Applications of Bayesian ROC Analysis in Psychometric Research

Authors: Liu Yuqing, Li Huiling, Zhou Qiang, Zhou Qiang

Date: 2021-04-24T00:00:00+00:00

Abstract

ROC (receiver operating characteristic) analysis is an important and widely applied method in diagnostic research. Although it has been extensively used in diagnostic research in recent years, its application in psychological measurement research has yet to be reported domestically. Within ROC analysis methods, this article primarily introduces the application of Bayesian theory-based ROC analysis (BROC) in psychological measurement. Accordingly, this article first reviews the development and evolution of ROC analysis methods, then systematically examines the application of BROC in psychological measurement with example simulations, and finally discusses its prospective applications in the field of psychological measurement.

Full Text

Application of Bayesian ROC Analysis in Psychometric Research

Liu Yuqing, Li Huiling, Zhou Qiang*
Department of Applied Psychology, School of Mental Medicine, Wenzhou Medical University, Wenzhou, 325000

Abstract

ROC (Receiver Operating Characteristic) analysis represents an important and widely applied method in diagnostic research. Although it has been extensively used in diagnostic studies in recent years, its application in psychological measurement research remains unseen domestically. This paper primarily introduces the application of Bayesian theory-based ROC analysis (BROC) in psychometrics. To this end, we first review the development and evolution of ROC analysis methods, then systematically examine BROC applications in psychological

measurement with practical simulations, and finally discuss its prospective applications in the field.

Keywords: ROC analysis; Bayesian; psychometrics; diagnostic testing; accuracy assessment

Psychological research commonly employs physiological or psychological indicators to quantify mental states and/or traits for predicting and controlling related behaviors. Consequently, the accuracy assessment of these indicators constitutes a critical prerequisite for ensuring predictive validity. Questionnaires and behavioral experiments are frequently used measurement tools in psychological and behavioral research—for instance, the Big Five Personality Inventory to reflect personality traits (Lui et al., 2020), the Beck Depression Inventory to assess individual depression levels (Macchi et al., 2020), and ERP to investigate internal mental states (Cui et al., 2021). The accuracy of psychometric tools is essential for research validity. However, previous psychological studies have typically relied on reliability and validity coefficients to reflect measurement effectiveness, yielding relatively singular results that cannot intuitively demonstrate predictive value or directly compare accuracy across different measurement tools. Therefore, developing superior methods to evaluate psychometric accuracy has become an urgent necessity.

ROC analysis originated in signal detection theory (SDT), initially developed for radar monitoring and later applied to behavioral responses such as sensory thresholds (e.g., auditory, visual, and tactile). Today, it is widely used in psychological and neuroscience experiments (Sumner et al., 2019) and various other fields including medical diagnostics and machine learning (Obuchowski & Bullen, 2018; Ma et al., 2019), with implementation available through the plotROC package in R (Sachs, 2017). In recent years, international research applying ROC analysis to assess psychometric tool accuracy has grown substantially (Ruddy et al., 2018; Bowers et al., 2019; Thapa et al., 2020), primarily adapting time-dependent ROC (tROC) and Bayesian ROC (BROC) methods from diagnostic research for psychological applications. For example, Levis and Sun (2020) used ROC analysis to compare the diagnostic accuracy of depression screening scales PHQ-2, PHQ-9, and their combined diagnosis. ROC analysis also facilitates binary conversion to identify optimal cutoff points for obtaining additional needed information. Richardson (2018) utilized ROC analysis to determine optimal thresholds for the Problematic Smartphone Use Scale (PSUS), calculating AUC to evaluate PSUS accuracy and using cutoff points to find optimal critical values for continuous outcomes.

With advances in computer technology, ROC analysis in medical diagnostics has evolved to relax gold standard requirements—from binary and ordinal gold standards to scenarios with no gold standard—while incorporating more covariates, such as time-dependent ROC analysis and Bayesian methods for gold-standard-free conditions. The latter enables diagnostic evaluation without gold standards, completely overcoming the traditional ROC analysis barrier requiring gold standard presence, thus providing possibilities for research lacking gold standards.

This development offers important insights for ROC applications in psychometric accuracy assessment. Meanwhile, although ROC analysis has long been applied in psychology, its use in domestic psychological measurement remains limited.

Based on this context, this paper first briefly summarizes existing ROC analysis methods, particularly Bayesian ROC (BROC), then reviews its specific applications in psychometrics, and proposes future development prospects. The aim is to "transplant" BROC methodology into psychometrics, thereby expanding its application scope in psychological research, particularly in psychological measurement.

2. Overview of Common ROC Methods

The ROC curve is a graphical plot with 1-specificity on the x-axis and sensitivity on the y-axis (see Figure 1 [Figure 1: see original paper]), primarily using cutoff points and area under the curve (AUC) to reflect diagnostic outcomes (Mandrekar et al., 2010).

The cutoff point, representing the tangent value at the curve' s inflection point, is clinically selected as the value corresponding to the maximum Youden index —the optimal cutoff—for classifying test results as positive or negative. The Youden index represents a diagnostic method' s overall ability to distinguish patients from non-patients (sensitivity + specificity - 1), with higher values indicating better screening effectiveness and greater validity (Martínez-Camblor et al., 2019).

AUC is formally defined as: AUC = y(x)dx, representing the average sensitivity across all possible specificity values. Tests with higher AUC are considered more accurate. However, AUC has low sensitivity to indicator changes, making comparisons based solely on AUC insufficient; thus, reference sensitivity and specificity values must also be considered (Janssens & Martens, 2020).

Traditional ROC analysis for diagnostic evaluation typically employs the Yerushalmy paradigm, which compares measured results against a gold standard. This requires a reliable, stable binary gold standard; otherwise, sensitivity and specificity cannot be calculated, precluding accuracy evaluation. Although crucial for ROC analysis, obtaining a stable, appropriate binary gold standard is challenging. Many clinical disease gold standards are not binary but ordinal or continuous variables. Moreover, some gold standards are prohibitively expensive, procedurally complex, ethically problematic, or simply lack established standards, severely limiting ROC analysis applications (Wang et al., 2019). To address these issues, researchers have used expert experience to subjectively convert ordinal variables into binary ones. For instance, Numan et al. used expert experience to merge three categories into two (Numan et al., 2019) for ROC analysis, but subjective conversion introduces substantial error. Alternatively, Chen et al. divided ordinal variables into multiple binary groups for pairwise comparison, providing direction for ordinal gold standard research,

but this essentially only extends AUC application without utilizing other ROC information such as cutoff values (Chen, 2012). In fact, as early as the late 20th century, researchers (Peng et al., 1996) introduced Bayesian theory into ROC analysis to enable gold-standard-free evaluation. Unlike the traditional Yerushalmy paradigm, this approach leverages Bayesian theory, focusing not on finding gold standards but on collecting prior information combined with clinically validated relevant information while incorporating multiple covariates to effectively estimate posterior distributions. Flor et al. demonstrated that Bayesian estimation methods outperform traditional frequentist approaches (Flor et al., 2020).

Based on different gold standard characteristics, this paper summarizes three common new methods: ROC analysis under ordinal gold standard conditions, time-dependent ROC analysis, and ROC analysis without gold standards. The following sections detail these three approaches, including gold standard characteristics, clinical applications, and evaluations.

2.1 ROC Analysis Under Ordinal Gold Standard Conditions

ROC analysis under ordinal gold standard conditions can evaluate diagnostic accuracy for ordinal or continuous data and convert ordinal variables into binary ones as required. The basic process involves pairwise comparison of data across ordinal states, calculating AUC for each comparison, and finally comparing AUCs to achieve evaluation (Chen, 2012; Obuchowski et al., 2005). For example, in evaluating the diagnostic value of oxidized low-density lipoprotein ELISA kits for coronary heart disease, Chen divided subjects into three categories (diseased, non-diseased, suspicious) based on gold standard. AUC estimation and comparison can be implemented using the nonbinROC package in R (Paul Nguyen et al., 2007), with more package details and operation methods available in that study.

2.2 Time-Dependent ROC Analysis (tROC)

tROC analysis extends the concepts of sensitivity and specificity by observing disease status at each time point, generating different sensitivities and specificities to obtain a time-dependent ROC curve (Kamarudin et al., 2017). Additionally, it can directly obtain AUC at different time points, yielding an AUC(t) function plot for intuitive and effective comparison of accuracy across different observation times for the same or different measurement indicators. This method was first proposed by Heagerty and Zheng (2005), whose research utilized three different definitions—cumulative sensitivity and dynamic specificity (C/D), incident sensitivity and dynamic specificity (I/D), and incident sensitivity and static specificity (I/S)—to evaluate sensitivity and specificity for time-dependent events, applicable to different contexts.

tROC analysis can observe continuous disease states in individuals, incorporate information about disease onset time, construct ROC curves across time

points, and compare predictive capabilities of various measurement indicators. It has broad clinical applications. For example, Suzuki et al. used survival analysis to evaluate the prognostic impact of SIS and mGPS, employing time-dependent ROC analysis to compare the prognostic effects of various scoring systems (Suzuki et al., 2018). Lima et al. used ROC methods combining oscillating gradient spin echo (OGSE) and pulsed gradient spin echo (PGSE) with different diffusion times to explore ADC value changes in differentiating benign and malignant head and neck tumors (Lima et al., 2019). tROC can be implemented through R packages, with specific references available in Díaz-Coto et al. (2020).

2.3 Bayesian Theory-Based ROC Analysis Without Gold Standards (BROC)

The aforementioned ROC methods depend on gold standards, yet many diseases lack gold standards or have prohibitively expensive ones in clinical practice. Consequently, Peng (1996) proposed introducing Bayesian theory into ROC analysis, enabling consideration of multiple covariates and calculation of AUC under different covariate influences even without gold standards, thereby comparing diagnostic accuracy.

Bayesian theory differs from frequentist statistics by treating probability as subjective and 主张将个体经验信息作为重要部分 incorporating individual experiential information as a crucial component to derive posterior distributions. The fundamental principle combines prior distributions with sample likelihood functions to derive posterior distributions—posterior probability equals prior probability multiplied by likelihood value. With recent computer technology advances, Bayesian theory has been widely applied across many fields, particularly in medical diagnostic research and psychometric accuracy assessment (Arora & Thorlund, 2019; Goyal & Yolcu, 2019; Park & Lee, 2019). In diagnostic accuracy evaluation studies, the first and most critical step is determining prior information based on target population data. Subsequently, prior distributions are adjusted through likelihood functions to derive posterior distributions, enabling estimation of sensitivity and specificity for relevant diagnostic methods; see McClean et al. (2014) for more details.

Therefore, for diagnostic test evaluation without gold standards, as long as there is some prior information about the diagnostic test combined with clinically validated current observational data (though not gold standards), Bayesian theory can derive posterior distributions for diagnostic evaluation indicators, thus eliminating dependence on gold standards. For example, Amini (2020) used Bayesian latent class models (LCMs) to link diagnostic test observations with latent disease states, assessing diagnostic accuracy without perfectly accurate disease classification. Additionally, BROC can simultaneously consider multiple covariate influences. Compared to previous methods, it essentially liberates ROC analysis from gold standard constraints, expanding its applications across medicine, psychology, computer science, and other fields. For instance, Zi-Hui

Tang (2014) used Bayesian models to evaluate baroreflex sensitivity (BRS) for predicting cardiovascular autonomic neuropathy (CAN). Without a CAN gold standard, 2,092 suspected cases were selected with age, blood pressure, and other covariates, using BRS as the diagnostic criterion and Bayesian latent class models to assess BRS sensitivity and specificity. Results showed BRS had high sensitivity and specificity in CAN diagnostic testing, suggesting it as a valuable diagnostic tool (Zi-Hui Tang et al., 2014). As early as 2012, QiuWang et al. proposed applying BROC analysis in education and psychology, combining Bayesian hierarchical models with ROC analysis to assess how interest strength (IS) and interest differentiation (ID) predict interest-major congruence (IMC) in low socioeconomic status (SES) youth (QiuWang et al., 2012). This paper only introduces Bayesian methods for diagnosis without gold standards, though Bayesian theory applications extend far beyond, including deep learning, latent variable modeling, multilevel structural modeling, and experimental data analysis. With increasing interdisciplinary integration, Bayesian theory applications continue to grow. However, Bayesian models are not perfect, as their emphasis on experiential information can introduce subjective bias and affect result accuracy.

Overall, ROC analysis is a comprehensive method for accurately assessing diagnostic accuracy and predictive value, widely applied in medicine and psychology. In recent years, building on traditional binary gold standard ROC analysis, new methods have been developed for different clinical conditions, with research results fully demonstrating their validity.

3. BROC Applications in Psychometrics

As noted in the introduction, although ROC analysis has been applied in psychology for many years, its use has been largely limited to sensory thresholds and cognitive processing research, fundamentally constrained by binary gold standards. However, with further development in computer science, ROC analysis has made significant advances, particularly in diagnostic research. This paper therefore focuses on "transplanting" ROC methodology into psychology, especially the insights offered by Bayesian ROC analysis, and summarizes existing relevant research.

3.1 Quantifying Predictive Value (Accuracy) of Psychometric Tools

In clinical psychology applications, measurement results often require classification to facilitate yes/no or presence/absence judgments. For example, in psychological selection testing, continuous result data must be classified to determine whether candidates meet organizational requirements; this is particularly crucial in mental illness measurement, such as using depression scale scores compared against specific values to diagnose depression. Previous studies often used means or medians for binary conversion, while in mental illness diagnosis, such as depression scales, fixed scores are typically used for classification. However, such

classification accuracy cannot be evaluated. BROC analysis can obtain threshold values (cut-off) from curve inflection point tangents and identify optimal cut-off values using the Youden index to dichotomize continuous variables. As the best classification indicator in diagnostic research for decades, cut-off values hold substantial potential for binary conversion in psychology. For instance, in depression, anxiety, and obsessive-compulsive disorder assessments, ROC curves can be generated from test results, and combined with physician judgment, can inform diagnostic decisions. Beyond clinical diagnosis, ROC analysis is also suitable for universal psychological screening. For example, Battaglia et al. used ROC analysis to obtain optimal cutoff points for ESAS physical, psychological, and global subscales and compare ESAS scores with KTR, ICD-10, and DCPR diagnoses (Battaglia et al., 2020). Thapa et al. used ROC curve analysis to evaluate the accuracy of three-dimensional psychological pain (DPPS) in detecting high suicide risk among depressed patients with suicidal ideation and attempts (Thapa et al., 2020).

Questionnaires, as one of psychology' s most commonly used measurement tools, are widely applied in trait measurement and mental illness diagnosis research. Assessing their accuracy and predictive value is essential for ensuring measurement validity. For example, the Big Five personality questionnaire is used to predict personality traits, measure susceptibility to affective disorders (Wilks et al., 2020), and predict subjective and psychological well-being (Anglim et al., 2020). Previous studies primarily used reliability and validity tests, with coefficients like Cronbach' s alpha reflecting reliability, but these methods cannot intuitively reflect accuracy and predictive value. BROC analysis can directly quantify accuracy in specific studies through AUC, compensating for traditional reliability/validity limitations. For example, Zeinab et al. used BROC analysis to determine personality traits' predictive value for psychological problems in Iranian adults, obtaining ROC curves for three questionnaires and comparing AUCs to conclude that neuroticism has good predictive value for common psychological problems (Zeinab et al., 2017). Kassing et al. used BROC analysis to predict adult convictions from early childhood behavior problems (Kassing et al., 2019). Lin GM (2020) used BROC analysis to evaluate machine learning models' accuracy in predicting suicide ideation among military personnel.

Furthermore, BROC analysis is not limited to questionnaire research but also applies to experimental studies such as magnetic resonance imaging and brain research. Stevens et al. used BROC to investigate fMRI reliability in preoperative brain tumor mapping (Stevens et al., 2016). Raes et al. used BROC to evaluate transcranial magnetic stimulation (TMS) accuracy and diagnose equine spinal cord dysfunction using Bayesian latent class models (Raes et al., 2020). Gu et al. also used BROC analysis to assess MRI diagnostic performance (Gu, 2019). In mental illness diagnosis research, ICD has traditionally served as the diagnostic standard, but many psychological traits lack gold standards. BROC analysis enables accuracy assessment without gold standards, opening new avenues for psychometric accuracy evaluation.

3.2 Comparing Measurement Tools Under Different Conditions

ROC analysis can obtain AUC values to enable meaningful comparisons across different screening or diagnostic tests (Walker, 2019). The same psychological trait is often measured using different tools—for instance, over five different questionnaires exist for measuring psychological craving, such as the Dependence Scale, Drug Craving Questionnaire, Drug Relapse Risk Scale, and Addiction Craving and Automatic Behavioral Response Scale. Additionally, the same tool may yield different results under different external conditions. Therefore, relying solely on reliability/validity for effectiveness conclusions is one-sided and unscientific, and using a single fixed value for binary conversion can bias results, which ROC analysis effectively avoids. Our review identifies three main types of measurement tool comparisons: across different subjects, across different time points, and across different measurement tools.

In psychological research, comparing the same psychological factor across different samples has significant theoretical and practical importance. Previous comparisons of the same psychological characteristic across different subject samples typically used parametric tests, which require normally distributed data. BROC analysis has no distribution requirements and can directly compare differences across samples using curve plots for more intuitive and clear result presentation. For certain psychological traits, differences may exist across populations, leading to accuracy variations across subject groups—for example, occupational burnout differences across professions can be compared using ROC analysis.

ROC curve methods can independently compare the accuracy of two or more measurement tools. The same psychological phenomenon or factor may be measured using different tools due to varying theoretical foundations and dimensions. Different researchers studying the same psychological issue may use different scales, but few compare accuracy across scales, potentially leading to inconsistent research results that hinder replication and meta-analysis. Researchers therefore need to compare accuracy across measurement tools, which BROC analysis can achieve by independently comparing different tools without gold standards. For example, Chenneville et al. used ROC analysis to compare PHQ and CES-D utility for depression screening in HIV-infected youth (Chenneville et al., 2019). Hartung et al. used ROC analysis to evaluate and compare HADS and PHQ-9 as depression screening tools for cancer patients (Hartung et al., 2017).

While the above focuses on BROC applications in psychology, tROC analysis is also an important and valuable method for longitudinal psychological research. It can not only compare the accuracy of individual measurement tools but also consider covariates such as time. tROC is currently commonly used in survival analysis, particularly for predicting survival time in advanced cancer patients. Although numerous studies have applied it in survival analysis, its use in other longitudinal research remains rare, and its application in psychology is even scarcer, despite its considerable potential. For example, Liu et al. used tROC

analysis to evaluate the dynamic predictive performance of muscle activity over time, obtaining optimal cut-off values from ROC curves to classify tonic and phasic muscle activity into mild and severe categories (Liu et al., 2019). Similarly, craving measurement scales for individuals in drug rehabilitation may show varying effectiveness at different time points during treatment, which is directly relevant to our subsequent research.

In summary, ROC analysis is a research method applicable to psychology, medicine, and numerous other fields. In recent years, its application in psychology has expanded beyond information processing to include comparison and evaluation of psychometric tools, though overall its use in psychology remains in its early stages. Systematically and comprehensively reviewing ROC analysis advances in psychometric accuracy assessment will facilitate broader application of the method.

4. Practical Demonstration

To better illustrate ROC analysis applications in psychometrics, this paper uses OpenBUGS software with artificial data to simulate BROC analysis implementation. BROC analysis first requires selecting an appropriate model, then choosing and setting different parameters based on practical needs, validating the model, and finally obtaining ROC curves, AUC, cut-off values, etc. This simulation examines heroin addiction status in 100 subjects to identify optimal addiction threshold values and quantify study accuracy. One hundred subjects completed the Heroin Dependence Scale, with scores recorded. The analysis process follows.

This simulation assumes 100 subjects: $i = 1, 2, \ldots, 100$. Subject questionnaire scores are denoted as Y_i, and demographic variables such as age as X_i. The true status of subject i is d_i (addicted=1, not addicted=0). Assuming questionnaire scores are continuous variables following normal distributions under both addicted and non-addicted conditions—two distinct normal distributions:

$$Y|d = 0 \sim N(\alpha, \tau)$$
$$Y|d = 1 \sim N(\alpha' = \alpha + \beta, \tau)$$

Thus d_i follows a binomial distribution: $d_i \sim Bern(\pi_i)$, where π_i is the probability of $d_i = 1$ (whether the person is addicted). Incorporating demographic covariates:

$$logit(\pi_i) = \eta + \psi * X_i$$

Under the Bayesian model, we assign appropriate prior distributions to these parameters:

$$\alpha \sim N(0,1)$$
$$\beta \sim N(0,1)$$
$$\eta \sim N(0,1)$$
$$\psi \sim N(0,1)$$
$$\tau \sim gamma(0.001, 0.001)$$

Assuming parameters "eta," "psi," etc., we use Gibbs sampling with repeated iterations for parameter convergence. This simulation iterates three times, with results shown in Figure 2 [Figure 2: see original paper]. The overlapping curves indicate good iteration performance. Additionally, rank correlation analysis of model parameters shows correlation coefficients approaching 0, indicating normal model performance (Figure 3 [Figure 3: see original paper]). Finally, the ROC curve is obtained (Figure 4 [Figure 4: see original paper]) with related information.

This simulation addresses ROC analysis for continuous variables without gold standards, applicable not only to questionnaire data accuracy measurement and classification but also to behavioral experiment results. Specific code for this simulation is available from the corresponding author.

5. Summary and Outlook

Since ROC analysis was applied to diagnostic research, it has continuously evolved with clinical needs and technological advances. While some studies have applied ROC analysis to psychological research, no systematic review of its specific applications in psychological research exists. This paper not only summarizes ROC analysis developments but also reviews its specific applications in psychological research.

We first organized ROC analysis applications under different conditions with brief implementation introductions, then detailed its applications in psychology. As described, ROC analysis itself is relatively mature, with increasing applications in psychometric accuracy assessment in recent years. However, the authors identify several issues requiring resolution for broader application in psychometric evaluation.

First, ROC analysis' s value in psychometric tool assessment requires more empirical research support. ROC analysis' s greatest advantage is obtaining ROC curves for intuitive, independent accuracy comparisons. While ROC analysis has served as an excellent diagnostic evaluation method in medicine, its role in psychological measurement—given differences between psychological and physiological indicators—still requires more empirical validation. Additionally, ROC analysis applications should consider practical contexts, particularly in mental illness diagnosis, where physicians' subjective judgments should be integrated for final decisions.

Second, BROC' s application value warrants further exploration. BROC is the least restrictive ROC method, enabling measurement tool accuracy assessment without gold standards. This provides a solid foundation for psychological applications. For example, in substance addiction research, psychological craving is primarily measured via questionnaires, while relatively objective tools like EEG still require questionnaire results for anchoring. However, different craving measurement questionnaires may yield different results, and tool accuracy may vary across different rehabilitation periods. Moreover, no method currently exists to quantify craving degree and determine "psychological addiction" status. Our subsequent research will address this by further applying ROC analysis to craving diagnosis. Thus, using BROC to evaluate craving measurement tool accuracy holds significant theoretical and practical importance. Furthermore, since psychology primarily uses questionnaires and experiments to measure phenomena and behaviors, BROC can independently calculate and compare scale and experimental result validity without gold standards.

Additionally, ROC analysis can integrate with machine learning and computational psychiatry for interdisciplinary research. With advances in computer science, machine learning and computational psychiatry have become research hotspots, widely applied in image recognition, language processing, data mining, and healthcare (Komura & Ishikawa, 2019; Goecks & Jalili, 2020; Kan, 2017; Crawley & Zhang, 2020), and serving as research tools for advanced psychological processes in psychometrics (Bleidorn & Hopwood, 2018; Shatte & Hutchinson, 2019). In machine learning, evaluating model accuracy and making judgments is essential—a step achievable through ROC analysis, where AUC is an important performance evaluation criterion widely used in imbalanced learning, cost-sensitive learning, ranking learning, and other tasks (Dwyer & Falkai, 2018). Overall, ROC analysis' s application scope remains worthy of promotion, possessing irreplaceable functions. Like a catalyst, ROC analysis can be applied across all fields requiring accuracy measurement, offering simple operation yet precise, rich results that can enhance numerous studies.

References

Chen, W., & Zhang, J. (2012). Evaluation of diagnostic tests with ordinal gold standard and its application in coronary heart disease diagnosis. *Chinese Journal of Health Statistics*, 29(2), 172-174.

Wang, X., Zhou, X., Liu, Q., & Gao, Y. (2019). Bayesian estimation of diagnostic accuracy for two methods without gold standard. *Chinese Journal of Health Statistics*, 36(5), 653-657.

Gong, X., Wang, S., Jiao, A., & Hua, X. (2020). Predictive value of PFM scoring system based on prothrombin time, fibrinogen, and mean platelet volume for survival in advanced pancreatic cancer patients. *Abdominal Surgery*, 33(5), 359-375.

Amini, M., Kazemnejad, A., Zayeri, F., Montazeri, A., Rasekhi, A., Amirian, A.,

& Kariman, N. (2019). Diagnostic accuracy of maternal serum multiple marker screening for early detection of gestational diabetes mellitus in the absence of a gold standard test. *BMC Pregnancy and Childbirth*, 20(1), 375–384.

Anglim, J., Horwood, S., Smillie, L. D., Marrero, R. J., & Wood, J. K. (2020). Predicting psychological and subjective well-being from personality: A meta-analysis. *Psychological Bulletin*, 146(4), 279–323.

Arora, P., Thorlund, K., Brenner, D. R., & Andrews, J. R. (2019). Comparative accuracy of typhoid diagnostic tools: A Bayesian latent-class network analysis. *PLOS Neglected Tropical Diseases*, 13(5), 1–23.

Bowers, A. J., & Zhou, X. (2019). Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk*, 24(1), 20–46.

Battaglia, Y., Zerbinati, L., Piazza, G., Martino, E., Provenzano, M., Esposito, P., Massarenti, S., Andreucci, M., Storari, A., & Grassi, L. (2020). Screening performance of Edmonton Symptom Assessment System in kidney transplant recipients. *Journal of Clinical Medicine*, 9(4), 995.

Blanche, P., Dartigues, J. F., & Jacqmin-Gadda, H. (2013). Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55(5), 687–704.

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, 23(2), 190–203.

Crawley, D., Zhang, L., Jones, E. J. H., Ahmad, J., Oakley, B., San José Cáceres, A., Charman, T., Buitelaar, J. K., Murphy, D. G. M., Chatham, C., den Ouden, H., Loth, E., & EU-AIMS LEAP group. (2020). Modeling flexible behavior in childhood to adulthood shows age-dependent learning mechanisms and less optimal learning in autism in each age group. *PLOS Biology*, 18(10).

Chenneville, T., Gabbidon, K., Drake, H., & Rodriguez, C. (2019). Comparison of the utility of the PHQ and CES-D for depression screening among youth with HIV in an integrated care setting. *Journal of Affective Disorders*, 140–

Cui, L., Dong, X., & Zhang, S. (2021). ERP evidence for emotional sensitivity in social anxiety. *Journal of Affective Disorders*, 279, 361–367.

Díaz-Coto, Martínez-Camblor, P., & Pérez-Fernández, S. (2020). Smooth ROC time: An R package for time-dependent ROC curve estimation. *Computational Statistics*, 35(3), 1231–1251.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, 14(1), 91–118.

Flor, M., Weiß, M., Selhorst, T., Müller-Graf, C., & Greiner, M. (2020). Comparison of Bayesian and frequentist methods for prevalence estimation under misclassification. *BMC Public Health*, 20(1), 1135.

Goecks, J., Jalili, V., Heiser, L. M., & Gray, J. W. (2020). How machine learning will transform biomedicine. *Cell*, 181(1),

Goyal, A., Yolcu, Y. U., Goyal, A., Kerezoudis, P., Brown, D. A., Graffeo, C. S., Goncalves, S., Burns, T. C., & Parney, I. F. (2019). The T2-FLAIR-mismatch sign as an imaging biomarker for IDH and 1p/19q status in diffuse low-grade gliomas: A systematic review with a Bayesian approach to evaluation of diagnostic test performance. *Neurosurgical Focus*, 47(6), 13.

Hartung, T. J., Friedrich, M., Johansen, C., Wittchen, H. U., Faller, H., Koch, U., Brähler, E., Härter, M., Keller, M., Schulz, H., Wegscheider, K., Weis, J., & Mehnert, A. (2017). The Hospital Anxiety and Depression Scale (HADS) and the 9-item Patient Health Questionnaire (PHQ-9) as screening instruments for depression in patients with cancer. *Cancer*, 123(21), 4236–4243.

Heagerty, & Zheng. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1), 92–105.

Janssens, A. C. J. W., & Martens, F. K. (2020). Reflection on modern methods: Revisiting the area under the ROC Curve. *International Journal of Epidemiology*, 49(4), 1397–1403.

Martínez-Camblor, P., & Pardo-Fernández, J. C. (2019). The Youden Index in the generalized receiver operating characteristic curve context. *International Journal of Biostatistics*, 15(1).

Kamarudin, A. N., Cox, T., & Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: Current methods and applications. *BMC Medical Research Methodology*, 17(1), 53.

Kassing, F., Godwin, J., Lochman, J. E., & Coie, J. D. (2019). Conduct Problems Prevention Research Group. Using early childhood behavior problems to predict adult convictions. *Journal of Abnormal Child Psychology*, 47(5), 765–

Komura, D., & Ishikawa, S. (2019). Machine learning approaches for pathologic diagnosis. *Virchows Archiv*, 475(2), 131–

Kan, A. (2017). Machine learning applications in cell image analysis. *Immunology and Cell Biology*, 95(6), 525–530.

Lin, G. M., Nagamine, M., Yang, S. N., Tai, Y. M., Lin, C., & Sato, H. (2020). Machine learning based suicide ideation prediction for military personnel. *IEEE Journal of Biomedical and Health Informatics*, 24(7), 1907–1916.

Lui, P. P., Samuel, D. B., Rollock, D., Leong, F. T. L., & Chang, E. C. (2020). Measurement invariance of the five factor model of personality: Facet-level analyses among Euro and Asian Americans. *Assessment*, 27(5), 887–902.

Mandrekar, J. N. (2010). Simple statistical measures for diagnostic accuracy assessment. *Journal of Thoracic Oncology*, 5(6), 763–764.

Macchi, C., Favero, C., Ceresa, A., Vigna, L., Conti, D. M., Pesatori, A. C., Racagni, G., Corsini, A., Ferri, N., Sirtori, C. R., Buoli, M., Bollati, V., & Ruscica, M. (2020). Depression and cardiovascular risk-association among Beck Depression Inventory, PCSK9 levels and insulin resistance. *Cardiovascular Diabetology*, 19(1), 187.

McClean, G., Riding, N. R., Pieles, G., Watt, V., Adamuz, C., Sharma, S., George, K. P., Oxborough, D., & Wilson, M. G. (2019). Diagnostic accuracy and Bayesian analysis of new international ECG recommendations in paediatric athletes. *Heart*, 105(2), 152–159.

Ma, Y., Ji, J., Huang, Y., Gao, H., Li, Z., Dong, W., Zhou, S., Zhu, Y., Dang, W., Zhou, T., Yu, H., Yu, B., Long, Y., Liu, L., Sachs, G., & Yu, X. (2019). Implementing machine learning in bipolar diagnosis in China. *Translational Psychiatry*, 9(1), 305.

Numan, T., van den Boogaard, M., Kamper, A. M., Rood, P. J. T., Peelen, L. M., & Slooter, A. J. C. (2019). Dutch Delirium Detection Study Group. Delirium detection using relative delta power based on 1-minute single-channel EEG: A multicentre study. *British Journal of Anaesthesia*, 122(1), 60–68.

Obuchowski, N. A. (2005). Estimating and comparing diagnostic tests' accuracy when the gold standard is not binary. *Statistics in Medicine*, 20, 3261–3278.

Obuchowski, N. A., & Bullen, J. A. (2018). Receiver operating characteristic (ROC) curves: Review of methods with applications in diagnostic medicine. *Physics in Medicine and Biology*, 63(7).

Peng, F., & Hall, W. J. (1996). Analysis of ROC curves using Markov-chain Monte Carlo methods. *Medical Decision Making*, 16(4), 404–11.

Paul Nguyen. (2007). nonbinROC: Software for evaluating diagnostic accuracies with non-binary gold standards. *Journal of Statistical Software*, 21(10), 1–10.

Park, M. H., Lee, S. H., Ko, Y. H., Kim, Y. K., Han, K. M., Jeong, H. G., & Han, C. (2019). Usefulness of the 15-item geriatric depression scale (GDS-15) for classifying minor and major depressive disorders among community-dwelling elders. *Journal of Affective Disorders*, 259, 370–375.

QiuWang, M. A. D. A. (2005). Applying Bayesian modeling and receiver operating characteristic methodologies for test utility analysis. *Educational and Psychological Measurement*, 73(2), 275–292.

Ruddy, J., Ciancio, D., Skinner, C. H., & Blonder, M. (2018). Receiver operating characteristic analysis of oral reading fluency predicting broad reading scores. *Contemporary School Psychology*.

Richardson, M., Hussain, Z., & Griffiths, M. D. (2018). Problematic smartphone use, nature connectedness, and anxiety. *Journal of Behavioral Addictions*, 7(1),

109–116.

Raes, E., Buczinski, S., Dumoulin, M., Deprez, P., Van Ham, L., van Loon, G., & Pardon, B. (2020). Accuracy of transcranial magnetic stimulation and a Bayesian latent class model for diagnosis of spinal cord dysfunction in horses. *Journal of Veterinary Internal Medicine*, 34(2), 964–971.

Suzuki, Y., Okabayashi, K., Hasegawa, H., Tsuruta, M., Shigeta, K., Kondo, T., & Kitagawa, Y. (2018). Comparison of preoperative inflammation-based prognostic scores in patients with colorectal cancer. *Annals of Surgery*, 267(3), 527–531.

Stevens, M. T., Clarke, D. B., Stroink, G., Beyea, S. D., & D' Arcy, R. C. (2016). Improving fMRI reliability in presurgical mapping for brain tumours. *Journal of Neurology, Neurosurgery, and Psychiatry*, 87(3), 267–74.

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, 49(9), 1426–1448.

Sumner, C. J., & Sumner, S. (2020). Signal detection: Applying analysis methods from psychology to animal behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375.

Sachs, M. C. (2017). plotROC: A tool for plotting ROC curves. *Journal of Statistical Software*.

Thapa, S., Sun, H., Pokhrel, G., Wang, B., Dahal, S., & Yu, S. (2020). Performance of Distress Thermometer and associated factors of psychological distress among Chinese cancer patients. *Journal of Oncology*, 1-8.

Tang, Z. H., Zeng, F., Yu, X., & Zhou, L. (2014). Bayesian estimation of cardiovascular autonomic neuropathy diagnostic test based on baroreflex sensitivity in the absence of a gold standard. *International Journal of Cardiology*, 171(3),

Wilks, Z., Perkins, A. M., Cooper, A., Pliszka, B., Cleare, A. J., & Young, A. H. (2020). Relationship of a big five personality questionnaire to the symptoms of affective disorders. *Journal of Affective Disorders*, 277, 14–20.

Zeinab Alizadeh A, B. (2017). The predictive value of personality traits for psychological problems (stress, anxiety and depression): Results from a large population-based study. *Journal of Epidemiology and Global Health*, 8, 124-

Author Contribution Statement:

Liu Yuqing: Responsible for conceptualization and manuscript writing
Li Huiling, Zhou Qiang: Revision of final manuscript version

Note: Figure translations are in progress. See original paper for figures.