Advances in Psychology 心理学进展, 2022, 12(7), 2612-2622
Published Online July 2022 in Hans. http://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2022.127311

Hans汉斯

无金标准ROC方法在心理研究中的发展应用

刘雨晴，李慧玲，余城昊，周　强*

温州医科大学，浙江　温州

收稿日期：2022年6月27日；录用日期：2022年7月20日；发布日期：2022年7月27日

摘　要

传统ROC (receiver operating characteristic)分析方法的核心是将所测的二分结果与“金标准”做比较，通过ROC曲线及其指标对测量工具的准确性进行评估。但心理学研究往往缺乏金标准。与传统手段相比，基于贝叶斯理论的无金标准ROC分析方法(BROC分析)无须依赖金标准，从而摆脱心理研究结果缺乏金标准的困局，为心理研究指标准确性评估中的应用提供了新方向。本文介绍BROC分析并概述其在问卷测量的临界值选择与量化测量工具准确性等心理学的应用价值，并通过实例演示模拟其在心理研究中的操作实现，进而讨论以BROC分析为主的ROC分析方法的应用前景及不足。

关键词

ROC分析，诊断测验，准确性评估，贝叶斯，金标准

The Development and Application of ROC Method without Gold Standard in Psychological Research

Yuqing Liu, Huiling Li, Chenghao Yu, Qiang Zhou*

Wenzhou Medical University, Wenzhou Zhejiang

Received: Jun. 27th, 2022; accepted: Jul. 20th, 2022; published: Jul. 27th, 2022

Abstract

The core of the traditional ROC (receiver operating characteristic) analysis method is to compare the measured dichotomous results with the “gold standard”, and to evaluate the accuracy of the measurement tool through the ROC curve and its indicators. But psychological research often

*通讯作者，心理学博士、副教授，Email: zq@wmu.edu.cn

文章引用: 刘雨晴, 李慧玲, 余城昊, 周强(2022). 无金标准 ROC 方法在心理研究中的发展应用. *心理学进展, 12(7),* 2612-2622. DOI: 10.12677/ap.2022.127311

lacks a gold standard. Compared with traditional methods, the non-gold standard ROC analysis method (BROC analysis) based on Bayesian theory does not need to rely on the gold standard, so as to get rid of the dilemma of lack of gold standard in psychological research results, and provides a new direction for the application of the accuracy evaluation of psychological research indicators. This manuscript firstly introduces BROC analysis and outlines its application value in psychology, such as the selection of critical value of questionnaire measurement and the accuracy of quantitative measurement tools, followed by demonstration and simulation of its operation in psychological research through examples, and finally discusses the application prospect and deficiency of ROC analysis method, especially the BROC analysis.

Keywords

ROC Analysis, Diagnostic Test, Accuracy Assessment, Bayesian, Gold Standard

1. 引言

ROC (receiver operating characteristic)分析发端于信号检测论(SDT)，最初用于研究感觉阈限(如听觉、视觉和触觉)等行为反应；而今则被广泛应用于分析心理学和神经科学实验(Sumner & Sumner, 2020)以及其他领域，如医学诊断、机器学习(Obuchowski & Bullen, 2018; Ma et al., 2019)等。ROC 分析一般通过二分类转换，寻找最佳临界点，从而获得更多所需要的信息。如，Levis 和 Sun 等人(2020)利用 ROC 分析方法来比较抑郁症筛查量表 PHQ-2、PHQ-9 及联合诊断之间的评估准确性。再如，Richardson 等(2018)在研究智能手机使用时，利用 ROC 分析来获得智能手机使用量表(PSUS)的最佳阈值，通过计算 AUC 来评估 PSUS 的准确性，并利用 cut-off 点去寻找连续性结果的最佳临界值。

心理学研究常利用生理或心理指标来量化心理状态和/或特质，从而达到预测和控制相关行为之目的，指标的准确性评估是保证预测结果的重要前提。但以往心理研究中利用信效度来反映测量工具的稳定性与有效性，无法直观反映结果的预测价值，且无法直接比较不同测量工具之间的准确性，更无法对连续变量的结果进行二分类转换(Diebig & Angerer, 2021)，因此，如何用更好的方法来评估心理测量工具的准确性迫在眉睫。与此同时，近年来国外应用 ROC 分析方法对心理测量工具进行准确性评估的研究越来越多(Bowers & Zhou, 2019; Thapa et al., 2020)，但它在国内心理研究领域中却尚未引起足够重视。究其原因，是因为传统 ROC 分析需要金标准(即公认的最可靠判断方法)，而心理学研究往往缺乏金标准。

随着计算机技术的发展，在医学诊断研究中，改进的 ROC 分析方法不仅实现了对金标准条件的放宽(从二分金标准、等级金标准到无金标准)，而且还能在研究过程中将更多协变量的影响考虑在内，如时间依赖相关的 ROC 分析、基于贝叶斯原理的无金标准 ROC 分析方法等。尤其后者，能够在无金标准下进行诊断评估，彻底摆脱过去 ROC 分析必须基于金标准存在的壁垒，从而为缺乏金标准的一些研究提供诊断评估的手段，这为 ROC 在心理研究指标准确性评估中的应用提供了可能性。

基于此，本文旨在将无金标准 ROC 分析方法“移植”到心理研究领域，从而拓宽其应用范围。首先介绍ROC分析的基本概念与相关指标的意义，其次阐述ROC分析的种类及其在心理测量中的应用进行，最后以贝叶斯为基础的 ROC 分析(BROC)为实例，模拟 BROC 分析在心理研究中的操作实现，并讨论以 BROC 分析为主的 ROC 分析方法的应用前景及不足。

2. 常见 ROC 分析方法的发展

ROC 曲线是一个以 1-特异性(specificity)为横坐标，敏感性(sensitivity)为纵坐标的曲线关系图(见图 1)。敏感性系通过阳性测试结果准确检测出实际风险(如抑郁)的速率；特异性被定义为检测通过阴性检测结果确定没有风险的比率(Behar et al., 2003; Lehr et al., 2010; Mandrekar, 2010)。研究者主要利用曲线的临界值(cut-off point)和曲线下面积(AUC，area under curves)来反应诊断结果。曲线的临界值(cut-off point)即曲线拐点处的正切值，在临床研究中常选择最大约登指数(Youden index)所对应的临界值，即最佳 cut-off 值，作为将测试结果划分为阳性和阴性的依据。约登指数表示诊断方法准确区分患者与非患者的总能力(灵敏度与特异度之和减去 1)，指数越大说明筛查实验的效果越好，真实性越大(Martínez-Camblor & Pardo-Fernández, 2019)。而曲线下面积(AUC)的形式定义是：$AUC = \int_0^1 y(x)\mathrm{d}x$，即对所有可能的特异性值进行检验的敏感度平均值，AUC 越高的测试被认为是准确性越好。但 AUC 的指标变化敏感性低，单靠 AUC 的比较无法直接得出结论，故仍需要结合参考敏感性和特异性值(Janssens & Martens, 2020)。

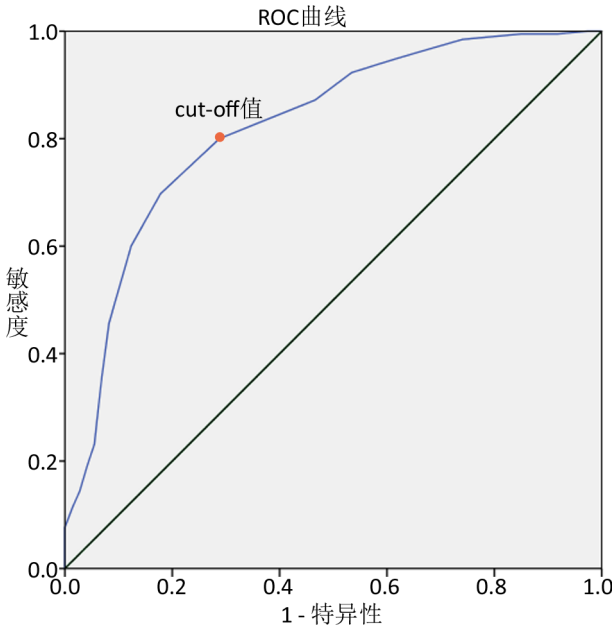


Figure 1. ROC curve
图 1. ROC 曲线图

传统的 ROC 分析在应用于诊断评估时通常使用 Yerushalmy 模式，其核心系将所测结果与“金标准”做比较。“金标准”被视为可正确地区分阴性或阳性症状。因此进行传统 ROC 分析的前提是存在一个可靠、稳定的二分金标准，否则将无法计算其灵敏度与特异度，进而无法评价和判断准确性。尽管金标准对 ROC 分析而言至关重要，但要获得一个稳定、合适的二分金标准并不容易。临床上很多疾病的金标准并非二分变量，而是等级或连续变量；而有些则金标准获取成本极高，过程繁琐复杂，或不符合伦理道德要求；有些甚至暂时没有成熟的金标准。由此可见，对金标准的严格要求极大地限制 ROC 分析方法的应用(王肖南等，2019)。研究者们尝试用不同的方法去突破这一限制，如陈卫中，张菊英(2012)和 Numan 等(2019)提出对等级变量进行两两分组分别进行比较。综上，根据金标准的特征不同，本文总结出以下三种非传统 ROC 方法，简单介绍等级金标准条件下的 ROC 分析方法以及时间依赖相关的 ROC 分析方法，重点推荐无金标准条件下的 ROC 分析方法。

等级金标准条件下的 ROC 分析方法不仅可用来对等级或连续数据的诊断方法进行准确度评价，还可

根据要求将等级变量转化为二分变量。其基本过程是先两两比较各等级状态下的数据，再分别计算曲线下的面积(AUC)，最后比较 AUC 以达到评价的效果(陈卫中，张菊英，2012；Obuchowski, 2005)。例如陈卫中，张菊英(2012)在评价氧化低密度脂蛋白 ELISA 检测试剂盒在冠心病诊断中的诊断价值中，按金标准将被试分为三类状态(阳性、可疑和阴性)。AUC 估计与互相比较可通过 R 软件中的 nonbinROC 包(Nguyen, 2007)实现，更多 R 包详情及操作方法详见该研究。但，这种方法本质上就是将等级变量转换成两两二分变量，并没有改变其本质，且因为其繁琐复杂，所以在临床研究中较少应用。

时间依赖相关的 ROC 分析方法(tROC)主要是通过拓展敏感性与特异性的概念，观察每个时间点的疾病状态，从而获取不同的敏感性和特异性，以此构建一个与时间相关的 ROC 曲线(Heagerty, Lumley, & Pepe, 2000; Kamarudin et al., 2017; Bansal & Heagerty, 2018)。此外还可直接得到不同时间点的 AUC，从而获得关于 AUC(t)的函数图，以便直观有效地比较同一测量指标以及不同测量指标之间在不同观测时间的准确性。此方法最早由 Heagerty 和 Zheng (2005)提出，他们的研究发现可利用每个时间点 t 的累计敏感性与动态特异性(C/D)、事件敏感性与动态特异性(I/D)以及事件敏感性与静态特异性(I/S)等三种不同定义评估上述时间观测事件的敏感性与特异性，从而适用于不同的情境。

在临床医学研究中，tROC 分析中可观察个体疾病的连续状态，增加个体发病时间的信息，还能在时间点间构建 ROC 曲线，并比较各测量指标的预测能力(Chambless & Diao, 2006; Shen, Ning, & Yuan, 2015)。这在临床上有颇为广泛的应用(Suzuki et al., 2018; Liu et al., 2019)，主要是利用患者个体收集的动态信息来预测他们未来健康状况的可能转变。Bansal 和 Heagerty (2019)的研究也发现时间依赖 ROC 方法来评估一个生物标志物的预后潜力比传统方法要好。近年来研究者在此基础上继续完善此方法，如 Schoop et al. (2011)与 Dey et al. (2020)考虑了存在竞争风险的条件下，如何保证时间－事件模型的预测准确性。

综上，上述的 ROC 分析仍无法完全摆脱金标准(GS)。随着贝叶斯学派的发展，一些研究者(Dendukuri & Joseph, 2001; Choi et al., 2006a; Ling et al., 2014)尝试将贝叶斯理论应用于解决金标准限制问题上，实现在无金标准条件下计算不同协变量影响下的 ROC 曲线下面积(AUC)，从而比较诊断准确性。事实上，早在 1996 年 Peng 等便已提出将贝叶斯理论引入 ROC 分析中以求达到该目的，但并未有实质进展(Peng & Hall, 1996)。Choi 等人(2006a)提出在缺乏金标准的情况下，利用 BROC 对相关 ROC 曲线进行比较。相比传统的 ROC 受到数据直接影响，贝叶斯 ROC 仅被数据间接影响，即其数据通过参数从而影响 ROC 曲线，因此相对更准确。与过去的 Yerushalmy 模式不同，该方法主要利用贝叶斯理论，不局限于寻找金标准，而强调收集先验信息，再结合临床经验获得的对疾病有效的相关信息。随着近年来计算机技术的进步，贝叶斯理论被广泛应用于许多领域，尤其是在医学的诊断研究和心理测量工具准确性评估中(Arora et al., 2019; Goyal et al., 2019; Jafarzadeh, Johnson, & Gardner, 2016)。Flor 等(2020)的研究亦表明贝叶斯估计方法显著优于传统的频率估计方法。具体来说，贝叶斯理论认为概率是主观的，并主张将个体经验信息作为重要部分来推导后验分布。其基本原理是先根据模型的样本似然函数，结合参数的先验分布，从而推导出后验分布，即由先验概率乘以似然值而获得后验概率。而在诊断准确性评估研究中，第一步也是最为关键的一步就是需要根据目标人群相关信息，确定先验信息；接着通过似然函数对参数的先验分布进行调整，从而推导出后验分布，实现对相关诊断方法灵敏度和特异度的估计。

因此，对无金标准诊断实验评价而言，只要有一定的实验诊断先验信息，再结合一些并非金标准但临床证实有效的观测数据，就可以通过贝叶斯理论推导出诊断实验评价指标的后验分布，从而摆脱对金标准的依赖。例如 Amini 等(2020)利用贝叶斯潜在分类模型(LCMs)以联系诊断测试观察结果与潜伏疾病状态，在无完全准确疾病状态分类的情况下评估诊断准确性。除此之外，BROC 还能同时考虑多个协变量的影响。相比前面几种 ROC 方法，其在本质上摆脱 ROC 分析方法受金标准的限制，从而拓展了 ROC 分析方法在医学、心理学、计算机等多个领域的应用。如 Tang et al. (2014)的研究，利用贝叶斯模型评估

压力反射敏感性(BRS)进而预测心血管自主神经病变(CAN)。在 CAN 无金标准的前提下，选取 2092 例疑似病例，将年龄、血压等作为协变量，以 BRS 为诊断标准，使用贝叶斯潜在类模型来评估 BRS 的敏感性和特异性。结果发现 BRS 在 CAN 诊断试验中具有较高的敏感性和特异性，具有一定的参考价值，提示 BRS 检验是诊断 CAN 的有效工具。

总的来说，ROC 分析是一种全面的，且准确评估诊断准确性和预测价值的方法，广泛应用于临床医学，也逐渐应用于心理学研究(Thapa et al., 2020; Richardson et al., 2018; Artieda-Urrutia et al., 2015)。其实，早在 2012 年 Qiu Wang 等人就提出将 BROC 分析应用在教育学与心理学当中，结合贝叶斯层次模型和接受者操作特征分析(BROC)来评估兴趣强度(IS)和兴趣分化(ID)如何预测低社会经济地位(SES)青年的兴趣－专业一致性(IMC) (Wang et al., 2012)。近年来，结合实际需求，在传统二分金标准条件下 ROC 分析的基础上，发展出适用于不同临床条件下的适用方法，研究结果也充分证明其合理性。相比传统的信效度检验，ROC 分析不仅仅能够通过 ROC 曲线图直观的反应其准确性，还能对其连续性结果进行二分类转换。

3. BROC 分析方法在心理研究中的应用

如前所述，ROC 分析虽然在心理学领域已经应用颇多年，但主要局限于感知觉阈限及认知加工等领域(Yonelinas & Parks, 2007; Benjamin, 2013; Fleming & Lau, 2014; Wixted, 2020)。在诸多心理疾病的诊断研究中，过去常将国际疾病诊断标准(ICD)作为诊断标准，但诸多心理特质的测量结果是不具备金标准的，而 BROC 分析可以实现在无金标准的条件下对其进行准确性评估。本文关注 BROC 分析在准确性评估和量化临界值上的应用，尤其是在心理学测量上的应用，并就已有的相关研究进行总结梳理。

3.1. 量化特定心理研究工具的预测价值(准确性)

BROC 分析可以用来确定诊断试验的诊断准确性(Fawcett, 2006)，即对测量工具的准确性进行评估，这是其最为重要的功能之一。在心理学研究中，心理测量工具的准确性以及预测价值的评估是研究中最为重要的环节之一。例如利用大五人格问卷来预测人格特征(如 Lui et al., 2020)，测量个体情感障碍的人格特征易感性(如 Wilks et al., 2020)，以及预测主观幸福感和心理幸福感(如 Anglim et al., 2020)。而 BROC 分析方法可以通过曲线下面积(AUC)直接量化其在该研究中的准确性，如比较人格问卷对常见心理问题的预测价值(Alizadeh, 2017)、探究儿童对外化行为的评分是否能准确预测成人信念(Kassing et al., 2019)、判断相关机器学习模型对军事人员自杀意念预测的准确性的好坏评价(Lin et al., 2020)、探究三维心理痛苦(DPPS)量表作为评估高自杀风险抑郁症患者有效筛查量表的准确性(Thapa et al., 2020)等等。

此外，BROC 分析方法还可独立比较两个或多个测量工具的准确性。同一个心理现象或者心理因素由于理论基础和维度不同所使用的测量工具可能存在差异。不同的研究者对同一心理问题的研究可能采用不同的量表，却很少有人将不同的量表之间进行准确性的比较，而测量结果的差异可能是有测量工具本身的差异带来的，且对不同测量工具的准确性进行比较能够很好地帮助我们选择更为适合的测量工具。所以研究者存在比较不同测量工具准确性的需求，通过 ROC 分析可独立比较不同测量工具之间的差异，并评价其准确性。如 Chenneville 等人利用 ROC 分析探讨比较 PHQ 和 CES-D 对艾滋病毒感染者青少年抑郁症筛查的效用(Chenneville et al., 2019)；再如 Hartung 等人(2017)利用 BROC 分析方法来评估医院焦虑抑郁量表(HADS)和 9 项患者健康问卷(PHQ-9)作为筛查癌症患者抑郁的工具的有效性比较)。

当然，BROC 分析方法不仅适用于问卷研究，还适用于实验研究，如磁共振研究等。如 Stevens 等利用 BROC 分析能够预测功能性磁共振成像(fMRI)数据结果的可靠性(Stevens et al., 2016)；而 Higham & Higham (2019)等提出利用 BROC 分析替代元认知领域测量准确性的 Gamma 值。

3.2. 量化问卷测量结果的临界值(Cut-Off)

在心理的临床应用中，常需要根据测量结果对数据进行分类，从而有助于做出是或否、有或无的判断。例如心理学的相关选拔测试中时，需要对连续性结果数据进行分类，从而做出是否符合企业要求的判断；在心理疾病的测量中也是尤为重要，如根据抑郁量表的得分多少，最终将其与特定值比较，从而做出是否患抑郁症的判断。在过去的研究中，我们常使用平均数或者中位数进行二分转换，而在心理疾病诊断中，例如抑郁量表得分中，我们常将其与固定的得分作比较，对其做出分类。但事实上，大多数筛查工具缺少将工作条件划分为不严重和严重的临界值(Diebig & Angerer, 2021)，而 BROC 分析可根据曲线上拐点的正切值来获得阈限值(cut-off)，并参考约登指数找到最佳 cut-off 值从而将连续变量的结果划分为两类(Diebig & Angerer, 2021)。Cut-off 值作为诊断研究中多年来最佳分类指标，将其应用于心理学的二分类转换中是具有十分大潜力的。例如抑郁症，焦虑症，强迫症等评估中可在测试中得出 BROC 曲线，根据 cut-off 值，结合医生的意见即可做出是否有患病的诊断。除临床诊断外，BROC 分析还适用于心理普测。例如 Battaglia 等利用 BROC 分析方法，获得埃德蒙顿症状评估系统(ESAS)情绪困扰等量表对 MINI6.0 定义的精神病患者的检测能力的最佳分界点(Battaglia et al., 2020)。

综上所述，BROC 分析方法是一种适用于心理学，医学等诸多领域的研究方法。近年来 BROC 分析方法在心理学中的应用不仅限于信息加工，还用于心理测量工具的比较与评价，但总体来说其应用在心理学领域方兴未艾。系统全面地梳理 BROC 分析方法在心理测量准确性评估领域的新进展有利于全面推动该方法的应用。

4. 实例演示

为更好地说明 ROC 分析方法在心理学领域的应用，笔者利用 OpenBUGS 软件，采用人工数据，模拟 BROC 分析方法在心理测量中的应用实操。首先需要选择合适的模型(Choi et al., 2006b)，BROC 最常用的是潜在类模型(LCM)，通过统计模型结合多个诊断检验的结果，以在没有单一、准确的参考标准的情况下获得疾病流行率和诊断检验准确性的估计(Yang & Becker, 1997; van Smeden et al., 2013; Collins & Huynh, 2014)，再根据实际的需要选择并设置不同的参数，然后验证模型，最后利用软件获得其 ROC 曲线以及 AUC、cut-off 值等等。本次实验模拟的是对 100 名受试者的海洛因成瘾情况进行分析。

本次模拟假设有 100 个受试者：$i=1,2,\cdots,100$。其中受试者的成瘾问卷分数计为 Y_i，年龄等人口学变量计为 X_i。假设第 i 个人的真实情况为 d_i (成瘾 = 1，不成瘾 = 0)且他在成瘾的情况下和不成瘾的情况下测试得到的问卷分数是连续变量且其得分的分布属于不同正态分布，即：

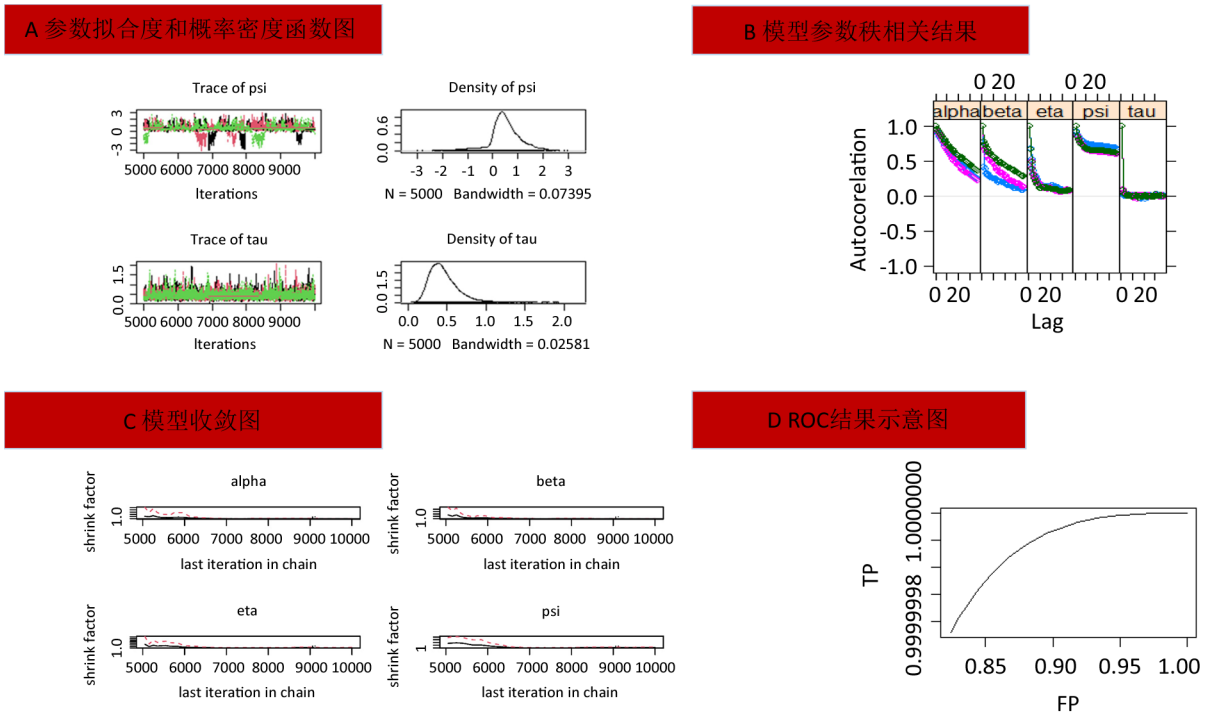
$$Y|d=0\sim N(\alpha,\tau)\quad Y|d=1\sim N(\alpha'=\alpha+\beta,\tau)$$

由上可知 d_i 实际上是二项分布，即 $d_i\sim Bern(\pi_i)$ 是 $d_i=1$ 的概率(这个人是否成瘾)，加入人口学等协变量的影响即：$\log it(\pi_i)=\eta+\psi*X_i$。

在贝叶斯模型下，我们给予这些参数适当的先验分布(prior)：$\alpha\sim N(0,1)\beta\sim N(0,1)\eta\sim N(0,1)$ $\psi\sim N(0,1)$ 正态分布 $\tau\sim gamma(0.001,0.001)$ gamma 分布(因为 tau 是正数)。假设我们选择“eta”、“psi”等参数，使用 Gibbs 抽样的方法通过反复迭代来让参数收敛，此次模拟使用了三条抽样链迭代三次，其结果如图 2(A)所示，其相互重叠，说明参数收敛。此外计算模型各参数的自相关结果发现相关系数趋于 0，说明模型正常，结果如图 2(B)所示，图 2(C)模型收敛因子趋于 1，说明其收敛良好。最后可输出 ROC 曲线图(见图 2(D))及相关信息，此次模拟数据结果 AUC = 0.375，说明该模拟数据的准确性较低(最佳的临界值是 45)。

综上，本次操作成功模拟在无金标准的前提下，对连续变量进行 ROC 分析的解决过程。该方法适用

于问卷数据结果的准确性测量以及结果分类。此次模拟的具体代码可联系本文通讯作者。



图注：图 A、B、C 均反应模型结果情况，D 表示的是 ROC 结果图。图 A 反映的是参数拟合度与密度函数分布情况，如 A 所示，其三条链互相重叠说明参数之间拟合情况良好；图 B 代表的是模型参数秩相关结果，其均趋于 0 代表说明模型正常；图 C 代表的是模型收敛情况，各其最后均趋于 1 代表收敛良好。在图 D 中，横坐标 TP 代表 True positives，反映的是灵敏度(Sensitivity)，代表分类器预测的正类中实际正实例占所有正实例的比例；FP 代表 False positives，反映的是 1-特异度，代表分类器预测的正类中实际负实例占所有负实例的比例。

Figure 2. Related results
图 2. 相关结果图

5. 总结与展望

本文首先整理不同条件下的 ROC 分析方法的应用，并简单介绍其实现方式，继而总结梳理具体方法上的进展以及在心理学中的具体应用，认为基于贝叶斯的无金标准 ROC (BROC)虽在心理测量领域具有较大的发展潜力，但仍存在以下问题需在未来研究中给予关注。

首先，认为 BROC 的应用价值值得进一步深入探讨。BROC 分析方法是各个 ROC 分析方法中限制条件最为宽松的，无需金标准即可评估测量工具准确性。该方法的提出和使用为 ROC 在心理学中的应用打下良好的基础。例如目前在物质成瘾的研究中，相对客观的心理渴求测量方式(如脑电等)仍需要问卷结果予以锚定。但由于测量心理渴求的问卷不同，结果可能会因为测量方式的不同而存在差异，而在不同戒毒时间段不同的测量工具的准确性亦可能存在差异。此外，缺乏一种可以量化心理渴求感程度，并且做出是否有“心瘾”判断的方法。我们的后续研究将会与此相关，进一步将 ROC 分析应用于心理渴求感的诊断研究。由此可见，利用 BROC 分析来评估心理相关指标(如心理渴求)的准确性具有重要的理论和实践意义。此外，在心理学研究中主要通过问卷法和实验法来测量心理现象与行为活动，而 BROC 分析可以在无金标准的条件下独立计算比较量表和实验结果的有效性。

其次，ROC 分析尚可以融合机器学习、计算精神病学等交叉学科进行研究。随着计算机科学的进一步发展，机器学习和计算精神病学逐渐成为研究热点，不仅广泛应用于图像识别、语言处理和数据挖掘、

医疗领域等(Komura & Ishikawa, 2019; Goecks et al., 2020; Kan, 2017; Crawley et al., 2020)，还在心理测量领域成为高级心理过程的研究工具(Bleidorn & Hopwood, 2019; Shatte et al., 2019)。在机器学习的过程中评估模型准确性，并做出判断是必不可少的步骤，而这一步骤可以通过 ROC 分析来实现，其中 AUC 是机器学习中一种重要的性能评价准则，广泛应用于类别不平衡学习、代价敏感学习、排序学习等诸多学习任务(Dwyer et al., 2018)。

当然，BROC 分析在心理测量工具评估中的使用价值需要更多的实际研究支撑。ROC 分析方法最大的优势是可以获得 ROC 曲线图，从而直观的独立比较其准确性差异。过去的 ROC 分析方法作为评估诊断价值的良好手段主要用于医学领域中，尽管近年来国外逐渐有研究出现心理测量评估中的应用研究，但基于心理指标与生理指标的特点不同，BROC 在心理测量中的作用仍然需要更多的实证研究来证明。

总的来说 BROC 分析的应用范围仍然值得推广，本身具备有不可替代的作用。BROC 分析能够应用于各个需要测量准确度的领域，并且由于它本身操作简单，结果却精确丰富，能够为诸多研究增添色彩。

致　谢

感谢徐适可学长在本论文代码部分提供的帮助。

基金项目

本研究受到国家社会科学基金(20BSH047)的支持。

参考文献

陈卫中, 张菊英(2012). 金标准为等级变量时诊断试验的评价及其在冠心病诊断试验中的应用. *中国卫生统计, 29(2),* 172-174.

王肖南, 周晓华, 刘强, 高颖(2019). 无金标准下两种诊断方法准确度的贝叶斯估计. *中国卫生统计, 36(5),* 653-657.

Alizadeh, Z., Feizi, A., Rejali, M. et al. (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,* 124-133. https://doi.org/10.1016/j.jegh.2017.11.003

Amini, M., Kazemnejad, A., Zayeri, F., Montazeri, A., Rasekhi, A., Amirian, A., & Kariman, N. (2020). 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,* Article No. 375. https://doi.org/10.1186/s12884-020-03068-7

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,* 279-323. https://doi.org/10.1037/bul0000226

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,* e0007303. https://doi.org/10.1371/journal.pntd.0007303

Artieda-Urrutia, P., Delgado-Gómez, D., Ruiz-Hernández, D., García-Vega, J. M., Berenguer, N., Oquendo, M. A., & Blasco-Fontecilla, H. (2015). Short Personality and Life Event Scale for Detection of Suicide Attempters. *Revista de Psiquiatría y Salud Mental (English Edition), 8,* 199-206. https://doi.org/10.1016/j.rpsmen.2015.10.001

Bansal, A., & Heagerty, P. J. (2018). A Tutorial on Evaluating the Time-Varying Discrimination Accuracy of Survival Models Used in Dynamic Decision Making. *Medical Decision Making, 38,* 904-916. https://doi.org/10.1177/0272989X18801312

Bansal, A., & Heagerty, P. J. (2019). A Comparison of Landmark Methods and Time-Dependent ROC Methods to Evaluate the Time-Varying Performance of Prognostic Markers for Survival Outcomes. *Diagnostic and Prognostic Research, 3,* Article No. 14. https://doi.org/10.1186/s41512-019-0057-6

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,* Article No. 995. https://doi.org/10.3390/jcm9040995

Behar, E., Alcaine, O., Zuellig, A. R., & Borkovec, T. D. (2003). Screening for Generalized Anxiety Disorder Using the Penn State Worry Questionnaire: A Receiver Operating Characteristic Analysis. *Journal of Behavior Therapy and Experimental Psychiatry, 34,* 25-43. https://doi.org/10.1016/S0005-7916(03)00004-1

Benjamin, A. S. (2013). Where Is the Criterion Noise in Recognition? (Almost) Everyplace You Look: Comment on Kellen, Klauer, and Singmann (2012). *Psychological Review, 120,* 720-726. https://doi.org/10.1037/a0031911

Bleidorn, W, & Hopwood, C. J. (2019). Using Machine Learning to Advance Personality Assessment and Theory. *Personality and Social Psychology Review, 23,* 190-203. https://doi.org/10.1177/1088868318772990

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,* 20-46. https://doi.org/10.1080/10824669.2018.1523734

Chambless, L. E., & Diao, G. (2006). Estimation of Time-Dependent Area under the ROC Curve for Long-Term Risk Prediction. *Statistics in Medicine, 25,* 3474-3486. https://doi.org/10.1002/sim.2299

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, 250,* 140-144. https://doi.org/10.1016/j.jad.2019.03.023

Choi, J. Y., Kim, M. J., Kim, J. H., Kim, S. H., Ko, H. K., Lim, J. S., Oh, Y. T., Chung, J. J., Yoo, H. S., Lee, J. T., & Kim, K. W. (2006). Detection of Hepatic Metastasis: Manganese- and Ferucarbotran-Enhanced MR Imaging. *European Journal of Radiology, 60,* 84-90. https://doi.org/10.1016/j.ejrad.2006.06.016

Choi, Y. K., Johnson, W. O., Collins, M. T., & Gardner, I. A. (2006). Bayesian Inferences for Receiver Operating Characteristic Curves in the Absence of a Gold Standard. *Journal of Agricultural, Biological, and Environmental Statistics, 11,* 210-229. https://doi.org/10.1198/108571106X110883

Collins, J., & Huynh, M. (2014). Estimation of Diagnostic Test Accuracy without Full Verification: A Review of Latent Class Methods. *Statistics in Medicine, 33,* 4141-4169. https://doi.org/10.1002/sim.6218

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,* e3000908. https://doi.org/10.1371/journal.pbio.3000908

Dendukuri, N., & Joseph, L. (2001). Bayesian Approaches to Modeling the Conditional Dependence between Multiple Diagnostic Tests. *Biometrics, 57,* 158-167. https://doi.org/10.1111/j.0006-341X.2001.00158.x

Dey, R., Sebastiani, G., & Saha-Chaudhuri, P. (2020). Inference about Time-Dependent Prognostic Accuracy Measures in the Presence of Competing Risks. *BMC Medical Research Methodology, 20,* Article No. 219. https://doi.org/10.1186/s12874-020-01100-0

Diebig, M., & Angerer, P. (2021). Description and Application of a Method to Quantify Criterion-Related Cut-Off Values for Questionnaire-Based Psychosocial Risk Assessment. *International Archives of Occupational and Environmental Health, 94,* 475-485. https://doi.org/10.1007/s00420-020-01597-4

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology, 14,* 91-118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters, 27,* 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Fleming, S. M., & Lau, H. C. (2014). How to Measure Metacognition. *Frontiers in Human Neuroscience, 8,* Article No. 443. https://doi.org/10.3389/fnhum.2014.00443

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,* Article No. 1135. https://doi.org/10.1186/s12889-020-09177-4

Goecks, J., Jalili, V., Heiser, L. M., & Gray, J. W. (2020). How Machine Learning Will Transform Biomedicine. *Cell, 181,* 92-101. https://doi.org/10.1016/j.cell.2020.03.022

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,* E13. https://doi.org/10.3171/2019.9.FOCUS19660

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,* 4236-4243. https://doi.org/10.1002/cncr.30846

Heagerty, P. J., & Zheng, Y. (2005). Survival Model Predictive Accuracy and ROC Curves. *Biometrics, 61,* 92-105. https://doi.org/10.1111/j.0006-341X.2005.030814.x

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-Dependent ROC Curves for Censored Survival Data and a Diagnos-

tic Marker. *Biometrics, 56,* 337-344. https://doi.org/10.1111/j.0006-341X.2000.00337.x

Higham, P. A., & Higham, D. P. (2019). New Improved Gamma: Enhancing the Accuracy of Goodman-Kruskal's Gamma Using ROC Curves. *Behavior Research Methods, 51,* 108-125. https://doi.org/10.3758/s13428-018-1125-5

Jafarzadeh, S. R., Johnson, W. O., & Gardner, I. A. (2016). Bayesian Modeling and Inference for Diagnostic Accuracy and Probability of Disease Based on Multiple Diagnostic Biomarkers with and without a Perfect Reference Standard. *Statistics in Medicine, 35,* 859-876. https://doi.org/10.1002/sim.6745

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,* 1397-1403. https://doi.org/10.1093/ije/dyz274

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,* Article No. 53. https://doi.org/10.1186/s12874-017-0332-6

Kan, A. (2017). Machine Learning Applications in Cell Image Analysis. *Immunology and Cell Biology, 95,* 525-530. https://doi.org/10.1038/icb.2017.16

Kassing, F., Godwin, J., Lochman, J. E., & Coie, J. D. (2019). Using Early Childhood Behavior Problems to Predict Adult Convictions. *Journal of Abnormal Child Psychology, 47,* 765-778. https://doi.org/10.1007/s10802-018-0478-7

Komura, D., & Ishikawa, S. (2019). Machine Learning Approaches for Pathologic Diagnosis. *Virchows Archiv, 475,* 131-138. https://doi.org/10.1007/s00428-019-02594-w

Lehr, D., Koch, S., & Hillert, A. (2010). Where Is (Im)balance? Necessity and Construction of Evaluated Cut-Off Points for Effort-Reward Imbalance and Overcommitment. *Journal of Occupational and Organizational Psychology, 83,* 251-261. https://doi.org/10.1348/096317909X406772

Levis, B., Sun, Y., He, C., Wu, Y., Krishnan, A., Bhandari, P. M. et al. (2020). Accuracy of the PHQ-2 Alone and in Combination with the PHQ-9 for Screening to Detect Major Depression: Systematic Review and Meta-Analysis. *Journal of the American Medical Association, 323,* 2290-2300. https://doi.org/10.1001/jama.2020.6504

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,* 1907-1916. https://doi.org/10.1109/JBHI.2020.2988393

Ling, D. I., Pai, M., Schiller, I., & Dendukuri, N. (2014). A Bayesian Framework for Estimating the Incremental Value of a Diagnostic Test in the Absence of a Gold Standard. *BMC Medical Research Methodology, 14,* Article No. 67. https://doi.org/10.1186/1471-2288-14-67

Liu, G. M., Zeng, H. D., Zhang, C. Y., & Xu, J. W. (2019). Identification of a Six-Gene Signature Predicting Overall Survival for Hepatocellular Carcinoma. *Cancer Cell International, 19,* 138. https://doi.org/10.1186/s12935-019-0858-2

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,* 887-902. https://doi.org/10.1177/1073191119873978

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,* 305. https://doi.org/10.1038/s41398-019-0638-8

Mandrekar, J. N. (2010). Simple Statistical Measures for Diagnostic Accuracy Assessment. *Journal of Thoracic Oncology, 5,* 763-764. https://doi.org/10.1097/JTO.0b013e3181dab122

Martínez-Camblor, P., & Pardo-Fernández, J. C. (2019). The Youden Index in the Generalized Receiver Operating Characteristic Curve Context. *International Journal of Biostatistics, 15,* Article ID: 20180060. https://doi.org/10.1515/ijb-2018-0060

Nguyen, P. (2007). NonbinROC: Software for Evaluating Diagnostic Accuracies with Non-Binary Gold Standards. *Journal of Statistical Software, 21,* 1-10. https://doi.org/10.18637/jss.v021.i10

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,* 60-68. https://doi.org/10.1016/j.bja.2018.08.021

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,* Article ID: 07TR01. https://doi.org/10.1088/1361-6560/aab4b1

Peng, F., & Hall, W. J. (1996). Analysis of ROC Curves Using Markov-Chain Monte Carlo Methods. *Medical Decision Making, 16,* 404-411. https://doi.org/10.1177/0272989X9601600411

Richardson, M., Hussain, Z., & Griffiths, M. D. (2018). Problematic Smartphone Use, Nature Connectedness, and Anxiety. *Journal of Behavioral Addictions, 7,* 109-116. https://doi.org/10.1556/2006.7.2018.10

Schoop, R., Beyersmann, J., Schumacher, M., & Binder, H. (2011). Quantifying the Predictive Accuracy of Time-to-Event Models in the Presence of Competing Risks. *Biometrical Journal, 53,* 88-112. https://doi.org/10.1002/bimj.201000073

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine Learning in Mental Health: A Scoping Review of Methods and Applications. *Psychology Medicine, 49,* 1426-1448. https://doi.org/10.1017/S0033291719000151

Shen, W., Ning, J., & Yuan, Y. (2015). A Direct Method to Evaluate the Time-Dependent Predictive Accuracy for Biomarkers. *Biometrics, 71,* 439-449. https://doi.org/10.1111/biom.12293

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,* 267-274. https://doi.org/10.1136/jnnp-2015-310307

Sumner, C. J., & Sumner, S. (2020). Signal Detection: Applying Analysis Methods from Psychology to Animal Behaviour. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 375,* Article ID: 20190480. https://doi.org/10.1098/rstb.2019.0480

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,* 527-531. https://doi.org/10.1097/SLA.0000000000002115

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,* 78-80. https://doi.org/10.1016/j.ijcard.2013.11.100

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, 2020,* Article ID: 3293589. https://doi.org/10.1155/2020/3293589

van Smeden, M., Naaktgeboren, C. A., Reitsma, J. B., Moons, K. G., & de Groot, J. A. (2013). Latent Class Models in Diagnostic Studies When There Is No Reference Standard—A Systematic Review. *American Journal of Epidemiology, 179,* 423-431. https://doi.org/10.1093/aje/kwt286

Wang, Q., Diemer, M. A., & Maier, K. (2012). Applying Bayesian Modeling and Receiver Operating Characteristic Methodologies for Test Utility Analysis. *Educational and Psychological Measurement, 73,* 275-292. https://doi.org/10.1177/0013164412455027

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 Affect Disorder, 277,* 14-20. https://doi.org/10.1016/j.jad.2020.07.122

Wixted, J. T. (2020). The Forgotten History of Signal Detection Theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46,* 201-233. https://doi.org/10.1037/xlm0000732

Yang, I., & Becker, M. P. (1997). Latent Variable Modeling of Diagnostic Accuracy. *Biometrics, 53,* 948-958. https://doi.org/10.2307/2533555

Yonelinas, A. P., & Parks, C. M. (2007). Receiver Operating Characteristics (ROCs) in Recognition Memory: A Review. *Psychological Bulletin, 133,* 800-832. https://doi.org/10.1037/0033-2909.133.5.800