Received: 29 May 2020 | Revised: 10 October 2020 | Accepted: 27 January 2021

DOI: 10.1002/mpr.1868

ORIGINAL ARTICLE

WILEY

Using Bayes theorem to estimate positive and negative predictive values for continuously and ordinally scaled diagnostic tests

Felix Fischer

Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany

Correspondence

Felix Fischer, Department of Psychosomatic Medicine, Charité - Universitätsmedizin Berlin, Chariteplatz 1, Berlin 10098, Germany.
Email: felix.fischer@charite.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: Fi1999/6-1

Abstract

Objectives: Positive predictive values (PPVs) and negative predictive values (NPVs) are frequently reported to put estimates of accuracy of a diagnostic test in clinical context and to obtain risk estimates for a given patient taking into account baseline prevalence in the population. In order to calculate PPV and NPV, tests with ordinally or continuously scaled results are commonly dichotomized at the expense of a loss of information.

Methods: Extending the rationale for the calculation of PPV and NPV, Bayesian theorem is used to calculate the probability of disease given the outcome of a continuously or ordinally scaled test. Probabilities of test results conditional on disease status are modeled in a Bayesian framework and subsequently transformed to probabilities of disease status conditional on test result.

Results: Using publicly available data, probability of a clinical depression diagnosis given PROMIS Depression scores was estimated. Comparison with PPV and NPV based on dichotomized scores shows that a more fine-grained interpretation of test scores is possible.

Conclusions: The proposed method bears the chance to facilitate accurate and meaningful interpretation of test results in clinical settings by avoiding unnecessary dichotomization of test scores.

KEYWORDS

depression, methodology, psychometrics, statistics

1 | INTRODUCTION

The diagnostic performance of a test is commonly evaluated by its sensitivity and specificity. These estimates of test accuracy have been reported to be misleading in clinical practice, since a high sensitivity does not necessarily imply that a condition is likely given a positive test result (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007). Therefore, positive predictive values (PPVs) and negative predictive values (NPVs) help to put the results of a diagnostic text in clinical context (Altman & Bland, 1994).

PPV (probability for the presence of a disease given a positive test result) and NPV (probability for the absence of a disease given a negative test result) are usually calculated for a given cutoff using this specific cutoff's sensitivity and specificity as well as the expected prevalence of the disease in the population of interest. However, many tests used in clinical practice give ordinal or continuous results and are subsequently dichotomized. This is true for laboratory parameters such as the HBa1c value as screening tool for diabetes (Burlingame, Bartholomew, Brink-Wong, Sampaga, & Dye, 2015; Higgins et al., 2011), clinical risk scores in cancer (Li, Khalighi, Wu, & Garcia, 2018) as well as for patient questionnaires such as the Patient Health Questionnaire-9 (PHQ-9), a depression screening tool (Levis, Benedetti, & Thombs, 2019).

Dichotomization into negative and positive test result obviously leads to a loss of information, as the risk of having the condition in question is likely to be substantially different for a patient with a result close to the cutoff compared to a patient with a result much lower or higher than the cutoff. Calculation of PPV and NPV for each potential test result is therefore desirable to obtain more fine-grained estimates of the risk of disease for a patient in a clinical context.

The aim of this paper is to outline a Bayesian method to calculate the PPV and NPV for each potential outcome of continuously and ordinally scaled tests. Following the logic of the calculation of PPV and NPV for dichotomous test, the general idea is to model the probabilities of each potential test result stratified by disease status and use these conditional probabilities in turn to calculate PPV and NPV using Bayes' theorem. After the derivation of the approach, an example is provided and strengths and limitations of the approached are discussed.

2 | METHODS

In the following, we denote positive and negative test results with *T*+ and T-, and true and negative disease status with *D*+ and *D*-, respectively. *P*(*T*+) is then the probability of a positive test result and *P*(*T*+|*D*+) is the probability of a positive test result given the true disease status is positive (sensitivity).

Sensitivity (*P*(*T*+|*D*+)), specificity (*P*(*T*−|*D*−)), PPV (*P*(*D*+|*T*+)), and NPV (*P*(*D*−|*T*−)) are conditional probabilities linked through Bayes' theorem:

$$\begin{aligned}\text{PPV} = \ & P(D+|T+) = P(T+|D+) * \frac{P(D+)}{P(T+)} \\ & = \frac{P(T+|D+) * P(D+)}{P(T+|D+) * P(D+) + P(T+|D-) * (1 - P(D+))}\ .\end{aligned}$$

$$\begin{aligned}\text{NPV} = \ & P\left(D-|T-\right) = \frac{P(T-|D-) * (1 - P(D+))}{P(T-)} \\ & = \frac{P(T-|D-) * (1 - P(D+))}{P(T-|D-) * (1 - P(D+)) + (1 - P(T+|D+)) * P(D+)}.\end{aligned}$$

Sensitivity and specificity for a given cutoff are usually estimated as the proportion of test positives in the diseased and test negatives in the undiseased, assuming two independent binomial distributions.

2.1 | Continuous test results

In order to extend PPV and NPV to ordinal and continuous test results, the challenge is to model *P*(Test result|*D*+), respectively *P*(Test result|*D*−). These are essentially the distributions of test results stratified by disease status. Continuous tests results can be modeled using a suitable continuous distribution such as the normal:

$$(\text{Test score} = i|D+) \sim N(\mu_{\text{diseased}}, \text{ sigma}_{\text{diseased}}),$$

$$(\text{Test score} = i|D-) \sim N\left(\mu_{\text{healthy}}, \text{ sigma}_{\text{healthy}}\right).$$

The conditional probabilities can be then calculated by integrating over the respective probability density function and—using Bayes' theorem—in turn used to derive the probability of the disease being present given a specific test score:

$$\begin{aligned}& P(D+|\text{Test score} = i) \\ & = \frac{P(\text{Test score} = \ i|D+) * P(D+)}{P(\text{Test score} = i|D+) * P(D+) + \ P(\text{Test score} = i|D-) * (1 - P(D+))},\end{aligned}$$

$$\begin{aligned}& P(D-|\text{Test score} = i) \\ & = \frac{P(\text{Test score} = \ i|D-) * (1 - P(D+))}{P(\text{Test score} = i|D-) * (1 - P(D+)) + \ P(\text{Test score} = i|D+) * P(D+)}.\end{aligned}$$

2.2 | Ordinal test results

The same reasoning applies to ordinal test outcomes. Here, distributions of test results must follow a bounded, discrete distribution. Questionnaires and risk scores used in screening are commonly scored by adding up all observed item responses. An IRT model such as the graded response model (GRM; Samejima, 1969) can be used to model such data (Embretson & Reise, 2000). The basic assumption of any IRT model is that the observed response to each item probabilistically depends on a person parameter on an underlying latent trait (Θ) and item parameters. For the unidimensional GRM, each item has a single slope parameter (a) and *J* = *K* - 1 threshold parameters (b), where *K* is the number of response options of the item. The probability of observing response category 1 or higher in an item is $P(x \geq 1|\theta) = 1$. The probability of observing a response of *k* or higher is given by:

$$P_i(x \geq k|\theta, a_i, b_i) = \frac{1}{1 + e^{\left(-a_i * \theta + \ b_{ij}\right)}} \ \mathit{for} \ 2 \leq k \leq K \text{ and } j = k - 1,$$

and the probability of a specific response *k* therefore is:

$$P_i(x = k|\theta, a_i, b_i) = P_i(x \geq k|\theta, a_i, b_i) - P_i(x \geq k + 1|\theta, a_i, b_i).$$

Item responses are considered independent and multiplying the probabilities of all observed item response gives the probability of the observed response pattern across the continuum of the latent trait.

In order to stratify by disease status, one can estimate a multi-group IRT model, where it is assumed that the latent trait follows two distinct distributions:

$$\Theta_{\text{healthy}} \sim N(0, 1).$$

$$\Theta_{\text{diseased}} \sim N(\mu_{\text{diseased}}, \sigma_{\text{diseased}}).$$

The mean and standard deviation of the latent trait Θ are fixed to 0, respectively 1, in the healthy group. Item parameters a and *b* are considered invariant across groups for identification of the model.

This model then gives *P*(Response Pattern|*D*+) and *P*(Response Pattern|*D*−). A recursive algorithm can be used to average over all response patterns resulting in the same sum score to obtain *P*(Test score = *i*|*D*+) and *P*(Test score = *i*|*D*−) (Thissen, Pommerich, Billeaud, & Williams, 1995). Bayes' theorem then allows to calculate *P*(*D*+|Test score = *i*) and *P*(*D*−|Test score = *i*) for any given test score.

3 | RESULTS

The following example outlines how the proposed model can be used in practice. It uses publicly available data collected within the Patient-Reported Outcome Measurement Information System (PROMIS) Wave 2 (Pilkonis, 2016; Pilkonis et al., 2014). For this analysis, only the data collected at time point 3 (3 months after admission) is considered. Overall, 187 patients completed the PROMIS Emotional Distress Depression computer-adaptive test and clinical diagnosis of major depressive disorder served as reference standard (134 negative, 53 cases).

PROMIS employs a GRM for scoring depression severity (Cella, Gershon, Lai, & Choi, 2007) and test results are therefore individual estimates of the continuous latent trait. This scale has been initially calibrated that 50 is approximately the mean depression score in the general population, with a standard deviation of 10.

For each respondent, Θ was sampled from the respective individual posterior distribution of the latent trait, approximated by a normal distribution. On the sample level, Θ was concurrently modeled normally distributed stratified by disease state. The parameters (mean μ, standard deviation σ) of both distributions along with their 95% credible intervals are given in Figure 1. Posterior predictive checks (Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019) indicated appropriate model fit as the data randomly generated by this model (thin lines in Figure 1) resembles the observed data (thick lines) reasonably well.

The probabilities of each T-score from 40 to 90 given diagnosis were estimated by integrating over the respective slice of the probability density functions. Assuming population prevalence of 5%, 15%, and 25%, Figure 2 shows *P*(*D*+|Test score = *i*) and *P*(*D*−|Test score = *i*) along with the respective 95% credible intervals. For comparison, NPV and PPV were also calculated by dichotomizing the test result using the optimal (maximizing combined sensitivity and specificity) cutoff (PROMIS T-score = 58.9) in this sample. 95% CIs were obtained using the bootstrap with 1000 iterations.

It is apparent that for many potential test results, PPV and NPV differ substantially from the observed *P*(*D*+|Test score = *i*) and *P*(*D*−|Test score = *i*), as PPV and NPV average over a wide spectrum of potential test results. For example, given an expected prevalence of 15%, the probability having the disease of a person with a positive test (score > 58.86) is 46% (95% CI: 36%–58%). This compares to 21% (16%–26%) if the actual score is 60 and 80% (63%–93%) if the actual score is 70.

All analysis were conducted using Stan (Carpenter, 2017) and R (R Development Core Team 3.0.1., 2013). MCMC sampling was done in three chains using 2000 iterations. Weakly informative priors were imposed on μ and σ to constrain the parameter space to meaningful values. Examination of traceplots, Rhat (all parameters < 1.005) and effective sample size (1095–7261) indicated appropriate exploration of the posterior distribution. Stan and R code is provided as supplementary material.

4 | DISCUSSION

This paper presents a method to obtain a probability of disease given the results of a continuous, ordinal, or categorical diagnostic test and the expected baseline prevalence of the disease in the population. Similar to PPV and NPV, the prevalence is marginalized—therefore, the probability of disease given any population prevalence can be modeled within a single study. Unlike PPV and NPV, *P*(*D*+|Test score = *i*) and *P*(*D*−|Test score = *i*) take into account all available information provided by a diagnostic test and therefore bear the chance to facilitate accurate and meaningful interpretation of test results in clinical settings.

The choice of the underlying distribution appears crucial for the application of the proposed model. Scores of questionnaires and risk scores can be readily modeled as continuous variables, if an appropriate measurement model is available. PROMIS employs such models for scoring of PROMIS instruments. Also, frequently used measures of depression (e.g., PHQ-9, CES-D [Choi, Schalet, Cook, & Cella, 2014]), anxiety (PANAS, GAD-7 [[Schalet, Cook, Choi, & Cella, 2014]), and other patient-reported outcomes have been calibrated on the PROMIS scales and can be scored on a continuous scale. If such a measurement model is not available, one can estimate such a model for the sole purpose to derive the probability of a given response.

Modeling complex distributions, for example coming from an IRT model as suggested, involves a substantially higher number of parameters and therefore larger sample sizes will be needed compared to estimation of PPV and NPV based on dichotomized test results. Also, it must be ensured that common IRT model assumptions regarding dimensionality of the test and local independence of the items are fulfilled (Embretson & Reise, 2000). Estimation of such a model in small and selected samples or with few items might be problematic in particular. A two-step approach of developing a valid

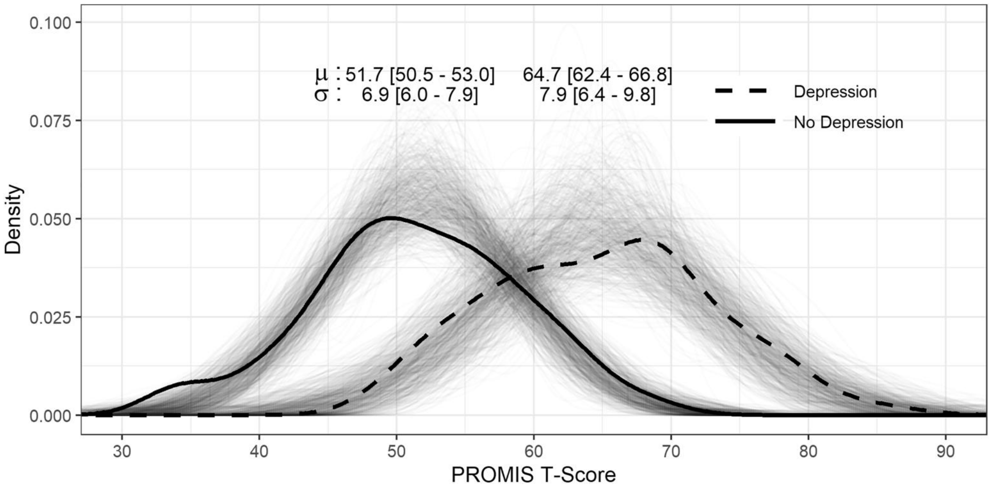

FIGURE 1 Distribution of PROMIS Depression T-scores by disease status. Thin lines depict density of randomly drawn data from the imposed distributions. Observed data (thick lines) resemble this random data well, indicating that the model is appropriate. PROMIS, Patient-Reported Outcome Measurement Information System

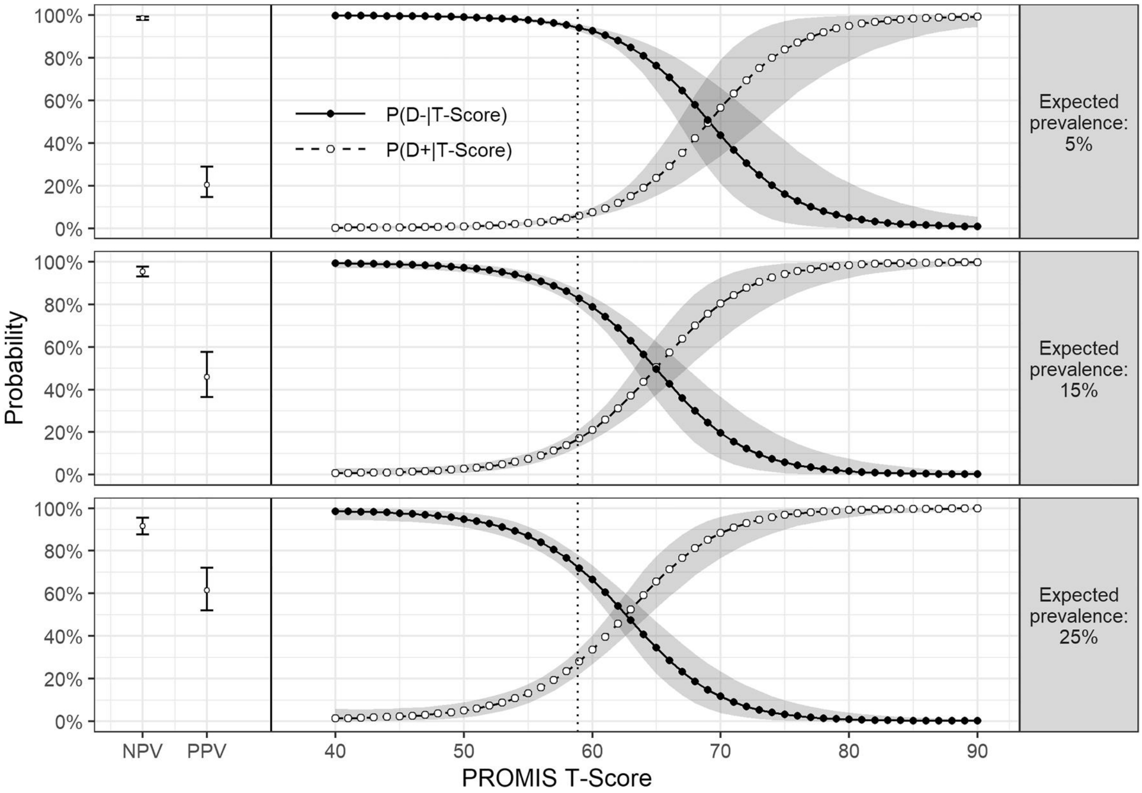

FIGURE 2 NPV, PPV, P(D-|PROMIS T-score) and P(D+|PROMIS T-score) for expected prevalence of 5%, 15%, and 25%. NPV and PPV are calculated on the basis of the optimal cutoff (indicated by pointed line) maximizing combined sensitivity and specificity. Error bars indicate 95% CIs and shaded areas 95% credible intervals. NPV, negative predictive values; PPV, positive predictive values; PROMIS, Patient-Reported Outcome Measurement Information System

measurement model and using this subsequently to obtain continuously distributed test results seems advisable.

Since the probability distributions are estimated independently in each disease status, they are not affected by any change in the marginal distribution of true disease status. Nonetheless, generalizability of the derived PPVs and NPVs could be threatened, if the underlying data does not reflect the full range of disease. For example, the distribution of test scores in the healthy participants can be expected to different in general population samples and clinical populations due to selection bias. This should be considered in applications.

In many cases, the true disease status is not exactly known, but diagnostic accuracy of screening tests is assessed in comparison to an imperfect reference test (Buck & Gart, 1966). In particular for major depressive disorder these reference tests (semi- and fully structured clinical interviews) are known to be imperfect (Levis et al., 2018). The concurrent estimation of test characteristics and prevalence in absence of a true gold standard has been addressed by Bayesian latent class models (Joseph, Gyorkos, & Coupal, 1995). In principle, the approach described in this paper could be further extended to imperfect gold standards by modeling true disease status as a latent variable using information on the reference tests sensitivity and specificity. It is, however, unclear, how such information could be obtained.

Risk of disease for a given test score can be modeled using logistic regression as well (Coughlin et al., 1992) and one can scale such a model to any expected population prevalence by simply adjusting the model intercept (Greenland, 1981). Compared to this approach, modeling the disease-specific test score distributions and subsequently calculating the conditional probabilities in a Bayesian framework as proposed has some advantages that are worth noting:

1. One can combine meaningful subsets of test results by simply integrating over the corresponding areas of the probability density (or mass) function to obtain the probability of disease given these subsets. For example, it is straightforward to calculate $P(D+|50 \leq \text{Test score} < 60)$.
2. Expected prevalence for a given population can be more realistically represented by a beta distribution than a single value of prevalence. While in logistic regression one needs to specify a point estimate of the population prevalence, in the Bayesian framework one could use a Beta(6, 34) to represent the assumption that the population prevalence is expected to be most likely 15% with 95% CI that it is smaller than 25%. Therefore, the approach allows to account for uncertainty about the population prevalence in a clinical setting.
3. In some situations, the data necessary to perform logistic regression might not be readily available. The Bayesian approach allows to flexibly combine data from different studies and sources, for example if the distribution parameters for controls and cases were estimated in different samples.

Readers should be aware of some potential shortcomings of the approach. The choice of distribution to model test results is crucial and although in the example provided a normal distribution seems to work well, this is probably not the case in every application. Furthermore, it is unclear to what extend misfit of the distribution could be tolerated. In particular, estimation of the distribution of the test score in the diseased group will be challenging in real world scenarios, as prevalence is typically low and therefore only few cases are observed. A limitation of the worked example is that it does not address validation of the model given the modest sample size.

Taken together, the Bayesian approach outlined in this paper allows a fine-grained assessment of risk of disease given results of continuously and ordinally scaled diagnostic tests by avoiding loss of information due to dichotomization at the optimal cutoff. It has some potential advantages compared to use of logistic regression to estimate risk of disease and therefore eventually bears the potential to improve interpretation and understanding of test results in clinical applications. A comparative assessment of this proposed method is needed in order to investigate, whether it can actually deliver such an improvement. Differences to traditional logistic regression models should be assessed and one should investigate whether a more fine-grained interpretation of screening test data can be successfully implemented in clinical practice.

ACKNOWLEDGMENTS

This work was supported by Deutsche Forschungsgemeinschaft (Fi 1999/6-1). The author likes to thank the DEPRESSD Team (https://www.depressd.ca/) for valuable discussions on the method and the team of the PROMIS Wave 2 study for making the data available.

CONFLICT OF INTEREST

The author declares that there are no conflict of interest.

AUTHOR CONTRIBUTIONS

Felix Fischer conceived the study, conducted the analysis and wrote the manuscript.

DATA AVAILABILITY STATEMENT

Data are openly available in a public repository that issues datasets with DOIs: The data used in this study are openly available in HealthMeasures Dataverse under the title "PROMIS 1 Wave 2 Depression" at https://doi.org/10.7910/DVN/ZDIITC.

ORCID

Felix Fischer https://orcid.org/0000-0002-9693-6676

REFERENCES

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 2: Predictive values. *BMJ*, *309*, 102 https://doi.org/10.1016/j.ebobgyn.2005.10.001

Buck, A. A., & Gart, J. J. (1966). Comparison of a screening test and a reference test in epidemiologic studies: I. Indices of agreement and their relation to prevalence. *American Journal of Epidemiology*, *83*(3), 586–592 https://doi.org/10.1093/oxfordjournals.aje.a120609

Burlingame, J. M., Bartholomew, L., Brink-Wong, T., Sampaga, S., & Dye, T. (2015). Can we really diagnose diabetes during pregnancy?. *Journal of Perinatal Medicine*, *43*(3), 277–282. https://doi.org/10.1515/jpm-2014-0162

Carpenter, B. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Cella, D., Gershon, R., Lai, J.-S., & Choi, S. W. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, *16*, 133–141. https://doi.org/10.1007/s11136-007-9204-6

Choi, S. W., Schalet, B. D., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. *Psychological Assessment*, *26*(2), 513–527. https://doi.org/10.1037/a0035768

Coughlin, S. S., Trock, B., Criqui, M. H., Pickle, L. W., Browner, D., & Tefft, M. C. (1992). The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. *Journal of Clinical Epidemiology*, *45*(1), 1–7. https://doi.org/10.1016/0895-4356(92)90180-U

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society - Series A: Statistics in Society*, *182*(2), 389–402. https://doi.org/10.1111/rssa.12378

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53–96 https://doi.org/10.1109/77.621897

Greenland, S. (1981). Multivariate estimation of exposure-specific incidence from case-control studies. *Journal of Chronic Diseases*, *34*(9–10), 445–453. https://doi.org/10.1016/0021-9681(81)90004-7

Higgins, T. N., Tran, D., Cembrowski, G. S., Shalapay, C., Steele, P., & Wiley, C. (2011). Is HbA 1c a good screening test for diabetes mellitus?. *Clinical Biochemistry*, *44*(17–18), 1469–1472. https://doi.org/10.1016/j.clinbiochem.2011.08.1138

Joseph, L., Gyorkos, T. W., & Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, *141*(3), 263–272. https://doi.org/10.1093/oxfordjournals.aje.a117428

Levis, B., Benedetti, A., Riehm, K. E., Saadat, N., Levis, A. W., Azar, M., …, Van Weert, H. C. (2018). Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews. *The British Journal of Psychiatry*, *212*, 377–385. https://doi.org/10.1192/bjp.2018.54

Levis, B., Benedetti, A., & Thombs, B. D. (2019). Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis. *BMJ*, *365*, 1476. https://doi.org/10.1136/bmj.l1476

Li, A., Khalighi, P., Wu, Q., & Garcia, D. (2018). External validation of the PLASMIC score: A clinical prediction tool for thrombotic thrombocytopenic purpura diagnosis and treatment. *Journal of Thrombosis and Haemostasis*, *16*(1), 164–169. https://doi.org/10.1111/jth.13882

Pilkonis, P. (2016). *PROMIS 1 Wave 2 depression*. https://doi.org/10.7910/DVN/ZDIITC

Pilkonis, P., Yu, L., Dodds, N. E., Johnston, K. L., Maihoefer, C. C., & Lawrence, S. M. (2014). Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS??) in a three-month observational study. *Journal of Psychiatric Research*, *56*(1), 112–119. https://doi.org/10.1016/j.jpsychires.2014.05.010

R Development Core Team 3.0.1. (2013). *The R Project for statistical computing*. http://www.r-project.org

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*, 1–97. https://doi.org/10.1007/BF02290599

Schalet, B. D., Cook, K. F., Choi, S. W., & Cella, D. (2014). Establishing a common metric for self-reported anxiety: Linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. *Journal of Anxiety Disorders*, *28*(1), 88–96. https://doi.org/10.1016/j.janxdis.2013.11.006

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*(1), 39–49. https://doi.org/10.1177/014662169501900105

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Fischer F. Using Bayes theorem to estimate positive and negative predictive values for continuously and ordinally scaled diagnostic tests. *Int J Methods Psychiatr Res*. 2021;30:e1868. https://doi.org/10.1002/mpr.1868