International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. (2015)
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1478

Concordance between Composite International Diagnostic Interview and self-reports of depressive symptoms: a re-analysis

TOM ROSENSTRÖM,[1,2] MARKO ELOVAINIO,[1,2] MARKUS JOKELA,[1] SAMI PIRKOLA,[2,3] KOSKINEN SEPPO,[2] OLAVI LINDFORS[2] & LIISA KELTIKANGAS-JÄRVINEN[1]

1 Institute of Behavioural Sciences, University of Helsinki, Finland
2 National Institute for Health and Welfare, Helsinki, Finland
3 School of Health Sciences, University of Tampere, Tampere, Finland

Key words
depressive disorder, validity, reliability, Beck's Depression Inventory, General Health Questionnaire

Correspondence
Tom Rosenström,
Siltavuorenpenger 1A (PO Box 9), 00014 Helsinki, Finland.
Telephone (+35) 850-448-4146
Fax (+35) 891-912-9521
Email: tom.rosenstrom@helsinki.fi

Received 4 November 2014;
revised 27 January 2015;
accepted 18 March 2015

Abstract

Concordance between sum scores of self-reported depressive symptoms and structured interview diagnoses has been studied extensively, but are these the best attainable self-report-based predictions for interview diagnoses? We maximized the cross-validated concordance between World Health Organization's Composite International Diagnostic Interview (CIDI) diagnosis and Beck's Depression Inventory (BDI), and General Health Questionnaire (GHQ), from the viewpoint of exploratory statistics, re-analysing Health 2000 general-population sample of adults over 30 years in mainland Finland ($N = 5200–5435$). BDI sum-score prediction of CIDI diagnosis could be superseded by using (1) weighted sums of items, (2) classification trees constructed from items, or (3) a single item. Best solution (2) yielded cross-validated Youden's Index 0.757 [standard error (SE) = 0.001, sensitivity = 0.907, specificity = 0.851], improving the concordance to 1.07-fold (1.18-fold for 12-month diagnosis). A single-item solution was best for the GHQ. All positive predictive values remained low (0.09–0.31). Thus, CIDI-to-questionnaire concordance can be improved by using all information in the questionnaires instead of just sum scores, but latent-trait theory for questionnaires is incompatible with interview diagnoses (single item achieved better concordance than summing all). Self-reports have low predictive value for CIDI diagnoses in the general population, but better in settings with higher major depressive disorder (MDD) base rates. *Copyright © 2015 John Wiley & Sons, Ltd.*

Introduction

Approximately 4–6% of people fulfil the criteria for major depressive disorder (MDD) during a randomly selected 12-month period (Bromet *et al.*, 2011; Pirkola *et al.*, 2005; Vos *et al.*, 2013; Weissman *et al.*, 1996). These prevalence estimates are based on diagnostic interviews that are costly to conduct. "Concordance studies" investigate the feasibility of replacing a source of diagnostic information with another one that produces quantitatively equivalent information; focusing on accuracy of information retrieval rather than theoretic validity of the sources (Haro *et al.*, 2006). Several studies have derived good predictions of interview results from self-reports (Aalto *et al.*, 2012; Caraveo-Anduaga *et al.*, 1998; Goldberg *et al.*, 1997; Hewitt *et al.*, 2011; Nuevo *et al.*, 2009; Viinamäki *et al.*, 2004). Good concordance might justify the use of inexpensive self-reports for determining who has a depressive disorder, and provide other benefits too. For example, self-reports could be useful for diagnosing in specific populations, such as offenders that are difficult to interview (Hewitt *et al.* 2011). This study explores the concordance of self-report questionnaires of depression and World Health Organization's (WHO's) Composite International Diagnostic Interview (CIDI).

Rather than trying to maximize the concordance between self-report items and interview results, previous studies have mostly investigated sum scores of self-reported symptoms. Although they provide convenient summaries, the sum scores collapse all the information in available items to a single number, thus necessarily losing information. A relevant unanswered question then becomes: how good questionnaire-to-interview concordance could be achieved, were one to maximize it using all the information from the individual items? We answer to this question using computationally intensive classification techniques, because the "data mining" algorithms of explorative statistics (Hastie *et al.*, 2009; Kuhn and Johnson, 2013) can provide additional information that would not be readily available from classical confirmatory statistics and psychiatric intuition (Baca-García *et al.*, 2006; Wardenaar *et al.*, 2014). These results are compared with a previous classification study that used the same data and more classic methods (Aalto *et al.*, 2012).

Because most people do not have depression at a given time, low prediction-error rates can be achieved in general-population samples simply by predicting "no depression" for everyone. Therefore, this classification task is "imbalanced" with respect to the depression-diagnosis outcome (Chawla *et al.*, 2002; Kuhn and Johnson, 2013; Liu *et al.*, 2009), and simple classification-error rate is not a good performance measure. In contrast, Youden's Index (YI), or informativeness (sensitivity + specificity − 1), is a reasonable measure to maximize, treating both positive and negative miss-classifications as equally 'important' and being independent of prevalence (Youden, 1950). We avoid a common performance metric, area under the receiver operating characteristic ("ROC") curve, due to its known methodological problems, such as using irrelevant information contained in the "area under the curve" and ambiguities resulting from possible curve crossings among the studied methods (Kuhn and Johnson, 2013; Lobo *et al.*, 2008).

To gain a comprehensive picture of optimal prediction based on self-reported questionnaire items, we tested several complementary approaches and methods. These were compared using cross-validation methodology (Hastie *et al.*, 2009; Kuhn and Johnson, 2013), yielding an improved estimate for "true" concordance between standard self-report instruments and a CIDI diagnosis of depression.

Method

Participants

The participants were derived from the Health 2000 study of the general Finnish population (Aromaa and Koskinen, 2003; Pirkola *et al.*, 2005). A two phase stratified cluster sample of adults over 30 years and living in the mainland Finland was collected between August 2000 and March 2001. We use the sub-sample of 6005 participants who reliably underwent the structured mental health interview. In line with the previous concordance study (Aalto *et al.*, 2012), we excluded participants older than 80 years, the remaining 5675 participants forming our base sample; which was reduced to 5200 (Beck's Depression Inventory, BDI) or 5435 (General Health Questionnaire, GHQ), depending on the availability of independent measures. These participants underwent a four hour investigation, incorporating a 30-minute mental-health interview. The self-report inventories were either filled out immediately or returned at a later date, this generated unfortunate temporal incongruences in the data collection. Because the previous study is used as a reference point for further exploration, our protocol was designed to be as similar as possible regarding the sample details.

Measures

The Finnish version of the CIDI was used for determining depression diagnoses (Aalto *et al.*, 2012). This computerized version (Wittchenn and Pfister, 1997) allows estimation of DSM-IV diagnoses for major mental disorders.

The Finnish translation was crafted by psychiatric professionals, and the interviews performed by persons trained by a WHO-authorized psychiatrist or physician (Aalto *et al.*, 2012; Pirkola *et al.*, 2005). We studied two different diagnostic recency periods: those who fulfilled the DSM-IV diagnostic criteria within the past two weeks from the interview time (a two-week diagnosis) and those who fulfilled them within 12 months (a 12-month diagnosis). Twelve-month diagnoses are most commonly used for prevalence estimates (Bromet *et al.*, 2011; Pirkola *et al.*, 2005), but shorter-recency diagnoses are more strongly associated with self-reports of depression and functional disability (Aalto *et al.*, 2012; Bromet *et al.*, 2011). As such, both recencies are of interest for different purposes and both have been studied in the previous concordance work from the same sample (Aalto *et al.*, 2012).

In addition to recency, we studied two different depression outcomes. "Any depression diagnosis" refers to one or several of the following: single episode of MDD, mild, moderate, or severe with or without psychotic symptoms, similarly defined recurrent MDDs, and dysthymic disorder. These outcomes always excluded depressive episodes during bipolar disorder, however, due to a limited accuracy, some undetected bipolar cases are likely to remain in the data. "Pure MDD" refers to an episode of MDD, excluding those with comorbid anxiety disorder and/or substance use or dependence disorder (the exclusions were based on the 12-month recency also when studying two-week Pure MDD outcome). Unfortunately, we did not have data on all DSM-IV disorders, but chose nevertheless to make the same distinction between detected comorbidity and lack of evidence for comorbidity, as in the relevant study based on the same sample (Pirkola *et al.*, 2005).

The earlier CIDI classifications were predicted using self-reported BDI (Beck, 1967; Beck *et al.*, 1961) and a 12-item version of the GHQ (Aalto *et al.*, 2012; Goldberg *et al.*, 1997). BDI is based on the idea that depression is manifested in 21 symptoms that vary in degree of severity from not present (score zero) to severe (score three), whereas GHQ concentrates on recent changes in 12 more general mental health statements (also scored from zero to three). Questions that constitute the items of BDI refer to the current situation of the respondent, whereas the GHQ items refer to "recent" symptoms of the individual. Both the instruments are typically used by summing the item scores together, but here we also study predictive power of the items themselves. The GHQ is of interest due to its frequent use in health care settings. It was originally designed for screening any kind of psychopathology and mental distress rather than for differentiating between them, but may also serve to provide differential information.

Model comparison

Ten-fold cross-validation is a common and frequently recommended method for comparing models and algorithms (Hastie *et al.*, 2009; Kuhn and Johnson, 2013). The available data is divided to 10 approximately equal-sized partitions, and one of the partitions (at a time) is used for testing and comparing the predictions of the models estimated in the other nine partitions. A small part of the variance in the end result derives from the randomly chosen partitions; we study this variation by repeating the procedure 20 times with different random partitions. For point estimates, predictions in each 10 test dataset are combined to form a single confusion matrix (Hastie *et al.*, 2009; Kuhn and Johnson, 2013), with the performance measures computed from the matrix.

Our main performance criterion is the (a) YI that has several desirable qualities: it is independent of relative and absolute sizes of minority and majority classes, and methods with the same YI make the same total percentage of miss-classifications (Youden, 1950). Because we did not have a specific data to weight the costs per miss-classification to depressed versus non-depressed group, the equal-importance weighting of YI was preferred over available weighting methods (Kuhn and Johnson, 2013).

In addition, we report (b) sensitivity and (c) specificity, decomposing the YI for interpretation. The indices a–c suppress prevalence/base-rate information (Kuhn and Johnson, 2013). Therefore we also report (d) positive predictive value (PPV) that directly answers the question "what is the probability that a population-dwelling individual classified as depressed by the self-report-based model is also so diagnosed by CIDI?" In contrast, (e) negative predictive value (NPV) provides the probability of an individual classified as non-depressed to be similarly classified by CIDI as well. We also report (f) Kappa coefficient, commonly applied when studying inter-rater agreement for diagnostic interviews. Here it is to be interpreted as agreement between a prediction method and the CIDI outcome. In addition, we studied correlations among performance criteria; in order to understand which performance criteria are simple trade-offs or are equivalents regarding total performance variance in the study.

Statistical models

Many alternative class-predictive models currently exist in data-mining and explorative-statistics literature, especially for the more difficult non-linear predictive tasks (Hastie *et al.*, 2009; Kuhn and Johnson, 2013; Rasmussen and Williams, 2006). Our aim was to gradually build from very simple to more complex models in order to maximize

trackability, the "high point" being represented by AdaBoost algorithm (Culp *et al.*, 2006; Hastie *et al.*, 2009), which necessarily leaves a "zoo" of possible algorithms for future exploration (Hastie *et al.*, 2009; Rasmussen and Williams, 2006). The reason we chose to rely on methods based on classification trees, AdaBoost specifically, was that in comparison to other non-linear data-mining algorithms it offers a good level of both performance and interpretability, sometimes pitted as the "best off-the-shelf classifier in the world" (Hastie *et al.*, 2009). It is recognized that Kernel-based methods, such as Support Vector Machines and Gaussian Process Regression, are frequently found effective, but they can require rather complex Eigenfunction Analyses to understand why, as well as a very specific choice of kernel to achieve the good performance (Rasmussen and Williams, 2006). In these respects, they are less "off-the-shelf" in their application than regression trees. Furthermore, during pilot testing AdaBoost demonstrated solid performance when compared to other tree-based methods; thus, offering a justifiable baseline for future research. All computations were performed using the R software for statistical computing, Linux 64-bit version 3.0.2 (R Core Team, 2012). The individual models/methods, and the logic behind specific choices, are detailed later.

First, we examined whether single questionnaire items were predictive of CIDI diagnosis. This is an important piece of baseline information: instead of automatically assuming that a sum score (let alone a more complex combination of items) has incremental value over single items, it should be explicitly verified. We programmed a simple script that chooses the best-predicting item and a cutoff for it that together maximize the YI in the training data. The script was just an exhaustive "brute-force" search by two nested for-loops, one over items and another over item scores, or possible cutoffs (note all models were tested in separate validation data using the 10-fold cross-validation only after this "estimation" step).

Second, sum score of items was studied, being the most widely used summary of self-report data on depression, and thus, another important baseline reference. Here, as in a previous study (Aalto *et al.*, 2012), the most CIDI-predictive cutoff value was chosen to maximize the YI. The maximization was easily performed using "optimize" function from the base-package of the R software. The "optimize" function is a general-purpose optimizer, seeking a function maximum at an interval using a combination of golden section search and parabolic interpolation (Brent, 1973; R Core Team, 2012). A sum score weights each and every item in the inventory equally; this weighting can be questioned both theoretically (e.g. interview diagnoses use stem and core symptoms) and empirically (Bringmann *et al.*, 2014; Cramer *et al.*, 2012; Rosenström *et al.*, 2012, 2013; Wakefield and Schmitz, 2013).

Third, a Generalized Linear Model (GLM), specifically the logistic regression model in this case ("glm" function in the base package of R), was used, because it yields a linear predictor that differentially weights the items [it has been previously used to improve concordance of CIDI and SCID (Haro *et al.*, 2006)]. A CIDI diagnosis was predicted for test-data observations when the model-derived probability exceeded the YI maximizing cutoff derived from the training data (defined by the "optimize" function). The logistic model differentially weights the items, but does not consider their interactions. Structured interviews present screening questions, however, implying that a hierarchical interaction model might eventually outperform this linearly superimposing model (note the "interaction" used here refers to a general multidimensional non-linearity, found by exploration, not to the pre-set product terms familiar from basic regression modelling).

A classification tree constructs a hierarchical set of rules for predicting the diagnostic status from the questionnaire items. Classification trees are efficient at modelling interactions and imitate medical diagnostic thinking, but suffer from high variance and limited flexibility (Kuhn and Johnson, 2013). Their performance can be considerably boosted, by adding together several rules, each aiming to achieve a progressively more fine-grained prediction of information missed by previously estimated rules. This special case of the more general "boosting" method is known as boosted additive trees, or AdaBoost (Hastie *et al.*, 2009; Kuhn and Johnson, 2013). It has been demonstrated that AdaBoost works badly for imbalanced data, consequently balancing modifications were used for the fourth and fifth method.

Particularly well-balanced AdaBoost classifiers have been obtained by repeatedly randomly re-sampling a subset of the majority class (non-depressed) that is equal in sample size to the minority class (the depressed); this Easy Ensemble method (Liu *et al.*, 2009) was applied here too. The commonly used Synthetic Minority Over-sampling Technique (SMOTE) instead constructs both under-samples and strategically chosen "synthetic" new (over-)samples as the training data (Chawla *et al.*, 2002; Kuhn and Johnson, 2013), and it was also used. AdaBoost was implemented by "ada" (version 2.0.3) R package (Culp *et al.*, 2006) and SMOTE by "DMwR" (version 0.4.1) R package, whereas the Easy Ensemble was our own script implemented according to the paper of Liu *et al.* (2009); freely available from the first author's personal web page or by direct request (http://www.iki.fi/tom.rosenstrom/softw/BDItoCIDI.zip).

Twenty AdaBoost models formed the Easy Ensemble, and the AdaBoost models themselves used default

shrinkage parameter ($\eta = 0.1$) and 20 boosting iterations. The number of boosting iterations (within each member of the Easy Ensemble) was determined by 10-fold cross-validation, as the one that yielded highest YI for BDI-predicted two-week CIDI depression diagnosis. Both the number of iterations (added trees) and the shrinkage parameter (weight for added trees) regularize an AdaBoost solution, with effects of "opposite directions". Thus, η was held constant in cross-validation, because it exhibits a trade-off with the number of iterations parameter, and the default value is already empirically determined as being a generally good one. Logistic (a.k.a. divergence) loss function was used, because it is "far more robust" when misspecifications in class labels exist (Hastie *et al.*, 2009), as was the case here (CIDI inter-rater reliability < 1).

Methods that weight or choose variables used in prediction also provide a relative importance index for the variables: we plotted the linear weights from GLM coefficients and the importance of variables as a part of AdaBoost classification trees (Hastie *et al.*, 2009; Kuhn and Johnson, 2013). The GLM coefficients (or importances) stand for the natural logarithm of the item-associated multiplier for CIDI diagnosis odds, which is the standard interpretation of logistic-regression coefficients; more complex measure of importance need to be introduced for the AdaBoost. A classification tree is constructed by a process of successive "branching", where a cutoff is made at a particular variable that gives maximal estimated improvement in squared error risk compared to a constant fit over the available predictor space. A variables importance is then the square root of the sum of the squared improvements over the tree's internal nodes that represent a branch on that variable. The importance is generalized to additive trees (including AdaBoost) by averaging over individual trees (Hastie *et al.*, 2009), and further to the Easy Ensemble by averaging over the ensemble members; thus, the unit of a variable's importance for AdaBoost is the average improvement in squared error risk, derived from branches on the variable.

When reported, distributions of importances or model parameter estimates were based on 200 "bootstrap" resamples of the original data (Efron and Tibshirani, 1993; Hastie *et al.*, 2009), but the main interest was the average performance of the methods in the earlier-explained

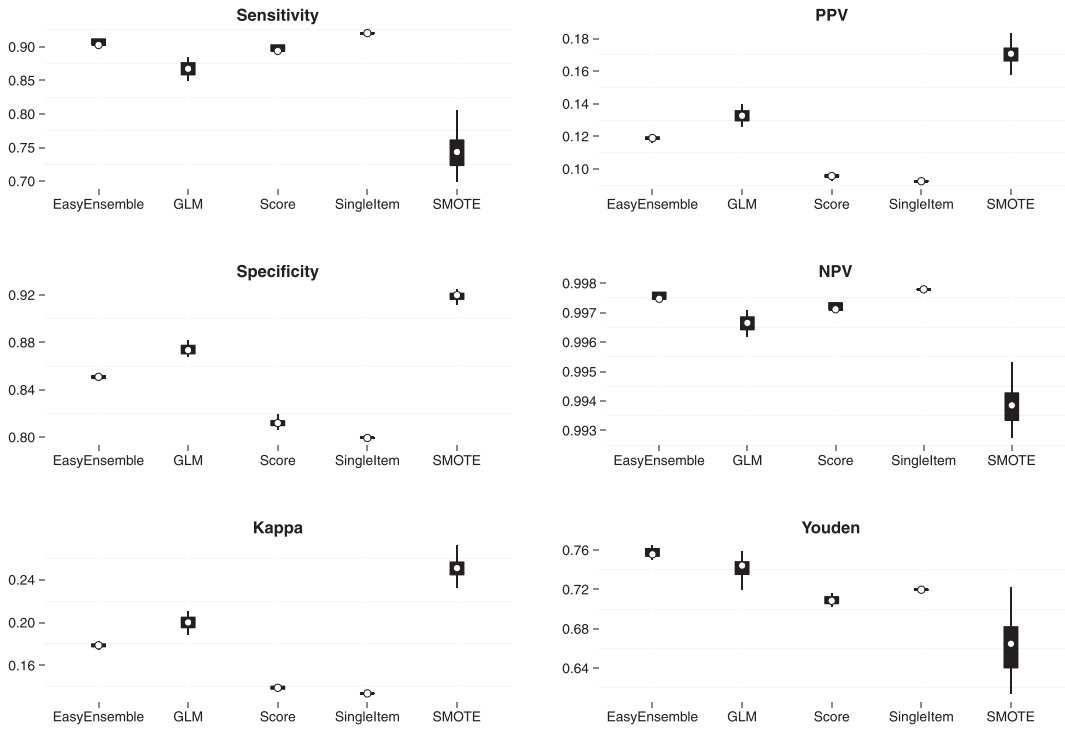


Figure 1. Performance metrics for predictive models: Beck's Depression Inventory items predicting depression diagnosis within two-week recency by Composite International Diagnostic Interview. Abbreviations: PPV, Positive Predictive Value; NPV, Negative Predictive Value; GLM, Generalized Linear (logistic regression) Model; Score, Sum Score of the items; SMOTE, Synthetic Minority Over-sampling Technique.

Int. J. Methods Psychiatr. Res. (2015). DOI: 10.1002/mpr

10-fold cross-validation. Because of the exploratory approach, results are presented by boxplots of entire resampling or cross-validation distributions instead of hypothesis tests and associated confidence intervals [note that good standard-error estimates require a factor of 10 less bootstrap resamples compared to confidence intervals (Efron and Tibshirani, 1993)]. In-text estimates are given as averages over the repeated cross-validation partitions, with their associated standard errors (SE).

Results

Altogether 5200 participants (46.9% men) had both the required CIDI information and a full set of BDI items: 2.3% obtained a two-week CIDI depression diagnosis and 1.2% obtained a two-week Pure MDD diagnosis, whereas 6.4% obtained a 12-month diagnosis and 4.3% obtained a 12-month Pure MDD diagnosis. The ages of the participants ranged from 30 to 79 years [mean = 50.0, standard deviation (SD) = 12.6]. Despite participants excluded due to lacking item data (n = 805), the prevalence numbers were close to a perfect agreement with previous studies (Aalto *et al.*, 2012; Pirkola *et al.*, 2005), and with the sample of 5435 participants who had the CIDI data and the GHQ items. For further sample details, see previous studies (Aalto *et al.*, 2012; Aromaa and Koskinen, 2003; Pirkola *et al.*, 2005).

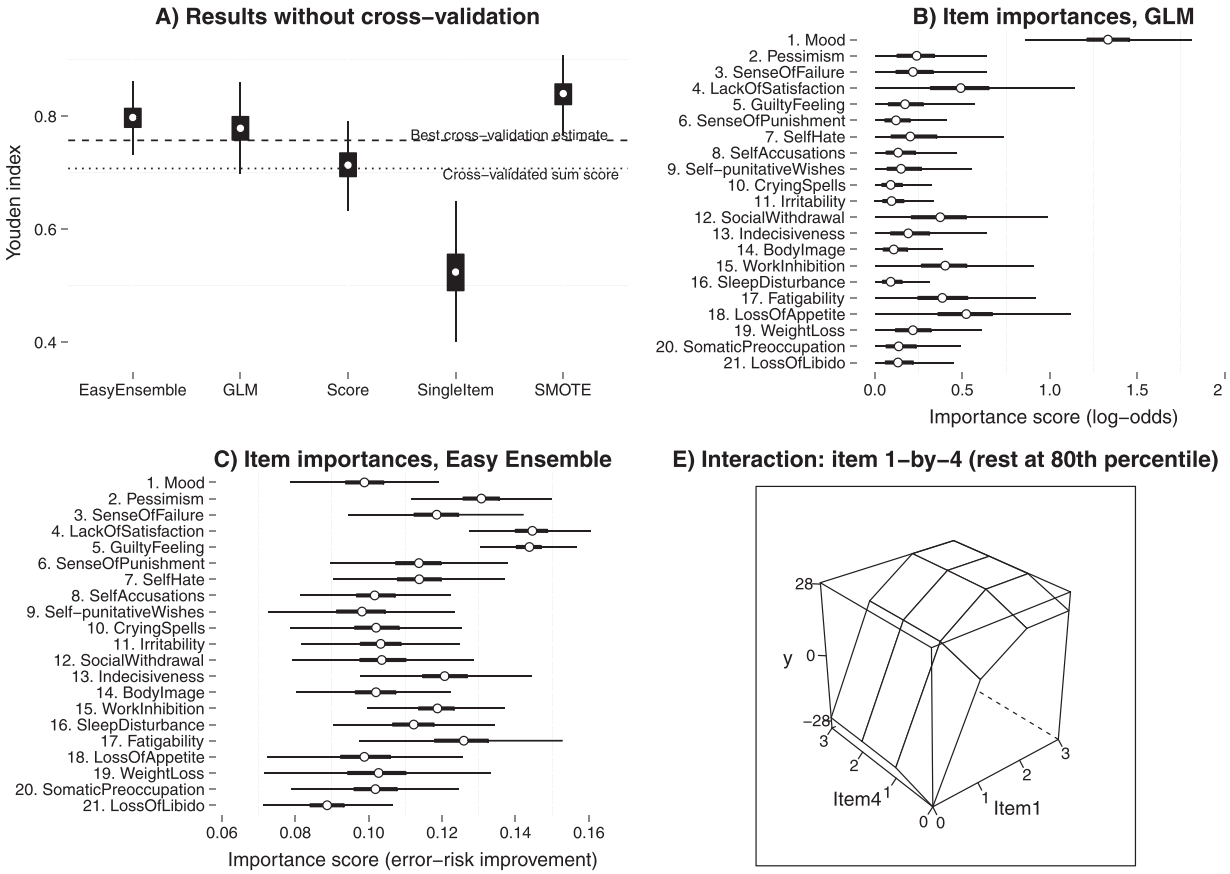


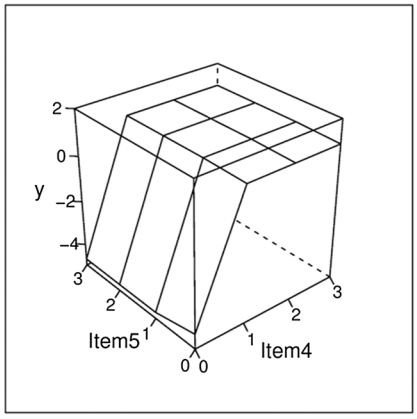
Figure 2. Analyses of predictive models. The models predict two-week diagnoses of depression using Beck's Depression Inventory items. (A) Youden's index (YI) from the Figure 1 when proper cross-validation is not used: 200 bootstrap replications of total data are shown instead. Dashed line shows the best cross-validated performance and dotted line the cross-validated performance of the sum score. (B) Bootstrap distributions of the item importances in the Generalized Linear Model (GLM). (C) Same as panel B but for the Easy Ensemble method. (D) Interaction plot for the two core-symptom items in the Easy Ensemble solution. The model's prediction surface is shown as a function of two items, with the other items being set to their 80th sample percentiles. The final predictions are thresholded from the surface by the sign function; e.g. sign (y) = 1 predicts depression.

Beck's Depression Inventory (BDI)

Figure 1 summarizes how well the BDI items were able to predict CIDI diagnosis of depression within two-week recency. The YI was highest for the Easy Ensemble method (0.757, SE = 0.001), with a sensitivity of 0.907 and specificity of 0.851. GLM method was nearly as good (0.741, SE = 0.002), and Single Item method ranked third. If one were to use the sample estimate instead of cross-validation, as done in most previous studies: the worst method (SMOTE) would appear as the best (Figure 2A)! The sum score estimates evaluated in the training data were almost unbiased, however. Even though sensitivities and specificities were acceptable for most methods (Figure 1), the PPV was low for all the methods, between 0.09–0.17. However, NPVs were above 0.99.

Figures 2B and 2C show item importance for GLM and Easy Ensemble methods, when predicting any depression diagnosis in the past two weeks using BDI items. GLM method, that linearly weights the item importance, utilized most the "Mood" item, whose scoring ranges from zero = "I do not feel sad" to three = "I am so sad or unhappy that I can't stand it". Also the Single Item method invariably used the "Mood" item. Easy Ensemble method, in contrast, most frequently made performance-improving classification-tree branches to the other core symptom, "Lack of Satisfaction", and to "Guilty Feeling". Despite the apparently low (non-linear) importance in

A) Item 4–by–5 (item 1 at 0, rest at 95th percentile}



B) Item 2–by–5 (items 1 & 4 at 1, rest at 60th percentile)

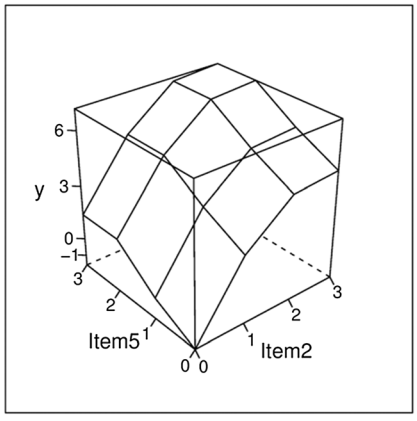


Figure 3. Further item-interaction plots. Different item configurations were explored for the Easy Ensemble model of Figure 2D. (A) Interaction between "Lack of Satisfaction" (item 4) and "Guilty Feeling" (item 5) was studied in a case where "Mood" item 1 was not endorsed, but the remaining items were a lot (i.e. they were set at 95th sample percentile; points where the surface exceeds zero represent predictions of Composite International Diagnostic Interview (CIDI) diagnosis by the model). (B) Interaction of "Pessimism" (item 2) and "Guilty Feeling" (item 5), when both the core symptoms (items 1 and 4) were mildly endorsed, and the remaining items to some extent as well.

classification tree branching, actual prediction surfaces of the Easy Ensemble model nonetheless strongly reflected the "Mood" item, as shown by the prediction surface of Figure 2D, whose sign yields the predicted classification.

Because of the importance scores (Figure 2C), however, we know that such configurations of the multidimensional item space must exist where the other core symptom, "Lack of Satisfaction", and the symptom "Guilty Feeling", make a difference. Indeed, Figure 3 shows two such configurations. For example, if one has many symptoms in general but no sad mood, then the Easy Ensemble model uses the other core symptom (item 4) in predicting CIDI diagnosis (Figure 3A); an expected outcome, as having at least one core symptom is a prerequisite for the diagnosis. "Guilty Feeling" and "Pessimism", however, predicted who had received a diagnosis when the core symptoms were only minimally endorsed (Figure 3B). Thus, general non-linear interactions of multiple variables appear to exist, explaining why AdaBoost with Easy Ensemble performed the best, however their effect on total prediction performance is of modest magnitude.

The pattern of BDI findings was similar for 12-month CIDI diagnoses as for the two-week diagnoses, but the performance was generally inferior, and clear differences between Easy Ensemble and GLM methods vanished (Supplementary Material Figure S1). Specificity declined when predicting 12-month Pure MDD compared to any depression diagnosis, as expected (e.g. 0.744 ± 0.001 for Pure MDD versus 0.804 ± 0.000 for any depression with the Easy Ensemble). Sensitivity declined less than specificity (0.757 ± 0.002 versus 0.780 ± 0.001). Also for the two-week diagnoses, predictions of Pure MDD had lower specificity than predictions of MDD (0.735 versus 0.848), but also clearly lower sensitivity (0.760 versus 0.871; note that only 1.2% of the participants had this condition, implying a highly imbalanced case).

General Health Questionnaire (GHQ)

Figure 4 shows the cross-validation results for the 12-item GHQ predicting CIDI diagnosis within two-week recency. In contrast to the BDI, for GHQ, the sum score and the Single Item method were the most informative ways to combine the items to a prediction of the CIDI diagnosis. Overall, the GHQ performed below BDI. The single most useful GHQ item in CIDI prediction was item 9, enquiring

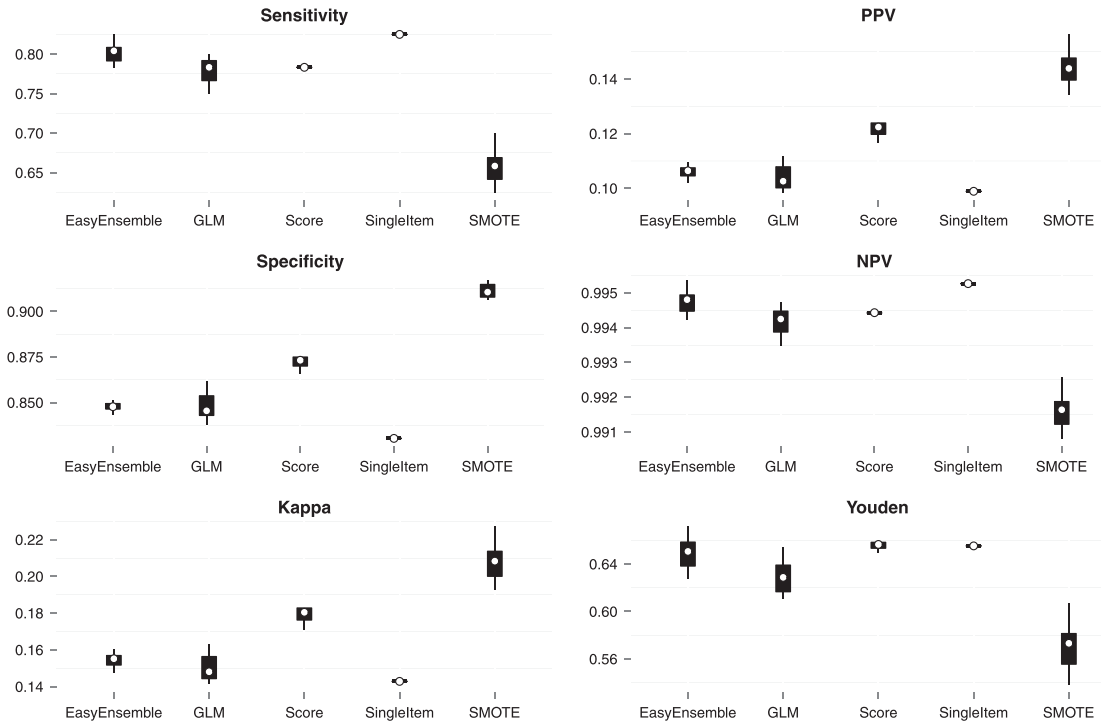


Figure 4. Performance metrics for predictive models: 12 General Health Questionnaire items predicting depression diagnosis within two-week recency by Composite International Diagnostic Interview. Abbreviations: PPV, Positive Predictive Value; NPV, Negative Predictive Value; GLM, Generalized Linear (logistic regression) Model; Score, Sum Score of the items; SMOTE, Synthetic Minority Over-sampling Technique.

"have you recently been feeling unhappy and depressed". The pattern of findings was similar for GHQ and 12-month CIDI diagnosis.

General patterns

Table 1 shows the correlations between performance metrics across the depression inventories, diagnostic intervals/recency, and methods. This analysis addressed which metrics varied together, representing either general trade-offs in ability to explain the data (negative correlations) or redundancy of information (positive correlations). One sees that a method's ability to boost sensitivity determined more of the general performance than its specificity. PPV and Kappa coefficient were highly related, that is, in this imbalanced task there is a lower probability of agreement given a prediction of depression compared to a prediction of no-depression.

Different item-scoring systems

While the Likert scoring of item-answering options (0–1–2–3) obviously contains more information, sometimes GHQ items are nevertheless scored by "lossy" coding systems known as the "usual" method (0–0–1–1) and the "C-GHQ" method that uses the same scoring as the usual method for items assessing health, but different scoring (0–1–1–1) for the items assessing illness (Goldberg *et al.*, 1997). Whereas for the Likert scales, the three best methods for predicting 2-week depression diagnosis were almost equally good (Easy Ensemble's YI = 0.650, Score's YI = 0.655, and Single Item's YI = 0.655), the usual scoring had only two clearly superior methods (GLM's YI = 0.654 and Single Item's YI = 0.655), and the C-GHQ scoring had just a single superior method, the total Score (YI = 0.640). In fact, for the Single Item method the exact same cutoff to the same mood item was induced by the "usual" scoring as by the Likert scoring, but the C-GHQ scoring forced a sub-optimal cutoff.

Discussion

Concordance between CIDI diagnoses and self-reports of depression has been studied extensively using sum scores of questionnaire items. This study assessed whether alternative, simpler or more complex, combinations of the items could supersede the sum-score approach. Regarding BDI, sum-score prediction could be improved by using (1) weighted sums of items, (2) nonlinear classification trees constructed from the items (Easy Ensemble), or (3) single item instead of the sum score of the items. For the 12-item GHQ, alternative methods could not supersede the sum score, but single item was as accurate a solution as the sum score. This is a non-trivial result, because GHQ was originally designed to detect general mental distress rather than depression per se; in our results, the general distress estimate appeared equally concordant with the specific CIDI depression diagnosis compared to the mood item that specifically probes for depressed mood.

The item assessing the core symptom of sad mood was the most important item in predicting a CIDI diagnosis of depression. Role of the other core symptom, anhedonia, was accentuated in nonlinear classification trees. Compared to BDI sum score, the new methods resulted in 1.07-fold Youden's index for 2-week diagnosis, and a 1.18-fold index for 12-month diagnosis. Predictions based on GHQ sum score were not significantly improved. With more complex methods it was important to (cross-)validate the performance criteria, but sum scores did not suffer from noticeable overfitting, meaning that the previous concordance estimates are reliable.

Table 1. Correlations and averages of the performance metrics over all the studied methods ($n = 5$), inventories ($n = 2$), and diagnostic recencies ($n = 2$, altogether 20 meta-observations)

	Sensitivity	Specificity	Kappa	PPV	NPV	Youden
Sensitivity	1	-	—	—	—	—
Specificity	−0.302	1	—	—	—	—
Kappa	−0.732	0.301	1	—	—	—
PPV	−0.785	0.134	0.972	1	—	—
NPV	0.837	0.162	−0.762	−0.884	1	—
Youden	0.902	0.138	−0.623	−0.754	0.943	1
Average	0.754	0.841	0.215	0.169	0.985	0.595
Standard deviation	0.100	0.043	0.058	0.060	0.011	0.097

Note: PPV, Positive Predictive Value; NPV, Negative Predictive Value; Youden, Youden Index.

The use of the several methods in the present analysis allowed us to also study which performance metrics generally offered room for improvement in the specific research question (high variance; e.g. sensitivity and PPV), which represented trade-offs difficult to overcome (negative correlations; e.g. PPV with Youden and sensitivity), and which were similarly affected by improvements (positive correlations; e.g. Youden, NPV and sensitivity). On this basis, the central challenge in the present imbalanced prediction problem is to provide methods that simultaneously obtain a good sensitivity without compromising Kappa and PPV (specificity was not easily compromised anyway).

Observed PPVs ranged from 0.09 to 0.31. This implies a less than 31% chance that a CIDI diagnosis predicted from the self-reported questionnaire items would be diagnosed by an interviewer as well. PPV is sensitive to base rate of depression: in a clinical study with half of the patients depressed, observed sensitivities and specificities (e.g. 0.91 and 0.85, respectively) would lead to a good PPV (i.e. 0.86). Kappa coefficient, that also takes base rates into account, was 0.18 between BDI (Easy Ensemble) and two-week CIDI in the present analysis, which is low compared to the $\kappa = 0.88$ agreement between ratings for MDD made by two different CIDI raters, reported previously in the same sample as used here (Pirkola *et al.*, 2005). An explanation is needed to understand the apparent conflict between claims for good instrument concordance (Aalto *et al.*, 2012; Goldberg *et al.*, 1997; Hewitt *et al.*, 2011; Nuevo *et al.*, 2009; Viinamäki *et al.*, 2004) and the less encouraging figures reported earlier.

The development of assessment tools is typically based on criteria deriving from sensitivity and specificity, which are conditional measures to a "gold-standard" observation. This conditioning ensures that the performance criteria are not sensitive to the base rate of events in the studied population, thus allowing an unbiased picture about the merits of the method itself, in any conceivable population (Kuhn and Johnson, 2013). Kappa coefficient, however, is designed to take into account the accuracy that would be generated simply by chance; it is sensitive to the base rate of events (Kuhn and Johnson, 2013). Both performance measures have their merits, but assessing inter-rater reliability by Kappa and inter-instrument concordance by sensitivity and specificity can result in a misleading double standard. Agreement between two raters using different instruments can appear too rosy to practitioners, because few studies directly assess this agreement by Kappa.

In addition to the earlier explanation, our comparisons had limitations that may decrease concordances to some extent: all self-report measures were not completed at the same date as the CIDIs, hindering prediction. It is also noted that elderly people may behave differently to young adults in the CIDI (O'Connor and Parslow, 2010). Furthermore, a new version of BDI, highly correlated with the presently used one (at $r = 0.94$), has been released (Beck *et al.*, 1996). Perhaps due to copyright issues, however, the older inventory was preferred by the Health 2000 Study designers; the data collaboration benefitted from a freely available online access to the item content. Finally, many alternative class-predictive methods exist in diverse fields of science (see, e.g. Hastie *et al.*, 2009; Kuhn and Johnson, 2013; Rasmussen and Williams, 2006), all of which could not be tested here; instead, we chose to concentrate on few well-justified methods (see Statistical models section).

Continuous measures are frequently found to be more valid and reliable than diagnostic thresholds (Haslam *et al.*, 2012; Markon *et al.*, 2011). But, this study demonstrated that the standard continuum theory for depression-questionnaire items is fundamentally incompatible with the diagnostic constructs of CIDI, because using a single BDI item consistently outperformed the sum score. It was better to throw away the information from more than 95% of the items than to use their sum, which grossly undermines the idea of the items "measuring" a latent continuum whose thresholds then would underlie the CIDI diagnoses. Several recent studies have questioned the latent-continuum assumption in general, not just in relation to the CIDI (Borsboom, 2008; Borsboom *et al.*, 2003; Bringmann *et al.*, 2014; Cramer *et al.*, 2012; Kendler *et al.*, 2011; Wigman *et al.*, 2013).

The GHQ's supposed latent continuum is "general mental distress" rather than "depression", therefore, one might have expected some GHQ items to supersede the GHQ sum score in depression-diagnosis prediction, but surprisingly, the sum score and the best (i.e. the mood) item fared equally well. If using the "usual" (0–0–1–1) item-scoring system (Goldberg *et al.*, 1997), then the single item indeed superseded the total score. For the "C-GHQ" (0–1–1–1 for illness-related and 0–0–1–1 for health-related items) item-scoring system (Goldberg *et al.*, 1997) the total Score was the best method, but only because the scoring system forced a sub-optimal cutoff to the mood item, decreasing the best item's performance. We did not find obvious reasons to introduce information loss by using the non-Likert coding systems, and would therefore advice against such practice. There has been plenty of work on GHQ predicting general presence of CIDI diagnoses (Caraveo-Anduaga *et al.*, 1998; Goldberg *et al.*, 1997, 1998), but detecting the specific diagnosis of depression has been studied to a lesser extent (but see Hewitt *et al.*, 2011). Based on this study, it seems that using the single mood item ("been feeling unhappy or depressed") with

the cutoff at score 2 ("rather more than usual") yields the most robust predictor for a CIDI depression diagnosis.

The fundamental incompatibility of the CIDI diagnoses and the questionnaire sum scores emphasizes the recent appeals for research endeavours aiming to find both psychometrically and causally justifiable basis for diagnostic systems (Borsboom, 2008; Kendler *et al.*, 2011; Nesse and Stein, 2012). Yet, all research must progress in manageable steps, and combining different sources of symptom and diagnostic information may well be among these manageable steps (Kendler *et al.*, 2011). To this end, we provide our estimated models: they may be used directly for maximizing the concordance of the CIDI and BDI, or as a reference for building even more concordant models (available at http://www.iki.fi/tom.rosenstrom/softw/BDItoCIDI.zip).

In summary, this study was able to improve the CIDI–BDI concordance achieved by previous studies by all the three tested means: (1) using just a single "Mood" item of the BDI, (2) using weighted linear model of all the items, and (3) using classification-tree-based non-linear models of all items. This also empirically demonstrated the lack of compatibility between the theoretical bases of questionnaire-item sum scores and interview diagnoses. Regardless of the method, the Kappa coefficients and PPVs of the item-based predictions did not support using self-reports as general proxies for CIDIs. For research samples with a high base rate of MDD, however, using item-based predictions may be useful and can have a high PPV; although, a separate validation study for special samples is generally recommended (Goldberg *et al.*, 1998).

Acknowledgements

This work was supported by the Academy of Finland (L.K. grant number 258711); (M.E. grant number 265977), and by the Emil Aaltonen Foundation (T.R., M.J.). The sponsor had no role in drafting of the manuscript or decision to submit. The authors are grateful to Abby Tabor for the expert language revision, and take full responsibility for any typographical errors that may have been introduced afterwards.

Declaration of interest statement

The authors have no competing interests.

References

Aalto A., Elovainio M., Kivimäki M., Uutela A., Pirkola S. (2012). The Beck Depression Inventory and General Health Questionnaire as measures of depression in the general population: a validation study using the Composite International Diagnostic Interview as the gold standard. *Psychiatry Research*, **197**(1), 163–171. DOI: 10.1016/j.psychres.2011.09.008

Aromaa A., Koskinen S. (2003) Health and Functional Capacity in Finland. Baseline Results of the Health 2000 Health Examination Survey, Helsinki: Publications of the National Public Health Institute.

Baca-García E., Perez-Rodriguez M., Basurte-Villamor I., Saiz-Ruiz J., Leiva-Murillo J.M., de Prado-Cumplido M., Santiago-Mozos R., Artés-Rodríguez A., de Leon J. (2006) Using data mining to explore complex clinical decisions: a study of hospitalization after a suicide attempt. *Journal of Clinical Psychiatry*, **67**(7), 1124–1132.

Beck A.T. (1967) Depression: Clinical, Experimental, and Theoretical Aspects. Philadelphia, PA: University of Pennsylvania Press.

Beck A.T., Ward C.H., Mendelson M., Mock J., Erbaugh J. (1961) An inventory for measuring depression. *Archives of General Psychiatry*, **4**(6), 561–571. DOI: 10.1001/archpsyc.1961.01710120031004

Borsboom D. (2008) Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, **64**(9), 1089–1108. DOI: 10.1002/jclp.20503

Borsboom D., Mellenbergh G.J., Heerden J. (2003) The theoretical status of latent variables. *Psychological Review*, **110**(2), 203–219. DOI: 10.1037/0033-295X.110.2.203

Brent, R. (1973) Algorithms for Minimization without Derivatives, Englewood Cliffs, NJ: Prentice Hall.

Bringmann L.F., Lemmens L.H.J.M., Huibers M.J.H., Borsboom D., Tuerlinckx F. (2014) Revealing the dynamic network structure of the Beck Depression Inventory – II. *Psychological Medicine*, **45**(4), 747–757. DOI: 10.1017/S0033291714001809

Bromet E., Andrade L., Hwang I., Sampson N., Alonso J., de Girolamo G., de Graaf R., Demyttenaere K., Hu C., Iwata N., Karam A.N., Kaur J., Kostyuchenko S., Lépine J.P.., Levinson D., Matschinger H., Mora M.E., Browne M.O., Posada-Villa J., Viana M.C., Williams D.R., Kessler R.C. (2011) Cross-national epidemiology of DSM-IV major depressive episode. *BMC Medicine*, **9**(1), 90. DOI: 10.1186/1741-7015-9-90

Caraveo-Anduaga J.J., Martinez N.A., Saldivar G., Lopez J.L., Saltijeral M.T. (1998) Performance of the GHQ-12 in relation to current and lifetime CIDI psychiatric diagnoses – GHQ-12 in relation to CIDI diagnoses. *Salud Mental*, **21**(4), 1–11.

Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. (2002) SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321–357. DOI: 10.1613/jair.953

Cramer A.O.J., Borsboom D., Aggen S.H., Kendler K.S. (2012) The pathoplasticity of dysphoric episodes: differential impact of stressful life events on the pattern of depressive symptom inter-correlations. *Psychological Medicine*, **42**(5), 957–965. DOI: 10.1017/S003329171100211X

Culp M., Johnson K., Michailidis G. (2006) ada: an R package for stochastic boosting. *Journal of Statistical Software*, **17**(2), 1–27.

Efron B., Tibshirani R.J. (1993) An Introduction to the Bootstrap. Boca Raton, CA: Chapman & Hall/CRC.

Goldberg D.P., Gater R., Sartorius N., Ustun T.B., Piccinelli M., Gureje O., Rutter C. (1997) The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychological Medicine*, **27**(1), 191–197.

Goldberg D.P., Oldehinkel T., Ormel J. (1998) Why GHQ threshold varies from one place to another. *Psychological Medicine*, **28**, 915–921.

Haro J.M., Arbabzadeh-Bouchez S., Brugha T.S., De Girolamo G., Guyer M.E., Jin R., Lepine J.P., Mazzi F., Reneses B., Vilagut G., Sampson N.A., Kessler R.C. (2006) Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health Surveys. *International Journal of Methods in Psychiatric Research*, **15**(4), 167–180. DOI: 10.1002/mpr.196

Haslam N., Holland E., Kuppens P. (2012) Categories versus dimensions in personality and psychopathology: a quantitative review of taxometric research. *Psychological Medicine*, **42**(5), 903–920. DOI: 10.1017/S0033291711001966

Hastie T., Tibshirani R., Friedman J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second edition, New York: Springer-Verlag.

Hewitt C.E., Perry A.E., Adams B., Gilbody S.M. (2011) Screening and case finding for depression in offender populations: a systematic review of diagnostic properties. *Journal of Affective Disorders*, **128**(1–2), 72–82. DOI: 10.1016/j.jad.2010.06.029

Kendler K.S., Zachar P., Craver C. (2011) What kinds of things are psychiatric disorders? *Psychological Medicine*, **41**(6), 1143–1150. DOI: 10.1017/S0033291710001844

Kuhn M., Johnson K. (2013) Applied Predictive Modeling, New York: Springer.

Liu X., Wu J., Zhou Z. (2009) Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics*, **39**(2), 539–550. DOI: 10.1109/TSMCB.2008.2007853

Lobo J.M., Jiménez-Valverde A., Real R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**(2), 145–151. DOI: 10.1111/j.1466-8238.2007.00358.x

Markon K.E., Chmielewski M., Miller C.J. (2011) The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review. *Psychological Bulletin*, **137**(5), 856–879. DOI: 10.1037/a0023678

Nesse R.M., Stein D.J. (2012) Towards a genuinely medical model for psychiatric nosology. *BMC Medicine*, **10**(1), 5. DOI: 10.1186/1741-7015-10-5

Nuevo R., Lehtinen V., Reyna-Liberato P.M., Ayuso-Mateos J.L. (2009) Usefulness of the Beck Depression Inventory as a screening method for depression among the general population of Finland. *Scandinavian Journal of Public Health*, **37**(1), 28–34. DOI: 10.1177/1403494808097169

O'Connor D.W., Parslow R.A. (2010) Differences in older people's responses to CIDI's depression screening and diagnostic questions may point to age-related bias. *Journal of Affective Disorders*, **125**(1–3), 361–364. DOI: 10.1016/j.jad.2010.01.008

Pirkola S.P., Isometsä E., Suvisaari J., Aro H., Joukamaa M., Poikolainen K., Koskinen S., Aromaa A., Lönnqvist J.K. (2005) DSM-IV mood-, anxiety- and alcohol use disorders and their comorbidity in the Finnish general population. *Social Psychiatry and Psychiatric Epidemiology*, **40**(1), 1–10. DOI: 10.1007/s00127-005-0848-7

Rasmussen C.E., Williams C.K.I. (2006) Gaussian Processes for Machine Learning, Cambridge: MA, The MIT Press.

R Core Team (2012) R: A Language and Environment for Statistical Computing, Vienna: R Foundation for Statistical Computing.

Rosenström T., Jokela M., Hintsanen M., Josefsson K., Juonala M., Kivimäki M., Pulkki-Råback L., Viikari J.S.A., Hutri-Kähönen N., Heinonen E., Raitakari O.T., Keltikangas-Järvinen L. (2013) Body-image dissatisfaction is strongly associated with chronic dysphoria. *Journal of Affective Disorders*, **150**(2), 253–260. DOI: 10.1016/j.jad.2013.04.003

Rosenström T., Jokela M., Puttonen S., Hintsanen M., Pulkki-Råback L., Viikari J.S., Raitakari O.T., Keltikangas-Järvinen L. (2012) Pairwise measures of causal direction in the epidemiology of sleep problems and depression. *PLoS ONE*, **7**(7), e50841. DOI: 10.1371/journal.pone.0050841

Viinamäki H., Tanskanen A., Honkalampi K., Koivumaa-Honkanen H., Haatainen K., Kaustio O., Hintikka J. (2004) Is the Beck Depression Inventory suitable for screening major depression in different phases of the disease? *Nordic Journal of Psychiatry*, **58**(1), 49–53. DOI: 10.1080/08039480310000798

Vos T., Flaxman A.D., Naghavi M., Lozano R., Michaud C., Ezzati M., Shibuya K., Salomon J.A., Abdalla S., Aboyans V., Abraham J., Ackerman I., Aggarwal R., Ahn S.Y., Ali M.K., Alvarado M., Anderson H.R., Anderson L.M., Andrews K.G., Atkinson C., Baddour L.M., Bahalim A.N., Barker-Collo S., Barrero L.H., Bartels D.H., Basáñez M.G., Baxter A., Bell M.L., Benjamin E.J., Bennett D., Bernabé E., Bhalla K., Bhandari B., Bikbov B., Bin Abdulhak A., Birbeck G., Black J.A., Blencowe H., Blore J.D., Blyth F., Bolliger I., Bonaventure A., Boufous S., Bourne R., Boussinesq M., Braithwaite T., Brayne C., Bridgett L., Brooker S., Brooks P., Brugha T.S., Bryan-Hancock C., Bucello C., Buchbinder R., Buckle G., Budke C.M., Burch M., Burney P., Burstein R., Calabria B., Campbell B., Canter C.E., Carabin H., Carapetis J., Carmona L., Cella C., Charlson F., Chen H., Cheng A.T., Chou D., Chugh S.S., Coffeng L.E., Colan S.D., Colquhoun S., Colson K.E., Condon J., Connor M.D., Cooper L.T., Corriere M., Cortinovis M., de Vaccaro K.C., Couser W., Cowie B.C., Criqui M.H., Cross M., Dabhadkar K.C., Dahiya M., Dahodwala N., Damsere-Derry J., Danaei G., Davis A., De Leo D., Degenhardt L., Dellavalle R., Delossantos A., Denenberg J., Derrett S., Des Jarlais D.C., Dharmaratne S.D., Dherani M., Diaz-Torne C., Dolk H., Dorsey E.R., Driscoll T., Duber H., Ebel B., Edmond K., Elbaz A., Ali S.E., Erskine H., Erwin P.J., Espindola P., Ewoigbokhan S.E., Farzadfar F., Feigin V., Felson D.T., Ferrari A., Ferri C.P., Fèvre E.M., Finucane M.M., Flaxman S., Flood L., Foreman K., Forouzanfar M.H., Fowkes F.G., Franklin R., Fransen M., Freeman M.K., Gabbe B.J., Gabriel S.E., Gakidou E., Ganatra H.A., Garcia B., Gaspari F., Gillum R.F., Gmel G., Gosselin R., Grainger R., Groeger J., Guillemin F., Gunnell D., Gupta R., Haagsma J., Hagan H., Halasa Y.A., Hall W., Haring D., Haro J.M., Harrison J.E., Havmoeller R., Hay R.J., Higashi H., Hill C., Hoen B., Hoffman H., Hotez P.J., Hoy D., Huang J.J., Ibeanusi S.E., Jacobsen K.H., James S.L., Jarvis D., Jasrasaria R., Jayaraman S., Johns N., Jonas J.B., Karthikeyan G., Kassebaum N., Kawakami N., Keren A., Khoo J.P., King C.H., Knowlton L.M., Kobusingye O., Koranteng A., Krishnamurthi R., Lalloo R., Laslett L.L., Lathlean T., Leasher J.L., Lee Y.Y., Leigh J., Lim S.S., Limb E., Lin J.K., Lipnick M., Lipshultz S.E., Liu W., Loane M., Ohno S.L., Lyons R., Ma J., Mabweijano J., MacIntyre M.F., Malekzadeh R., Mallinger L., Manivannan S., Marcenes W., March L., Margolis D.J., Marks G.B., Marks R., Matsumori A., Matzopoulos R., Mayosi B.M., McAnulty J.H., McDermott M.M., McGill N., McGrath J., Medina-Mora M.E., Meltzer M., Mensah G.A., Merriman T.R., Meyer A.C., Miglioli V., Miller M., Miller T.R., Mitchell P.B., Mocumbi A.O., Moffitt T.E., Mokdad A.A., Monasta L., Montico M., Moradi-Lakeh M., Moran A., Morawska L., Mori R., Murdoch M.E., Mwaniki M.K., Naidoo K., Nair M.N., Naldi L., Narayan K.M., Nelson P.K., Nelson R.G., Nevitt M.C., Newton C.R., Nolte S.,

Norman P., Norman R., O'Donnell M., O'Hanlon S., Olives C., Omer S.B., Ortblad K., Osborne R., Ozgediz D., Page A., Pahari B., Pandian J.D., Rivero A.P., Patten S.B., Pearce N., Padilla R.P., Perez-Ruiz F., Perico N., Pesudovs K., Phillips D., Phillips M.R., Pierce K., Pion S., Polanczyk G.V., Polinder S., Pope C.A. 3rd, Popova S., Porrini E., Pourmalek F., Prince M., Pullan R.L., Ramaiah K.D., Ranganathan D., Razavi H., Regan M., Rehm J.T., Rein D.B., Remuzzi G., Richardson K., Rivara F.P., Roberts T., Robinson C., De Leòn F.R., Ronfani L., Room R., Rosenfeld L. C., Rushton L., Sacco R.L., Saha S., Sampson U., Sanchez-Riera L., Sanman E., Schwebel D. C., Scott J.G., Segui-Gomez M., Shahraz S., Shepard D.S., Shin H., Shivakoti R., Singh D., Singh G.M., Singh J.A., Singleton J., Sleet D. A., Sliwa K., Smith E., Smith J.L., Stapelberg N.J., Steer A., Steiner T., Stolk W.A., Stovner L.J., Sudfeld C., Syed S., Tamburlini G., Tavakkoli M., Taylor H.R., Taylor J.A., Taylor W.J., Thomas B., Thomson W.M., Thurston G.D., Tleyjeh I.M., Tonelli M., Towbin J.A., Truelsen T., Tsilimbaris M.K., Ubeda C., Undurraga E.A., van der Werf M.J., van Os J., Vavilala M.S., Venketasubramanian N., Wang M., Wang W., Watt K., Weatherall D.J., Weinstock M.A., Weintraub R., Weisskopf MG, Weissman M.M., White R.A., Whiteford H., Wiersma S.T., Wilkinson J.D., Williams H.C., Williams S.R., Witt E., Wolfe F., Woolf A.D., Wulf S., Yeh P.H., Zaidi A.K., Zheng Z.J., Zonies D., Lopez A.D., Murray C.J., AlMazroa M.A., Memish Z.A. (2013) Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, **380**(9859), 2163–2196. DOI: 10.1016/S0140-6736(12)61729-2

Wakefield J.C., Schmitz M.F. (2013) When does depression become a disorder? Using recurrence rates to evaluate the validity of proposed changes in major depression diagnostic thresholds. *World Psychiatry*, **12**(1), 44–52. DOI: 10.1002/wps.20015

Wardenaar K.J., van Loo H.M., Cai T., Fava M., Gruber M.J., Li J., de Jonge P., Nierenberg A.A., Petukhova M.V., Rose S., Sampson N.A., Schoevers R.A., Wilcox M.A., Alonso J., Bromet E.J., Bunting B., Florescu S.E., Fukao A., Gureje O., Hu C., Huang Y.Q., Karam A.N., Levinson D., Medina Mora M.E., Posada-Villa J., Scott K.M., Taib N.I., Viana M.C., Xavier M., Zarkov Z., Kessler R.C. (2014) The effects of co-morbidity in defining major depression subtypes associated with long-term course and severity. *Psychological Medicine*, **44**(15), 3289–3302. DOI: 10.1017/S0033291714000993

Weissman M.M., Bland R.C., Canino G.J., Faravelli C., Greenwald S., Hwu H.G., Joyce P.R., Karam E.G., Lee C.K., Lellouch J., Lepine J., Newman S.C., Rubio-Stibec M., Wells E., Wickramaratne P.J., Wittchen H., Yeh E. (1996) Cross-national epidemiology of major depression and bipolar disorder. *JAMA*, **276**(4), 293–299.

Wigman J.T.W., van Os J., Thiery E., Derom C., Collip D., Jacobs N., Wichers M. (2013) Psychiatric diagnosis revisited: towards a system of staging and profiling combining nomothetic and idiographic parameters of momentary mental states. *PLoS ONE*, **8**(3), e59559. DOI: 10.1371/journal.pone.0059559

Wittchenn H.U., Pfister H. (1997) DIA-X-Interviews: Manual fur screening-verfahren und Interview; Interviewheft Langsschnittuntersuchung (DIA-X-Lifetime); Ergan- zungsheft (DIAX- Lifetime); Interviewheft Querschnittuntersuchung (DIA-X-12 Monate); Erganzungsheft (DIA-X-12 Monate); PC-Programm zur Durchfuhrung des Interviews (Langs- und Querschnittuntersuchung); Auswertungsprogramm, Frankfurt: Swets and Zeitlinger.

Youden W.J. (1950) Index for rating diagnostic tests. *Cancer*, **3**(1), 32–35.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.