Journal of the American Medical Informatics Association, 28(10), 2021, 2258–2264
doi: 10.1093/jamia/ocab159
Advance Access Publication Date: 5 August 2021
Brief Communications

Brief Communications

Learning decision thresholds for risk stratification models from aggregate clinician behavior

Birju S. Patel, Ethan Steinberg, Stephen R. Pfohl, and Nigam H. Shah

Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California, USA

Corresponding Author: Nigam H. Shah, MBBS, PhD, Stanford Center for Biomedical Informatics Research, Stanford University, 1265 Welch Road, Stanford, CA 94305, USA; nigam@stanford.edu

Received 03 March 2021; Revised 26 June 2021; Editorial Decision 11 July 2021; Accepted 13 July 2021

ABSTRACT

Using a risk stratification model to guide clinical practice often requires the choice of a cutoff—called the decision threshold—on the model's output to trigger a subsequent action such as an electronic alert. Choosing this cutoff is not always straightforward. We propose a flexible approach that leverages the collective information in treatment decisions made in real life to learn reference decision thresholds from physician practice. Using the example of prescribing a statin for primary prevention of cardiovascular disease based on 10-year risk calculated by the 2013 pooled cohort equations, we demonstrate the feasibility of using real-world data to learn the implicit decision threshold that reflects existing physician behavior. Learning a decision threshold in this manner allows for evaluation of a proposed operating point against the threshold reflective of the community standard of care. Furthermore, this approach can be used to monitor and audit model-guided clinical decision making following model deployment.

Key words: decision threshold, operating point, decision-making, risk stratification model, real world data

INTRODUCTION

Physician informaticists are increasingly involved in the deployment of risk stratification models for clinical decision support. They may be asked if the predictive performance, such as the sensitivity and specificity, of a machine learning–derived model is appropriate for guiding the allocation of an intervention in their health system.[1] If so, the corresponding threshold on the predicted score from the risk stratification model's receiver-operating characteristic (ROC) curve becomes the operating point, which is the risk score cutoff that triggers a subsequent action, such as generating an alert for the early detection and treatment of sepsis.[2]

Medical decision analysis, one of the most recognized approaches for determining an operating point, calculates a decision threshold above which patients should be treated by finding the specific probability of disease at which expected benefits outweigh expected harms from an intervention.[3,4] Typically, the probability of disease for a particular patient is estimated by a physician, but leveraging advances in precision medicine, this probability can be replaced by the risk score generated by a risk stratification model. However, calculating a decision threshold using this procedure requires measuring the ratio of economic utility values for harms and benefits of treatment as an input. A significant limitation is that these utility values can be difficult to obtain in practice.[5,6] A different approach to choosing an operating point is using a clinical practice guideline, which may suggest risk-stratified treatment thresholds. However, the bases of how threshold recommendations are developed are often unreported, making it difficult to determine if these thresholds are appropriate for use.[7] Alternatively, the concept of a standard of care established by what other physicians have done or would do in similar situations has long been used in the legal context.[8] Echoing this notion, it is possible to obtain a decision threshold by having physicians respond to a series of clinical vignettes where risk scores are known.[9–13]

We recognized an opportunity to augment such evaluation beyond hypothetical clinical cases using real data from the collective

practice of many physicians recorded by the electronic health record.[14] We hypothesized that clinician behavior—reflecting how physicians across an organization balanced harms, benefits, costs, patient preferences, and resource constraints to make shared clinical decisions with actual patients in individual situations[15]—could be used to learn the latent decision threshold that was used in practice.[16] This learned threshold could then be used as a reference[17] to understand how a potential operating point compares to the current community standard of care when deploying or monitoring a risk stratification model. Our objective was to demonstrate the feasibility of a flexible mathematical approach that uses observational data to learn the underlying decision threshold implicit in physician practice. To illustrate the approach with a clinical example, we fit an equation that best captured clinical decision making from observational data, extracted decision thresholds from this equation, compared these empirical results with guideline-recommended thresholds, and assessed the stability of learned decision thresholds after the release of the risk stratification model to the public.

METHODS

As an example, we learned decision thresholds for statin treatment based on the 2013 pooled cohort equations (PCEs),[18] which predict 10-year atherosclerotic disease risk and have well-described decision thresholds.[19] We constructed a retrospective cohort from adult primary prevention patients who did not have a history of atherosclerotic disease or diabetes (as the decision to initiate statin treatment for these conditions is not conditional on the PCE risk score), underwent lipid screening by a primary care provider at Stanford Medicine before 2013, and otherwise met criteria to have their 10-year risk of atherosclerotic disease calculated. We ascertained whether patients were prescribed a statin within 180 days after lipid screening. We additionally ensured that the PCEs were generalizable to our cohort as evidenced by similar predictive performance; this ensured that it was meaningful to compare learned decision thresholds from our cohort to thresholds listed in guidelines. To calculate predictive performance, we first determined if patients developed major atherosclerotic disease within the following 10 years or before being lost to follow-up, whichever occurred first, and then calculated Harrell's C-statistic, which can be used to measure discrimination in censored survival data.[20] Our study utilized de-identified data from the Stanford Medicine Research Repository, and research on this dataset was classified as nonhuman subjects research by our institutional review board.

Our method to learn a decision threshold begins by developing a mathematical equation that fits the decision to prescribe a statin using the 10-year risk of atherosclerotic disease provided by the PCEs (Figure 1). The formulation of the equation is inspired by expected utility theory, in which a decision to treat is made if the net utility from treatment is >0.[21] For example, a provider will likely choose to prescribe a statin if the decreased risk of atherosclerotic disease outweighs potential adverse side effects such as myalgias, increased risk of diabetes, and other factors such as monetary costs. The expected utility increases as disease probability increases.[22] Similarly, we assume that the probability of treating a patient increases as disease probability increases. Thus, we search for a monotonic function that best describes the relationship between probability of treatment and PCE score across our cohort in the general form of the decision-making equation:

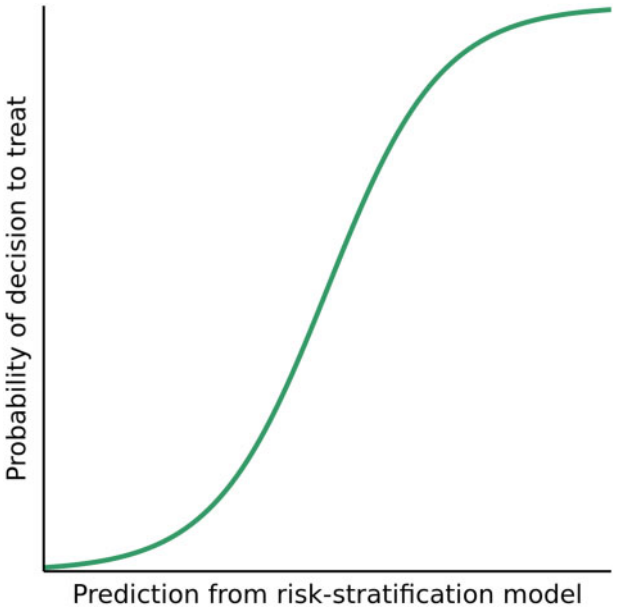

Figure 1. Hypothetical example of a decision-making equation. A decision-making equation (green line) predicts the probability of observing a decision to treat, shown here on the y-axis, as a monotonic function of the predictions from a risk stratification model, shown here on the x-axis.

$$\mathrm{P}(\mathit{treatment}) = \mathrm{function}_{\mathrm{monotonic}}(\mathit{PCEScore})$$

We test 2 alternative equations from this general form using real-world data. In the first, we use a linear transformation of the PCE risk score and leverage the standard logistic function to link this expression to the treatment probability:

$$\mathrm{P}(\mathit{treatment}) = \mathrm{logistic}(\mathrm{b}_1 \times (\mathit{PCEScore}) + \mathrm{b}_2)$$

where b_1 and b_2 are coefficients that are learned from the data. In seminal studies of decision making under risk, the relationship between utility and disease state may take the shape of a concave function.[23,24] Following this observation, we develop a second form of the decision-making equation that uses a logarithmic transformation of the PCE risk score to reflect a concave relationship:

$$\mathrm{P}(\mathit{treatment}) = \mathrm{logistic}(\mathrm{b}_1 \times \log(\mathit{PCEScore}) + \mathrm{b}_2)$$

We determine which equation empirically best describes real world decision making by calculating the Brier score, a measure of model fit to the observed data, for each equation and then selecting the equation which has the best score. This decision-making equation is then used to identify decision thresholds as follows: for a specified predicted probability of treatment, we use the equation coefficients to solve for the corresponding risk score. In our example, we examine 2 specific treatment probabilities, though other probabilities could have been chosen using different motivations. The first threshold we calculate corresponds to the PCE risk score in which the fitted decision-making equation predicts a 50% probability of treatment. This threshold corresponds to the point at which clinicians are indifferent between treating and not treating. This same threshold has been described in vignette-based studies of decision thresholds.[10,16] Because the equation predicts that half of patients with risk scores at this point are treated with statins, we refer to this threshold as the aggregate majority vote threshold. The second threshold we examine is where the probability of treatment is equal to the overall treatment proportion in the cohort.

For example, if 30% of all patients in the cohort are prescribed statins, this threshold would correspond to the PCE risk score in which the fitted decision-making equation predicts a 30% probability of treatment. While this threshold does not have a decision-theoretic interpretation, it may serve as a useful reference to identify the risk score at which the treatment probability exceeds the overall treatment rate in the population. We refer to this threshold as the aggregate treatment rate threshold. We then compare these empirically derived thresholds with thresholds stated in guidelines.

Finally, to evaluate the sensitivity of these learned decision thresholds to the release of the risk stratification model in 2013, we generate a cohort of patients screened for atherosclerotic cardiovascular disease risk after 2013 and examine whether there are differences in the equation fit or learned thresholds.

RESULTS

There were 4705 patients seen at Stanford Medicine between 2009 and 2013 who underwent primary prevention risk assessment (Table 1). Of these patients, 1045 (22.2%) were prescribed a statin. The PCEs had similar discriminative ability (C-statistic = 0.71) in this cohort compared with the original cohorts in which they were constructed.[18] The median 10-year risk score calculated by the PCEs was 2.4% in those not treated and 6.1% in those treated with statins.

As expected, we found that increasing 10-year atherosclerotic disease risk was associated with higher rates of prescribing statins (Figure 2). The log transformed decision-making equation better fit observed clinician decision making than the linear one (Brier score 0.159 for the log transformation vs 0.165 for the linear transformation). In the log-transformed equation, the proportion of the population treated with statins (22.2%) corresponded to an aggregate treatment rate threshold of 3.6% 10-year risk. The 50% probability of treatment corresponded to an aggregate majority vote threshold of 23.0% 10-year risk. The linear equation, which did not describe observed decision making as well as the log-transformed equation, produced similar decision thresholds of 5.9% and 20.7%, respectively.

The PCEs are essentially a predictive model whose continuous risk score is converted to a treatment recommendation by setting an operating point on its ROC curve (Figure 2). Based on the observed decision-making behavior from 2009 to 2013, an operating point set near the aggregate treatment rate threshold would capture patients in the borderline (5%) and intermediate (7.5%) risk categories of the treatment guidelines. An operating point set near the aggregate majority vote threshold would capture patients in the high-risk category, in which guidelines use a cutoff of 20% and document the large benefits associated with high-intensity statin therapy for these patients.

We then constructed a cohort of patients who had risk assessments performed after the publication of the PCEs in 2013 (Table 1). This cohort included 23 291 patients, of whom 4851 (20.8%) were treated with statins. We found a similar pattern of decision making (Figure 3) after the risk stratification model was released, with the recalculated log-transformed decision-making equation having similar performance (Brier score = 0.148). The aggregate treatment rate threshold was 3.9% and the aggregate majority vote threshold was 23.7%, which were 0.3% and 0.7% higher than the thresholds from the pre-2013 cohort, respectively.

Table 1. Baseline characteristics of the Stanford Medicine primary prevention population both before and after 2013

Characteristic	Pre-2013	Post-2013
Age, y	55.4 ± 9.2	56.0 ± 9.5
Female	2757/4705 (59)	13 812/23 291 (59)
Race		
White	3603/4705 (77)	13 871/23 291 (60)
Black or African American	60/4705 (1)	935/23 291 (4)
Asian	530/4705 (11)	4904/23 291 (21)
Other or unknown	512/4705 (11)	3581/23 291 (15)
Systolic blood pressure, mm Hg	126 ± 17	125 ± 17
Antihypertensive medication	1209/4705 (26)	6900/23 291 (30)
Total cholesterol, mg/dL	196 ± 35	197 ± 35

Values are mean ± SD or n/n (%). Patients were excluded from the cohort if they had a history of atherosclerotic cardiovascular disease or diabetes.

DISCUSSION

We found that an analysis of past, aggregate physician behavior can translate the combination of revealed patient preferences, clinical judgment, and decision rules used in practice into reference operating points on the ROC curve of a risk prediction model. To our knowledge, this is the first example describing how empirical decision thresholds learned from past clinician behavior of ordering an intervention could inform the deployment and monitoring of a risk stratification model. For the PCEs, our data-derived thresholds demonstrate how collective physician behavior before 2013 generally concurred with the decision thresholds later determined by an expert guideline panel. In particular, the aggregate majority vote threshold we found was located near the cutoff for the high-risk category for statin treatment, suggesting that physician practice tends to agree that benefits outweigh harms for patients in this risk category. In other evaluations, decision thresholds located near this high-risk cutoff were found to provide net benefit compared with lower thresholds, especially as patients become older.[25] On the other hand, it is interesting to note that the aggregate treatment rate threshold was located between 3.0% and 4.0%, which was the range suggested by a cost-effectiveness analysis that incorporated patient preferences.[26]

One major limitation of learning decision thresholds from aggregate decision making is that past clinical behavior could be flawed. For example, physicians may not be informed of the most recent clinical evidence or guidelines in the literature, meaning that the standard of practice in a given community may not align with national or international standards. There may be systematic bias in treatment for a particular group, or there may be such high prevalence of diagnostic errors that the learned decision thresholds are not meaningful as objective measurements of decision making. In addition, a preference that is shared by many patients toward or against a treatment could shift the aggregate majority vote threshold, even when clinical evidence suggests a different threshold. Deploying a learned threshold as an operating point could then institutionalize behaviors that conflict with efforts toward health equity or guidelines generated by experts who have reviewed the clinical evidence.

However, our approach to learn decision thresholds fulfills a unique role, especially as a reference that summarizes current behavior prior to deployment of a risk stratification model. The cholesterol management guidelines in effect before the PCEs[27] used a strategy based on tiered cholesterol targets to recommend medical therapy. When the PCEs were released as a new risk-based framework in 2013, our method could have been used to translate the

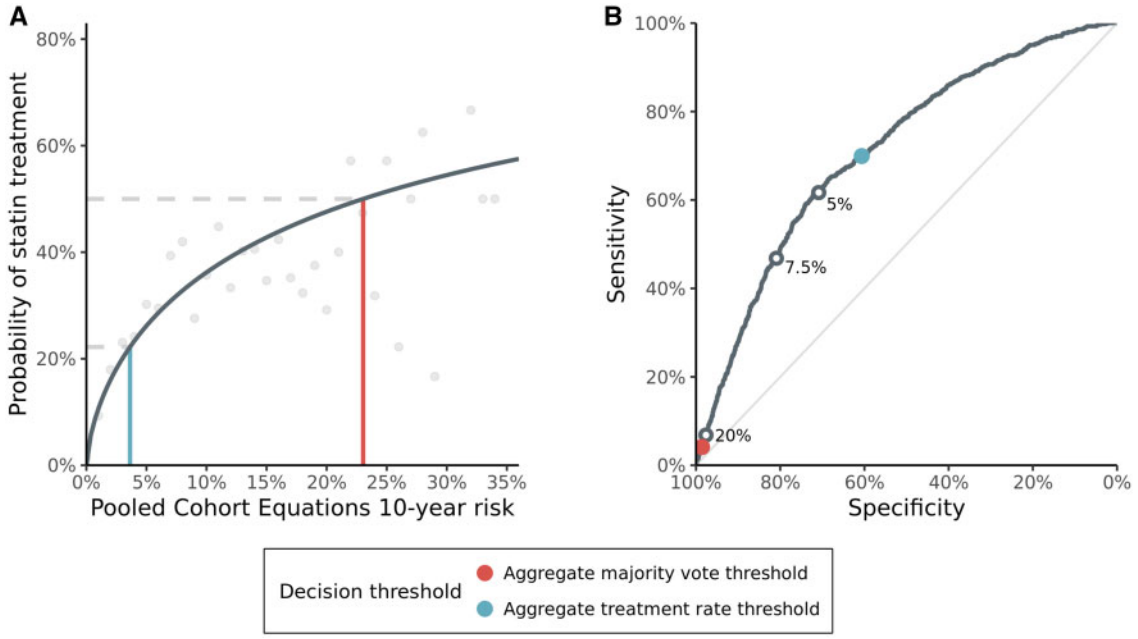

Figure 2. Decision thresholds derived from the decision-making equation. (A) The fitted decision-making equation (dark gray line) captures the relationship between the risk score and the decision to prescribe a statin (light gray circles are treatment rates for patients binned to the nearest whole number risk score percent). The proportion of the cohort treated with a statin (lower dashed line) corresponds to the aggregate treatment rate threshold of 3.6% 10-year risk (blue line), while the 50% probability of statin treatment (upper dashed line) corresponds to the aggregate majority vote threshold of 23.0% 10-year risk (red line). (B) The receiver-operating characteristic curve demonstrates the relationship between sensitivity and specificity at various potential thresholds. To make decisions based on the output of the pooled cohort equations, a continuous risk score is converted to a dichotomous recommendation by setting an operating point on the receiver-operating characteristic curve. Published clinical guidelines set operating points at 5%, 7.5%, and 20% 10-year risk (white circles), which are located near the aggregate treatment rate threshold (blue circle) and aggregate majority vote threshold (red circle) learned from the decision-making equation.

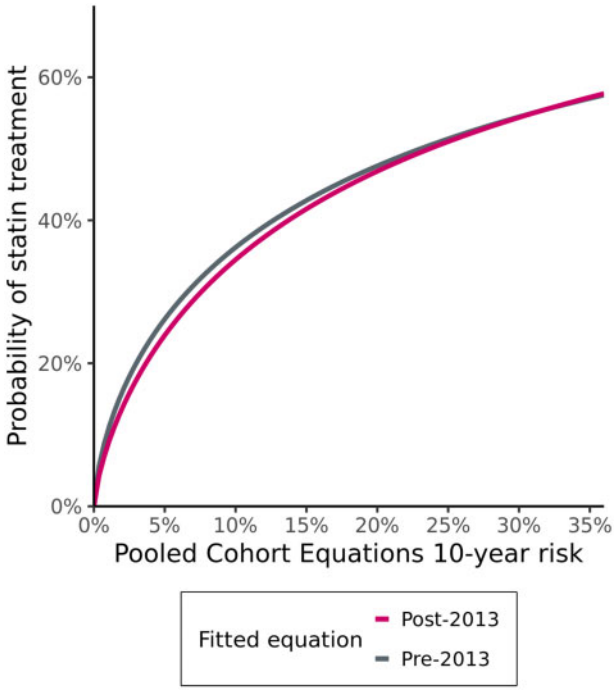

Figure 3. Comparison of fitted decision-making equations for the pre- and post-2013 cohorts. The fitted decision-making equation for the pre-2013 (gray line) and post-2013 cohorts (magenta line) closely overlap over the range of observed pooled cohort equation risk scores. This results in similar derived decision thresholds before and after publication of updated clinical guidelines in 2013.

existing community standard of practice into reference decision thresholds for the new risk stratification model. Such context could have been useful to communicate with physicians who were wary of how to reconcile these new risk scores with their practice.[28] One particular benefit of the approach that we describe is that it can be used even when a risk stratification algorithm does not already exist. With the proliferation of risk stratification models in health care,[29] there will be many novel use cases in which understanding current clinical practice could be useful for model implementers.

The proposed method can also be used to monitor patterns of decision making after the deployment of a risk stratification model to identify changes and biases in clinician behavior. For example, we observed only minor changes in learned decision thresholds after 2013, suggesting that physician behavior in practice was similar even when a new algorithm to calculate cardiovascular risk was recommended. This observation could simply reflect low adoption of the PCEs after 2013. However, even if physicians avoided using the new risk stratification model, our method would still be able to use the net decisions resulting from those other heuristics to find the corresponding risk score for the risk stratification model of interest. In other words, a decision threshold can be learned for a risk stratification model even when it is not being used in practice. The method could even be extended to evaluate differences in decision-making patterns among individual physicians,[10,16,30] helping to clarify sources of clinical variation.

The method could also be used to assess how decision making differs across relevant subgroups while controlling for the risk of the outcome.[11,13] Such threshold tests have been investigated to measure discrimination introduced by machine learning models.[31–34] Observing a difference in learned thresholds may be evidence of a differen-

tial standard of care and could generate hypotheses for further auditing to identify disparities in care processes and model derivation.[35] The combination of a learned decision threshold and a measure of calibration error for each subgroup[36] could be quite informative to characterizing sources of bias. For example, when examining for racial disparities in treatment, if the risk stratification model underestimates risk systematically for one group, it may appear that there is a different threshold being used for that group even though actual risk could be the same. Future work in this direction may help audit progress toward health equity when using risk stratification models to guide care.

We derived 2 different decision thresholds from the decision-making equation. It may be helpful to consider different thresholds when intending to understand if a risk stratification model reduces errors either due to omissions or commissions of care.[37] For example, in order to identify low-risk patients exposed to potential side effects and unnecessary costs from treatment, an operating point could flag deviances in care when patients with risk scores lower than a specified threshold are prescribed interventions. This kind of anomaly detection could use a learned decision threshold calculated at a predefined low treatment probability, such as 5%. On the other hand, to analyze omissions in care, an operating point near the aggregate majority vote threshold could identify high-risk patients who have not yet been prescribed a potentially beneficial treatment. The use of 2 different thresholds has a parallel in the threshold approach to clinical decision making.[4] In that theory, patients with disease probability below the "testing" threshold should not be treated, while those with disease probability above the "test-treatment" threshold should be treated, and patients with disease probabilities in between these thresholds should undergo further testing to determine the next course of action. Similarly, more information may be useful to determine the next step for a patient with a risk score generated by a risk stratification model that is between 2 relevant decision thresholds. Otherwise, a risk stratification model that uses an operating point between these thresholds may trigger excessive alarms with lower value information, resulting in overridden alerts[38] and provider fatigue.[39]

Our approach differs from other methods to evaluate the choice of decision thresholds by permitting flexibility for the multiple components of decision making that are not often explicitly measured, such as patient risk tolerance and individualized estimates of harms and benefits, and leveraging a descriptive approach to learn from real-world decisions. In contrast, normative theories like medical decision analysis requires an upfront measurement of utility values to complete the decision modeling process.[3,40] Multicriteria decision analysis extends this by querying stakeholders on multiple factors relevant to decision making.[41] However, these criteria also need to be determined in a transparent and reliable way, and a strategy to combine measurements across multiple criteria to choose a decision threshold is nontrivial. A different approach, decision curve analysis,[5] is useful to highlight the range of decision thresholds in which the model contributes predictive value but is not able to determine on its own if a clinician or patient would think those thresholds are reasonable given real-world context,[42] which our method directly observes. Similar to clinical vignettes, experiments in collective intelligence capture clinician decision making but can be based on real-world cases.[43] However, these studies are difficult to operationalize in practice outside of research settings since multiple physicians would have to evaluate the same patient case, which is rarely done outside of second opinions. Our approach uses individual clinical decisions but benefits from the aggregation of similar clinical scenarios that occur naturally in practice.

One important contribution of our method is the flexibility in the choice of mathematical equation to fit treatment decisions. Prior studies that learned decision thresholds from clinical vignettes used linear functions of risk scores.[10,16,44] However, we found that the logarithmic transformation generated a better descriptive equation for real-life statin prescribing as evidenced by the best Brier score. Other nonlinear monotonic functions may in fact produce better descriptive decision-making equations.[45] For example, one formulation of decision making under risk is the exponential utility function, which includes a parameter to capture the degree of risk aversion.[24] A drawback of our illustrated approach is that logistic regression would not be able to fit data to this particular monotonic function in a straightforward manner. Another approach to decision making under risk is prospect theory, which suggests that harms may be weighed more heavily than benefits by decision makers in real life.[46,47] However, it is not always clear how the output of a risk stratification model relates separately to the harms and benefits of the treatment that it is paired with in practice. The application of prospect theory to decision making with a risk stratification model warrants further investigation in the future. Nonetheless, the approach that we have described can be generalized to any monotonic function of predicted risk given that a procedure to fit the equation to data exists, and an implementer may consider using the equation that best describes real-life decision making measured by an objective scoring function such as the Brier score in order to learn decision thresholds.

Overall, learned decision thresholds can provide useful empirical information about the community standard of care to evaluate the context of a potential operating point when using a risk stratification model to guide care.

FUNDING

This work was supported by the National Heart, Lung, and Blood Institute under award number R01 HL144555. This research used data or services provided by the STAnford medicine Research data Repository (STARR), a clinical data warehouse containing live Epic data from Stanford Health Care, Stanford Children's Health, and the University Healthcare Alliance and Packard Children's Health Alliance clinics, and other auxiliary data from hospital applications such as radiology PACS. The STARR platform is developed and operated by the Stanford Medicine Research IT team and is made possible by the Stanford School of Medicine Research Office. The funding agency had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

AUTHOR CONTRIBUTIONS

BSP had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. BSP was involved in concept and design, drafting of the manuscript, and statistical analysis. All authors were involved in acquisition, analysis, or interpretation of data as well as in critical revision of the manuscript for important intellectual content. NHS was involved in administrative, technical, or material support as well in supervision.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to information that could compromise research participant privacy.

CONFLICT OF INTEREST STATEMENT

NHS reports being a cofounder of Prealize Health, which uses machine learning to better predict and responsibly contain health care costs, while also improving quality of care. BSP, ES, and SRP have no reported conflict of interest disclosures.

REFERENCES

1. Schwartz J, Moy A, Rossetti S , *et al.* Clinician involvement in research on machine learning–based predictive clinical decision support for the hospital setting: a scoping review. *J Am Med Inform Assoc* 2021; 28(3): 653–63.
2. Sendak M, Ratliff W, Sarro D, *et al.* Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. *JMIR Med Inform* 2020; 8 (7): e15182.
3. Sox H, Higgins M, Owens D. *Medical Decision Making*. Hoboken, NJ: Wiley; 2013.
4. Pauker S, Kassirer J. The threshold approach to clinical decision making. *N Engl J Med* 1980; 302 (20): 1109–17.
5. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; 26 (6): 565–74.
6. Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; 21 (1): 128–38.
7. Yu T, Vollenweider D, Varadhan R, *et al.* Support of personalized medicine through risk-stratified treatment recommendations - an environmental scan of clinical practice guidelines. *BMC Med* 2013; 11: 7.
8. Moffett P, Moore G. The standard of care: legal history and definitions: the bad and good news. *West J Emerg Med* 2011; 12 (1): 109–12.
9. Boland MV, Lehmann HP. A new method for determining physician decision thresholds using empiric, uncertain recommendations. *BMC Med Inform Decis Mak* 2010; 10: 20.
10. Ebell MH, Locatelli I, Senn N. A novel approach to the determination of clinical decision thresholds. *Evid Based Med* 2015; 20 (2): 41–7.
11. Eisenberg JM, Hershey JC. Derived thresholds. Determining the diagnostic probabilities at which clinicians initiate testing and treatment. *Med Decis Making* 1983; 3 (2): 155–68.
12. Poses RM, De Saintonge DM, McClish DK, *et al.* An international comparison of physicians' judgments of outcome rates of cardiac procedures and attitudes toward risk, uncertainty, justifiability, and regret. *Med Decis Making* 1998; 18 (2): 131–40.
13. Young MJ, Eisenberg JM, Williams SV, *et al.* Comparing aggregate estimates of derived thresholds for clinical decisions. *Health Serv Res* 1986; 20 (6 Pt 1): 763–80.
14. Longhurst CA, Harrington RA, Shah NH. A 'green button' for using aggregate patient data at the point of care. *Health Aff (Millwood)* 2014; 33 (7): 1229–35.
15. Djulbegovic B, Van den Ende J, Hamm RM, *et al.*; International Threshold Working Group (ITWG). When is rational to order a diagnostic test, or prescribe treatment: the threshold model as an explanation of practice variation. *Eur J Clin Invest* 2015; 45 (5): 485–93.
16. Plasencia CM, Alderman BW, Baron AE, *et al.* A method to describe physician decision thresholds and its application in examining the diagnosis of coronary artery disease based on exercise treadmill testing. *Med Decis Making* 1992; 12 (3): 204–12.
17. Poole S, Schroeder LF, Shah N. An unsupervised learning method to identify reference intervals from a clinical database. *J Biomed Inform* 2016; 59: 276–84.
18. Goff DC, Lloyd-Jones DM, Bennett G, *et al.*; American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association task force on practice guidelines. *Circulation* 2014; 129 (25 Suppl 2): S49–73.
19. Grundy SM, Stone NJ, Bailey AL, *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. *Circulation* 2019; 139 (25): e1082–143.
20. Harrell FE Jr, Califf RM, Pryor DB, *et al.* Evaluating the yield of medical tests. *JAMA* 1982; 247 (18): 2543–6.
21. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. *N Engl J Med* 1975; 293 (5): 229–34.
22. Habbema JD. Clinical decision theory: the threshold concept. *Neth J Med* 1995; 47 (6): 302–7.
23. Bernoulli D. Exposition of a new theory on the measurement of risk. *Econometrica* 1954; 22 (1): 23–36.
24. Howard RA. Risk preference. In: *Readings in Decision Analysis*. Menlo Park, CA: SRI International; 1977:429–65.
25. Yebyo HG, Aschmann HE, Puhan MA. Finding the balance between benefits and harms when using statins for primary prevention of cardiovascular disease: a modeling study. *Ann Intern Med* 2019; 170 (1): 1–10.
26. Pandya A, Sy S, Cho S, *et al.* Cost-effectiveness of 10-year risk thresholds for initiation of statin therapy for primary prevention of cardiovascular disease. *JAMA* 2015; 314 (2): 142–50.
27. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third Report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III) final report. *Circulation* 2002; 106 (25): 3143–421.
28. Virani SS, Pokharel Y, Steinberg L, *et al.* Provider understanding of the 2013 ACC/AHA cholesterol guideline. *J Clin Lipidol* 2016; 10 (3): 497–504.e4.
29. Challener DW, Prokop LJ, Abu-Saleh O. The proliferation of reports on clinical scoring systems: issues about uptake and clinical utility. *JAMA* 2019; 321 (24): 2405–6.
30. Hartz A, McKinney WP, Centor R, *et al.* Stochastic thresholds. *Med Decis Making* 1986; 6 (3): 145–8.
31. Pierson E, Corbett-Davies S, Goel S. Fast threshold tests for detecting discrimination. *Proc Mach Learn Res* 2018; 84: 96–105.
32. Simoiu C, Corbett-Davies S, Goel S. The problem of infra-marginality in outcome tests for discrimination. *Ann Appl Stat* 2017; 11 (3): 1193–216.
33. Pierson E. Assessing racial inequality in COVID-19 testing with Bayesian threshold tests. arXiv, doi: https://arxiv.org/abs/2011.01179, 2 Nov 2020, preprint: not peer reviewed.
34. Bakalar C, Barreto R, Bergman S, *et al.* Fairness on the ground: applying algorithmic fairness approaches to production systems. arXiv, doi: https://arxiv.org/abs/2103.06172, 10 Mar 2021, preprint: not peer reviewed.
35. Obermeyer Z, Powers B, Vogeli C, *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366 (6464): 447–53.
36. Yadlowsky S, Hayward RA, Sussman JB, *et al.* Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Ann Intern Med* 2018; 169 (1): 20–9.
37. Leape LL. Error in medicine. *JAMA* 1994; 272 (23): 1851–7.
38. Poly TN, Islam MM, Yang HC, *et al.* Appropriateness of overridden alerts in computerized physician order entry: systematic review. *JMIR Med Inform* 2020; 8 (7): e15653.
39. Co Z, Holmgren AJ, Classen DC, *et al.* The tradeoffs between safety and alert fatigue: Data from a national evaluation of hospital medication-related clinical decision support. *J Am Med Inform Assoc* 2020; 27 (8): 1252–8.
40. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. *Med Decis Making* 1988; 8 (4): 279–89.
41. Marsh K, Lanitis T, Neasham D, *et al.* Assessing the value of healthcare interventions using multi-criteria decision analysis: a review of the literature. *Pharmacoeconomics* 2014; 32 (4): 345–65.
42. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019; 3: 18.
43. Kurvers RH, Krause J, Argenziano G, *et al.* Detection accuracy of collective intelligence assessments for skin cancer diagnosis. *JAMA Dermatol* 2015; 151 (12): 1346–53.

44. Eraker SA, Eeckhoudt LR, Vanbutsele RJ, *et al*. To test or not to test–to treat or not to treat: the decision-threshold approach to patient management. *J Gen Intern Med* 1986; 1 (3): 177–82.
45. Eeckhoudt L, Lebrun T, Sailly JC. Risk-aversion and physicians' medical decision-making. *J Health Econ* 1985; 4 (3): 273–81.
46. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979; 47 (2): 263–91.
47. Attema AE, Brouwer WB, I'Haridon O. Prospect theory in the health domain: a quantitative assessment. *J Health Econ* 2013; 32 (6): 1057–65.