JAMIA Open, 00(0), 2021, 1–4
doi: 10.1093/jamiaopen/ooab017
Brief Communications

Brief Communications

Improving visual communication of discriminative accuracy for predictive models: the probability threshold plot

Stephen S. Johnston,[1] Stephen Fortin,[2] Iftekhar Kalsekar,[1] Jenna Reps,[2] and Paul Coplan[1]

[1]Epidemiology, Medical Devices, Johnson & Johnson, New Brunswick, New Jersey, USA, [2]Epidemiology; Janssen Research and Development, Titusville, New Jersey, USA

Corresponding Author: Stephen S. Johnston, PhD, Sr. Director, Real-World Data Analytics and Research, Medical Device Epidemiology and Real-World Data Science, Johnson & Johnson, 410 George Street, New Brunswick, NJ, USA; sjohn147@its.jnj.com

Received 8 December 2020; Revised 16 February 2021; Editorial Decision 19 February 2021; Received 8 December 2020;

ABSTRACT

Objectives: To propose a visual display—the probability threshold plot (PTP)—that transparently communicates a predictive models' measures of discriminative accuracy along the range of model-based predicted probabilities (*Pt*).

Materials and Methods: We illustrate the PTP by replicating a previously-published and validated machine learning-based model to predict antihyperglycemic medication cessation within 1–2 years following metabolic surgery. The visual characteristics of the PTPs for each model were compared to receiver operating characteristic (ROC) curves.

Results: A total of 18 887 patients were included for analysis. Whereas during testing each predictive model had nearly identical ROC curves and corresponding area under the curve values (0.672 and 0.673), the visual characteristics of the PTPs revealed substantive between-model differences in sensitivity, specificity, PPV, and NPV across the range of *Pt*.

Discussion and Conclusions: The PTP provides improved visual display of a predictive model's discriminative accuracy, which can enhance the practical application of predictive models for medical decision making.

Key words: predictive analytics, discriminative accuracy, receiver operating characteristic curve

BACKGROUND AND SIGNIFICANCE

Predictive models using machine learning hold promise as a tool to support precision medicine. Such models may be used to estimate an individual patient's predicted probability (Pr_i) of benefit and/or harm from a medical intervention, or of experiencing a future outcome.

A probability threshold (Pt) is required when applying a prediction model to inform a clinical decision such as whether to give an intervention. Patients with a predicted risk greater than the probability threshold, $Pr_i > Pt$, may be selected (or disqualified) for an intervention. For example, consider the situation where a model is developed to predict 5-year future risk of stroke. The model assigns each patient a personalized 5-year risk of stroke. To determine which patients should be prescribed statins, a clinician may apply the prediction model and then give patients with a predicted risk >10% a statin. In this case, the Pt is 10%.

LAY SUMMARY

The extent to which a predictive model can accurately predict an outcome is often communicated via visual displays, most commonly by receiver operating characteristic (ROC) curves. However, ROC curves are limited sensitivity and 1-specificity and omit measures such as positive predictive value (PPV) and negative predictive value (NPV). Furthermore, the ROC curve does not communicate the trade-offs among operating characteristics across the range of predicted probability thresholds, and the area under the curve (AUC) can be misrepresentative in the presence of rare outcomes, in which case PPV may be more important. We suggest a standard visual display that transparently communicates measures of discriminative accuracy, such as sensitivity, specificity, PPV and NPV, along with the range of predicted probability thresholds from a predictive model: the probability threshold plot (PTP). In the present study, two separate predictive models generated nearly identical ROC curves and corresponding AUCs; however, the visual characteristics of the PTPs revealed substantive between-model differences in operating characteristics across the range of predicted probability thresholds; the PTP provides improved visual display of a predictive model's discriminative accuracy, which can facilitate the selection of the Pt and thereby translate predictive models into more useful decision tools enhance the practical application of predictive models for medical decision-making.

The value of the *Pt* used for decision making impacts the operating characteristics when implementing the prediction model and the choice is subjective. Published prediction models generally only provide information about the operating characteristics for a limited selection of *Pt* values, but this knowledge is required to ensure the most suitable *Pt* is used when making a decision. Some previous researchers have adopted the strategy of reporting the *Pts* associated with a selected set of target performance characteristics in a tabular format; for example, Carrell et al[1] reported the *Pts* needed to achieve excellent (0.90), good (0.80), or acceptable (0.75) values of characteristics such as PPV, along with the values of various other operating characteristics associated with that specific *Pt*. Others have recommended to report operating characteristics for multiple selected *Pts*.[2]

Furthermore, a challenge in selecting *Pt* is that a tradeoff exists between false negative and false positive classification when increasing or decreasing the *Pt* for a clinical decision rule. Although the expected prevalence of an outcome, or perhaps a predicted probability >50%, may be selected as the default *Pt* for a clinical decision rule, there may be a desire to select an alternative *Pt* at which certain operating characteristics receive greater emphasis (e.g., PPV may be emphasized for interventions with risk of iatrogenic effects).

The ROC curve itself is perhaps the most ubiquitous visual display of discriminative accuracy for predictive modeling.[3] ROC curves are an elegant way to depict the tradeoff between sensitivity (true positive rate) and false-positive rates (1–specificity) across the distribution of *Pt*s for a given predictive model. The area under the ROC curve (AUC) is a summary of discriminative ability across all *Pt* values and provides no information on what *Pt* value to use when implementing a model.

Nevertheless, ROC curves are limited in communicating information to inform the selection of *Pt*; the value of the Pt corresponding to a point on the ROC curve is unclear, ROC curves are limited to sensitivity and the false positive rate but miss other important operating characteristics such as PPV; the ROC curve fails to visually display the dependence of PPV and NPV on the expected prevalence of the outcome in a population; the ROC curve does not communicate the simultaneous trade-offs among sensitivity, specificity, PPV and NPV for the range of *Pts*; and the specificity value is dominated by the true negatives when the outcome is rare, in which case PPV may be an important measure to consider.[4,5]

Whereas the ROC curve is a useful tool for the evaluation and comparison of predictive models, there is currently no standard visual display that directly conveys information to support the selection of *Pt* for application of a *final* predictive model for medical decision making. In this Brief Communication, we suggest a standard visual display that transparently and comprehensively communicates measures of discriminative accuracy, such as sensitivity, specificity, PPV and NPV, along with the complete range of *Pt* to visually assist in the selection of a *Pt* for a clinical decision rule: the PTP.

MATERIALS AND METHODS

To illustrate the PTP, we replicated a previously published and validated machine learning-based modeling approach to predict the outcome of antihyperglycemic medication cessation within 1–2 years following metabolic surgery, using the IBM MarketScan® Commercial insurance claims database.[6] The predictive modeling approach is described in Johnston et al, which identified insulin use within 6 months prior to metabolic surgery as the strongest predictor of antihyperglycemic medication cessation.[6] We created two predictive models, each one applied to a different population: one for patients with prior insulin use (model A) and one for patients without prior insulin use (model B). The R Code used to generate the PTP is provided in the Appendix.

At each *Pt* value (*x*-axis, range 0%–100%), the PTP plots a predictive model's measures of discriminative accuracy (*y*-axis, range 0%–100%). We show the sensitivity, specificity, PPV, and NPV; however, other measures such as the F1 score, accuracy, or queue rate may also be conveyed.

RESULTS

A 75%/25% training/test set split of the original sample was used to train and internally validate the models. For model A (patients without insulin use prior to metabolic surgery), 9 972 patients were used for training and 3 323 were used for testing; the N (%) of patients experiencing antihyperglycemic medication cessation within 1-2 years after metabolic surgery was 8 161 (81.8%) for training and 2 720 (81.9%) for testing. For model B (patients with insulin use prior to metabolic surgery), 4 194 patients were used for training and 1 398 were used for testing; the N (%) of patients experiencing antihyperglycemic medication cessation within 1–2 years after metabolic surgery was 1 951 (46.5%) for training and 650 (46.5%) for testing.

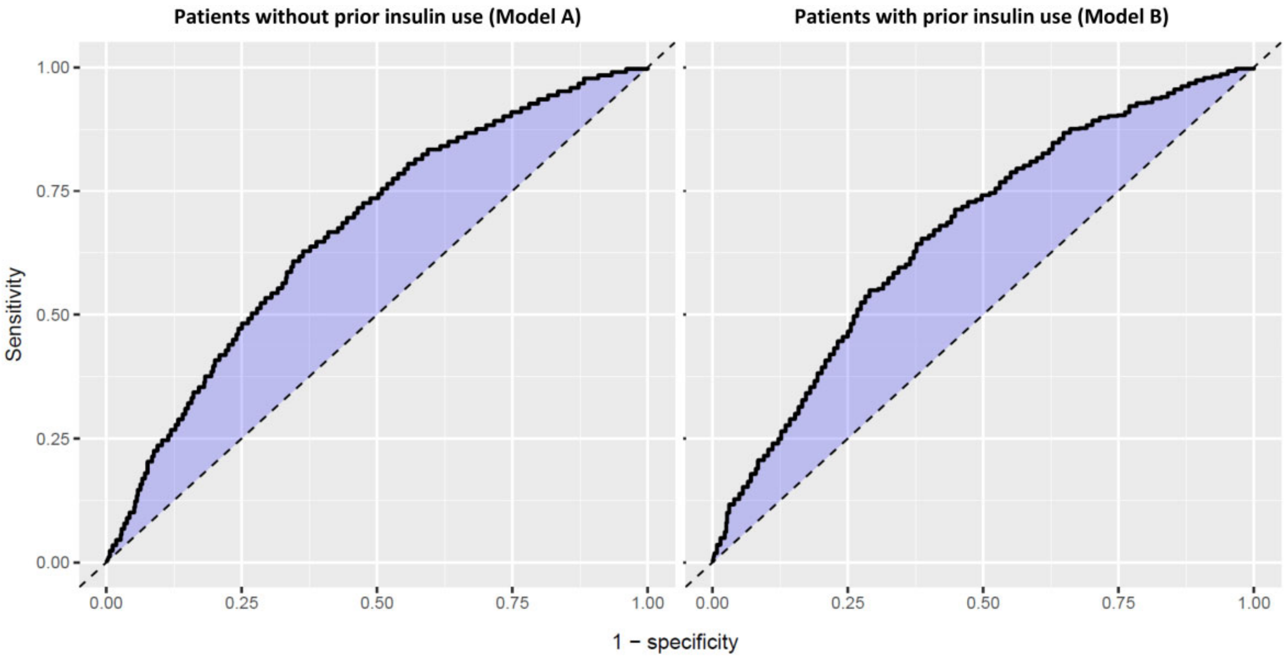

Figure 1. Receiver operating characteristic curves of two predictive models with similar areas under the receiver operating curve. Figures are based on internal validation of the models using a 25% test set after training the model in the other 75%. For model A (patients without insulin use prior to metabolic surgery), the area under the receiver operating curve was 0.725 for training and 0.673 for testing. For model B (patients with insulin use prior to metabolic surgery), the area under the receiver operating curve was 0.700 for training and 0.672 for testing.

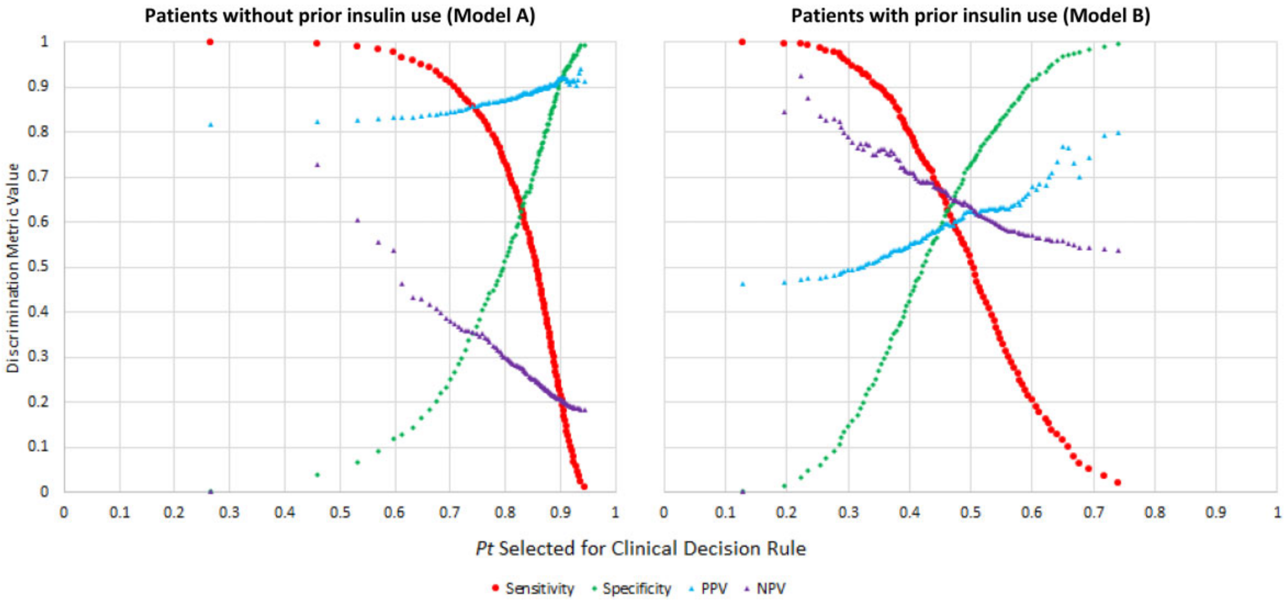

Figure 2. Probability threshold plot (PTP) of two predictive models with similar areas under the receiver operating curve. *Pt* for a clinical decision rule is the threshold above or below which patients are classified as either "test positive" or "test negative" based on their personalized Pr_i, and at which a model will possess a *Pt*-specific set of operating characteristics arising from the corresponding true/false classification. Whereas each predictive model had nearly identical ROC curves (Figure 1) and corresponding AUCs for the test set (0.672 and 0.673 among patients with vs. without prior insulin use, respectively), the PTPs for the test sets reveal substantive between-model differences in sensitivity, specificity, PPV, and NPV across the range of *Pt*. The PTP also transparently visualizes the simultaneous tradeoffs among various operating characteristics along with the distribution of *Pt* for a clinical decision rule (e.g., increasing PPV at the cost of decreasing sensitivity). For example, if *Pt* were selected to reflect the outcome prevalence within the testing and training data (e.g., *Pt* set to ∼82% for model A), a clinician selecting patients without prior insulin use to undergo metabolic surgery if they have $Pr_i > Pt$ can expect that 12.4% of patients may not experience antihyperglycemic medication cessation despite undergoing metabolic surgery (PPV = 87.6%), and they will have missed 31.4% of those may would have experienced such benefits (sensitivity = 68.6%); setting the *Pt* to a higher value (e.g., 90%) reduces sensitivity to 22.5% in favor of an increase in PPV to 92.0%. PPV, positive predictive value; NPV, negative predictive value.

Whereas each predictive model had nearly identical ROC curves (Figure 1) and corresponding AUCs for the test set (0.672 and 0.673 among patients with vs. without prior insulin use, respectively), the PTPs (Figure 2) reveal substantive between-model differences in sensitivity, specificity, PPV, and NPV across the range of *Pt*. We see in the PTP, for example, that at a $Pt = 0.60$, the model among patients with prior insulin use possesses PPV = 68.1%, NPV = 57.0%, sensitivity = 20.6%, and specificity = 91.5%. In other words, a clinician choosing patients with $Pr_i > 0.60$ to undergo metabolic surgery (i.e., setting $Pt = 0.60$) can expect that 32.9% of patients may not experience antihyperglycemic medication cessation despite undergoing metabolic surgery, and they will have missed 79.4% of those may would have experienced such benefits. In contrast, among patients without prior insulin use, PPV = 83.3% and sensitivity = 97.8% at $Pt = 0.60$. Although the between-model differences are accentuated by the underlying differences in the prevalence of the outcome between the samples used for model A and model B, the PTP illustrates how two models with nearly identical ROC curves and AUCs can possess very different operating characteristic at a given *Pt*.

The PTP also shows how adjusting *Pt* would affect the trade-off between dimensions of discriminative accuracy: e.g., increasing PPV at the cost of decreasing sensitivity. For example, if *Pt* were selected to reflect the outcome prevalence within the testing and training data (e.g., *Pt* set to ∼82% for model A), a clinician selecting patients without prior insulin use to undergo metabolic surgery if they have $Pr_i > Pt$ can expect that 12.4% of patients may not experience antihyperglycemic medication cessation despite undergoing metabolic surgery (PPV = 87.6%), and they will have missed 31.4% of those may would have experienced such benefits (sensitivity = 68.6%); setting the *Pt* to a higher value (e.g., 90%) reduces sensitivity to 22.5% in favor of an increase in PPV to 92.0%.

DISCUSSION AND CONCLUSIONS

The PTP transparently visualizes the simultaneous tradeoffs among various operating characteristics along the distribution of *Pt*. This can aid in the final selection of a *Pt* for a clinical decision rule, which should also be informed by context-dependent information on the relative importance of true/false and positive/negative classification, such as health economic evaluations or patient preference information.[2,7–9] A future area for research to further improve the PTP is the potential addition of measures of variability for each operating characteristic at a given *Pt*, such as 95% confidence intervals through the normal approximation or bootstrapping methods.

With the increasing use of machine learning-based predictive models to aid clinicians in precision medicine, improved visual displays of model results can enhance the practical application and communication of predictive models. The PTP is a new visual tool to facilitate the selection of the *Pt* based on values of the PPV, NPV, sensitivity, specificity and accuracy, and thereby translate predictive models into more useful decision tools for medical decision making.

FUNDING

This work was supported by Johnson & Johnson; however, it received no specific grant but rather was conducted as routine methodological research work.

AUTHOR CONTRIBUTIONS

Stephen S. Johnston, Stephen Fortin, Iftekhar Kalsekar, Jenna Reps, and Paul Coplan all made substantial contributions to the conception and design of this work, acquisition, analysis, and interpretation of data for the work, drafting the work and revising it critically for important intellectual content. All authors provided final approval of the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

DATA AVAILABILITY STATEMENT

The data underlying this article were provided by IBM and Optum® under licence/by permission. Data will be shared on request to the corresponding author with permission of IBM and Optum®.

CONFLICT OF INTEREST

Stephen S. Johnston, Stephen Fortin, Iftekhar Kalsekar, Jenna Reps, and Paul Coplan are employees and stockholders of Johnson & Johnson.

REFERENCES

1. Carrell DS, Albertson-Junkans L, Ramaprasan A, *et al.* Measuring problem prescription opioid use among patients receiving long-term opioid analgesic treatment: development and evaluation of an algorithm for use in EHR and claims data. *J Drug Assess* 2020; 9 (1): 97–105. doi: 10.1080/21556660.2020.1750419
2. Wynants L, van Smeden M, McLernon DJ, on behalf of the Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative, *et al.* Three myths about risk thresholds for prediction models. *BMC Med* 2019; 17 (1): 192. doi: 10.1186/s12916-019-1425-3
3. Pencina MJ, D'Agostino RB. Evaluating discrimination of risk prediction models: the C statistic. *JAMA* 2015; 314 (10): 1063–4. doi:10.1001/jama.2015.11082
4. Romero-Brufau S, Huddleston JM, Escobar GJ, *et al.* Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. *Crit Care* 2015; 19(1): 285.doi:10.1186/s13054-015-0999-1
5. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol Biogeography* 2008; 17 (2): 145–51. doi: 10.1111/j.1466-8238.2007.00358.x
6. Johnston SS, Morton JM, Kalsekar I, *et al.* Using machine learning applied to real-world healthcare data for predictive analytics: an applied example in metabolic surgery. *Value Health* 2019; 22 (5): 580–6. doi: 10.1016/j.jval.2019.01.011
7. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; 352i6. doi: 10.1136/bmj.i6
8. Salkever DS, Johnston S, Karakus MC, *et al.* Enhancing the net benefits of disseminating efficacious prevention programs: a note on target efficiency with illustrative examples. *Adm Policy Ment Health* 2008; 35 (4): 261–9. doi:10.1007/s10488-008-0168-9
9. Jiménez-Valverde A. Threshold-dependence as a desirable attribute for discrimination assessment: implications for the evaluation of species distribution models. *Biodivers Conserv* 2014; 23 (2): 369–85. doi:10.1007/s10531-013-0606-1