (98.7)
No European	7262 (24.2)	63 (1.3)

[a] Europe4 (UK, Spain, Slovenia and Portugal).

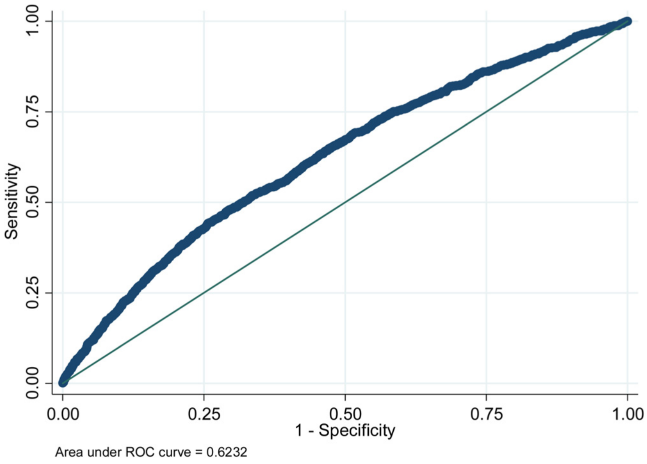

Fig. 1. Discrimination graph: receiver operating characteristics (ROC) graph.

Fig. 1 shows the ROC curve; the diagonal indicates no discrimination above chance. When we applied the PredictA algorithm in the NESARC data, using the exact same coefficients of the nine predictors in the original PredictA model, we found the *C*-statistic was 0.62 (95%CI: 0.61; 0.64), with poor calibration, as assessed by the Hosmer–Lemeshow (H-Lχ^2) (p < 0.001). While the *C*-statistics of PredictA sample (UK, Spain, Slovenia and Portugal) was 0.790 (95%CI: 0.767–0.813) with a perfect calibration. The observed and the mean predicted risk of GAD and/or panic disorders in the NESARC were 5.3% and 3.6%, respectively. Particularly, the observed and predicted risks of GAD and/or panic disorders in the highest (10th) decile of risk score in the NESARC participants were 13.3% and 10.4%, respectively. This suggests that the PredictA model tends to under estimate the risk of GAD and/or panic disorders in the US general population. Comparing the 10th (mean predicted risk = 10.4%) and the first (mean predicted risk = 1.1%) decile group, the PredictA model could identify over 10-fold of risk. Using the minimum risk of the 8th, 9th, and 10th decile group as cutoffs, the sensitivity (specificity) was 47.5% (70.9%), 35.3% (80.9%), and 19.4% (90.5%), respectively. Overall, the positive and negative predicted values were 50% and 94.7%, respectively.

With re-calibration, the *C*-index score (*C* = 0.62) remained the same. Although the agreement between the observed and predicted risks improved with re-calibration, the goodness of fit test remained significant (H-Lχ^2, p < 0.001) which indicates poor calibration. In Fig. 2A andB, we plotted the mean predicted probability vs the observed probability of GAD and panic disorders with and without re-calibration.

4. Discussion

We validated the PredictA algorithm for the new onset of GAD and panic disorders over two years in the U.S. general population. This multivariable algorithm had acceptable discriminative power (*C* = 0.62) but its calibration capacity with the NESARC data was poor. The PredictA was validated in Chile and a few European countries. To our knowledge, this is the first time that the PredictA algorithm was validated in the US general population. When the PredictA algorithm was applied in the NESARC, it under estimated the risk of GAD/panic disorder overall and in high risk groups. The absolute differences between the mean predicted and the observed risk of GAD and panic disorders were improved with re-calibration.

In prediction research, external validation is necessary since predictive models tend to perform better in the training or development

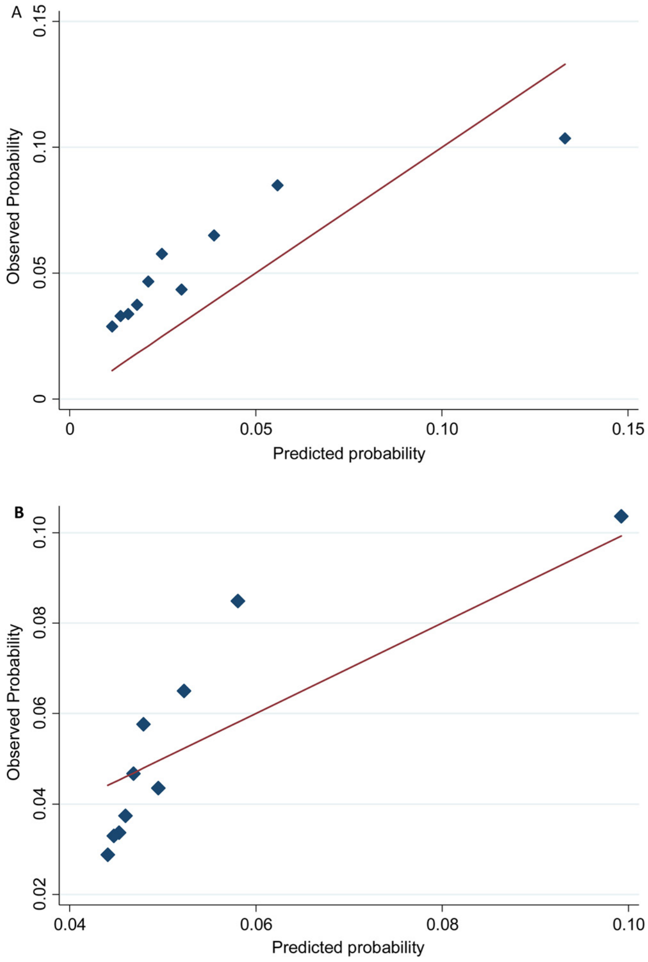

Fig. 2. Plots of mean predicted probability against observed probability of generalized anxiety disorder/Panic disorder within deciles of predicted risk without re-calibration. (A) Before calibration. (B) After recalibration.

sample than on new data. The concordance (*c*) statistic (equivalent to the area under the Receiver Operating Characteristic curve) provides a standard way of comparing the discrimination capacity of the tests that use different measurement units in different settings (Austin & Steyerberg, 2012; Steyerberg & Lingsma, 2008). It is known that greater discrimination is possible when the regression model contains independent explanatory variables that are strongly associated with the outcome (Austin & Steyerberg, 2012; Steyerberg & Harrell, 2016). However, when comparing the performance of the same regression model in different populations, a higher *c*-statistic is to be expected for the model fit in the population in which there is greater variation in the explanatory variable. In our study, we found that the PredictA multivariable algorithm seemed to perform well in the US general population ($C = 0.62$), which is lower than the *c*-statistic of the model in the UK sample (*C*-index = 0.69) (King et al., 2011).

The difference in the *C*-statistic between the PredictA and the current study may be due to many factors. First, in the present study, we validated the PredictA in the general population to examine the potential value of using this model in the US population as a whole while the PredictA model was developed in the primary care attendees, where the incidence of anxiety disorders might be high. We anticipate that this may not have a significant impact on the results because majority of the general population use primary care service in their lifetime. Second, the original PredictA has an outcome (PHQ-defined anxiety and/or panic syndrome) which is based on self-reported assessment, and might induce social desirability bias. In the NESARC, GAD/panic disorders were assessed using a fully structured diagnostic instrument based on the DSM-IV criteria. Third, the different performance of the PredictA algorithm in the NESARC may be due to timeframe. PredictA validation was carried out using 6-month outcome data from Estonia, The Netherlands and Chile whereas the present study was carried out using 24-month outcome data from the NESARC. However, the authors reported that time has no effect on the risk of GAD and/or panic disorders (King et al., 2011). In fact, the PredictA algorithm included a variable for time to predict the risk at intermediate times, but time had weak non-significant effect on the risk, thus, in practice, an individual's estimated risk is effectively the same at 6 and 24 months (King et al., 2011). Finally, we used 'experiencing difficulties with boss or coworker; and laid off' as a proxy of 'difficulties in paid and unpaid work' which might partly explain some of the discrepancies between the PredictA and NESARC studies. As difficulties with paid/unpaid work is associated with high total workload in terms of hours spend on paid/unpaid work, it might have a different impact on the risk of anxiety disorders. However, this could not be significant but more knowledge is needed on both changes in anxiety symptoms from workload in paid and unpaid work, and the interplay between these stressors over the life course. The differences in the model performance may also be due to different distributions of predictors (e.g. sex) in the European and American populations. For instance, women spend more time in unpaid work than men and the stress level could be double burden which could have some impacts on the risk of anxiety disorders.

The PredictA algorithm might perform well in the general population as much as in the primary care setting. But the calibration with the NESARC data was poor. In risk prediction research, calibration should receive more attention because it determines the model's potential clinical utility, in combination with the model's discriminative ability (Steyerberg & Harrell, 2016; Steyerberg, Van Calster, & Pencina, 2011; Steyerberg, van der Ploeg, & Van Calster, 2014; Van Calster, Steyerberg, & Harrell, 2015). The validation results showed that direct application of the PredictA algorithm would under estimate the risk of GAD and/or panic disorders in the NESARC participants, leading to more false negatives. With re-calibration, the performance of the PredictA algorithm improved but was still poor. This indicated that re-calibration and/or re-estimation might be needed to achieve optimal performance prior to applying a risk prediction algorithm in a new population. Furthermore, the development of sex-specific prediction algorithms for GAD and/or anxiety disorders might be important as the predictors for the risk of GAD and/or panic disorders and their predicted values may differ by sex.

The strength of this study is that the NESARC data were population-based and the sample size was large. To our knowledge, this is the first time that the PredictA algorithm was validated in a general population sample. This study also had limitations, including the fact that the NESARC relied on self-report. So reporting and recalling biases were possible. Such biases may also contribute to the inconsistencies in the predictive power of some factors in different populations. However, the instruments used in the NESARC have been validated and standardized as those in the PredictA study. Moreover, adding other risk factors when training PredictA model may refine risk assessment and improve the accuracy of the model in the general population.

5. Conclusions

The PredictA algorithm has acceptable discrimination, but the calibration capacity was poor in the US general population. Despite of re-calibration, the PredictA algorithm under estimated the risk of anxiety and panic disorders in the NESARC sample. Therefore, based on the results, at current stage, the use of PredictA in the US general population is not encouraged. In psychiatry, there have been many attempts in developing risk prediction algorithms. However, the developed tools need to be independently validated in different populations to ensure the generalizability of the models. More independent validation research is needed.

Conflict of interests

The authors have declared that they have no competing interests.

Acknowledgments

The National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) is funded by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) with supplemental support from the National Institute on Drug Abuse (NIDA).

References

Austin, P. C., & Steyerberg, E. W. (2012, January). Interpreting the concordance statistic of a logistic regression model: Relation to the variance and odds ratio of a continuous explanatory variable. *BMC Medical Research Methodology, 12*, 82. https://doi.org/10.1186/1471-2288-12-82.

Bandelow, B., & Michaelis, S. (2015). Epidemiology of anxiety disorders in the 21st century. *Dialogues in Clinical Neurosciences, 17*(3), 327–335.

Chan, H. N., Rush, A. J., Nierenberg, A. A., Trivedi, M., Wisniewski, S. R., Balasubramani, G. K., et al. (2012). Correlates and outcomes of depressed out-patients with greater and fewer anxious symptoms: A CO-MED report. *International Journal of Neuropsychopharmacology, 15*(10), 1387–1399.

Cornelius, B., van der Klink, J. J. L., Brouwer, S., & Groothoff, J. W. (2014). Under-recognition and under-treatment of DSM-IV classified mood and anxiety disorders among disability claimants. *Disability and Rehabilitation, 36*(14), 1161–1168.

de Wit, L. M., Fokkema, M., van Straten, A., Lamers, F., Cuijpers, P., & Penninx, B. W. (2010). Depressive and anxiety disorders and the association with obesity, physical, and social activities. *Depression and Anxiety, 27*(11), 1057–1065.

Duivis, H. E., Vogelzangs, N., Kupper, N., de Jonge, P., & Penninx, B. W. (2013). Differential association of somatic and cognitive symptoms of depression and anxiety with inflammation: Findings from the Netherlands Study of Depression and Anxiety (NESDA). *Psychoneuroendocrinology, 39*(9), 1573–1585.

Global Burden of Disease (GBD) 2015 Disease and Injury Incidence and Prevalence Collaborators (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015. *Lancet (London, England), 388*(10053), 1545–1602.

Grant, B. F., Goldstein, R. B., Chou, S. P., Huang, B., Stinson, F. S., Dawson, D. A., et al. (2009, November). Sociodemographic and psychopathologic predictors of first incidence of DSM-IV substance use, mood and anxiety disorders: Results from the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions. *Molecular Psychiatry, 14*(11), 1051–1066.

Hasin, D. S., Goodwin, R. D., Stinson, F. S., & Grant, B. F. (2005, October). Epidemiology of major depressive disorder: Results from the National Epidemiologic Survey on Alcoholism and Related Conditions. *Archives of General Psychiatry, 62*(10), 1097–1106.

Jenkinson, C., Layte, R., Jenkinson, D., Lawrence, K., Petersen, S., Paice, C., et al. (1997, June). A shorter form health survey: Can the SF-12 replicate results from the SF-36 in longitudinal studies? *Journal of Public Health Medicine, 19*(2), 179–186.

Karsten, J., Nolen, W. A., Penninx, B. W., & Hartman, C. A. (2011). Subthreshold anxiety better defined by symptom self-report than by diagnostic interview. *Journal of Affective Disorders, 129*(1–3), 236–243.

Kessler, R. C., Barber, C., Beck, A., Berglund, P., Cleary, P. D., McKenas, D., et al. (2003). The World Health Organization Health and Work Performance Questionnaire (HPQ). *Journal of Occupational and Environmental Medicine American College of Occupation Environmental Medicine, 45*(2), 156–174.

King, M., Bottomley, C., Bellon-Saameno, J. A., Torres-Gonzalez, F., Svab, I., Rifel, J., et al. (2011). An international risk prediction algorithm for the onset of generalized anxiety and panic syndromes in general practice attendees: PredictA. *Psychological Medicine, 41*(8), 1625–1639.

Martin, P. (2003). The epidemiology of anxiety disorders: A review. *Dialogues in Clinical Neurosciences, 5*(3), 281–298.

Moreno-Peral, P., Luna, J., de, D., Marston, L., King, M., Nazareth, I., et al. (2014). Predicting the onset of anxiety syndromes at 12 months in primary care attendees. The predictA-Spain study. *PLOS ONE, 9*(9), e106370.

Prins, M. A., Verhaak, P. F., Hilbink-Smolders, M., Spreeuwenberg, P., Laurant, M. G., van der Meer, K., et al. (2011). Outcomes for depression and anxiety in primary care and details of treatment: A naturalistic longitudinal study. *BMC Psychiatry, 11*. https://doi.org/10.1186/1471-244X-11-180.

Ruan, W. J., Goldstein, R. B., Chou, S. P., Smith, S. M., Saha, T. D., Pickering, R. P., et al. (2008, January). The alcohol use disorder and associated disabilities interview schedule-IV (AUDADIS-IV): Reliability of new psychiatric diagnostic modules and risk factors in a general population sample. *Drug and Alcohol Dependence, 92*(1–3), 27–36.

Steyerberg, E. W., & Harrell, F. E. (2016, January). Prediction models need appropriate internal, internal–external, and external validation. *Journal of Clinical Epidemiology, 69*, 245–247.

Steyerberg, E. W., & Lingsma, H. F. (2008, April). Prediction citations: Validation predictions models. *BMJ, 336*(7648), 789.

Steyerberg, E. W., Van Calster, B., & Pencina, M. J. (2011 Sep). Performance measures for prediction models and markers: Evaluation of predictions and classifications. *Revista Espanola de Cardiologia, 6*(9), 788–794.

Steyerberg, E. W., van der Ploeg, T., & Van Calster, B. (2014, July). Risk prediction with machine learning regression methods. *Biometrical Journal Biometrische, 5*(4), 601–606.

Van Calster, B., Steyerberg, E. W., & Harrell, F. H. (2015 November). Risk prediction for individuals. *JAMA, 314*(17), 1875.