Learning Decision Thresholds for Risk-Stratification Models from Aggregate Clinician Behavior

Birju S. Patel, MD, MPH[1], Ethan Steinberg, MCS[1], Stephen R. Pfohl, BS[1], Nigam H. Shah, MBBS, PhD[1]

[1] Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California

Corresponding author:

Birju Patel, MD, MPH

Stanford Center for Biomedical Informatics Research, Stanford University

750 Welch Road, Suite 315, Palo Alto, CA 94304

birju@stanford.edu

Word Count: 1934

Abstract

Deploying a risk-stratification model to guide clinical practice often requires the choice of a cutoff on the model's output - called the decision threshold - to trigger a subsequent action such as an electronic alert. Choosing this cutoff is not always straightforward. Leveraging the collective information in treatment decisions made in real life, we propose an approach that learns 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 standard of care. Furthermore, this approach can be used to monitor and audit model-guided clinical decision-making following model deployment.

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 acceptable for guiding the allocation of an intervention in their health system.[1] Given this information, a corresponding operating point on the receiver operating characteristic (ROC) curve is used as a cutoff for an action trigger, such as generating an alert for the early detection and treatment of sepsis.[2]

Medical decision analysis, one of the most recognized approaches to choosing an operating point, requires economic utility values for correct and incorrect predictions as inputs to calculate a decision threshold.[3,4] However, these inputs can be difficult to obtain in practice.[5,6] Instead, physician informaticists often rely on their personal clinical experience as a guide to evaluate the merits of a particular operating point.

We recognized an opportunity to augment such evaluation beyond an individual clinical perspective using data from the collective practice of many physicians recorded by the electronic health record.[7] We hypothesized that analyzing real world clinician behavior—reflecting how physicians across an organization balanced harms, benefits, costs, patient preferences, and resource constraints to make clinical decisions with patients in individual situations—could be used to learn the latent decision threshold that was used in practice.[8] This learned threshold could then be used as a reference to understand how a potential operating point compares to current practice when deploying or monitoring a risk-stratification model. Our objective was to demonstrate the feasibility of using observational data to learn the underlying decision threshold implicit in physician practice. To illustrate the approach with a clinical example, we developed a mathematical equation that captures clinical decision-making, extracted decision thresholds, compared these empirical results to guideline recommended thresholds, and assessed the stability of learned decision thresholds over time after the release of a practice changing guideline.

Methods

As an example, we learned a decision threshold for statin treatment based on the 2013 Pooled Cohort Equations (PCEs),[9] which predict 10-year atherosclerotic disease risk and have well described decision thresholds.[10] We constructed a pre-2013 retrospective cohort from the

Stanford Medicine Research Repository of adult primary prevention patients who underwent lipid screening by a primary care provider and 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 and if they developed major atherosclerotic disease within the following 10 years.

Our method 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 form of the equation is inspired by decision analysis, where a decision to treat is made if the net utility from treatment is greater than zero. For example, a provider will likely choose to prescribe a statin if the decreased risk of atherosclerotic disease outweighs the potential side effects of the medication. We then specify net utility as a function of disease probability,[11] which is the output of the risk-stratification model, and use the sigmoid function to link increasing net utility to increasing probability of treatment. This produces the general form of the decision-making equation:

$$P(treatment\ decision) = \text{sigmoid}\big(net\ utility(PCE\ Score)\big)$$

We fit two alternative equations from this general form. In the first, we use a linear transformation of the PCE risk score as the net utility:

$$P(treatment\ decision) = \text{sigmoid}(b_1 \times (PCE\ Score) + b_2)$$

where b_1 and b_2 are coefficients that are learned from the data. In the second form of the equation, motivated by observations from economic theory,[12] we use a log transformation of the PCE risk score as the net utility:

$$P(treatment\ decision) = \text{sigmoid}(b_1 \times \log(PCE\ Score) + b_2)$$

We evaluate which equation best captures real world decision-making as evidenced by the lowest Brier score, a measure of model fit. We then use this decision-making equation to identify the corresponding decision thresholds as follows: for a specified predicted probability of treatment, we use the equation coefficients to solve for the corresponding risk score. We examine two specific treatment probabilities. The first, where the probability of treatment is equal to the overall treatment proportion in the cohort, reflects when clinicians commonly prescribe statins. We refer to this threshold as the *popular vote*. The second, where probability of treatment is equal to fifty percent or equivalently where net utility equals zero, reflects the risk score where net benefits become greater than harms. Since the equation predicts that more

than half of patients with risk scores at or above this point are treated with statins, we refer to this threshold as the *majority vote*.

Finally, to evaluate if these thresholds were sensitive to changes in clinical guidelines, we generate a cohort of patients screened for ASCVD risk after 2013 and examine whether there are differences in the equation fit or derived thresholds after the release of updated treatment guidelines in 2013.

Results

Of the 4,705 patients seen at Stanford Medicine between 2009 and 2013 who underwent primary prevention risk assessment, 1,045 (22.2%) were prescribed a statin. The PCEs had similar discriminative ability (c-statistic 0.71) in this cohort compared to the original cohorts of which they were constructed.[9] 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 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 logarithm equation, the average treatment rate (22.2%) corresponded to a popular vote decision threshold of 3.6% 10-year risk. The fifty percent probability of treatment corresponded to a majority vote decision threshold of 23.0% 10-year risk.

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 3). Based on the observed decision-making behavior from 2009 to 2013, an operating point set near the popular vote decision threshold would capture patients in the borderline (5%) and intermediate (7.5%) risk categories of the 2013 guidelines. An operating point set near the majority vote decision threshold would capture patients in the high-risk category, where the guideline uses a cutoff of 20% and documents 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. This cohort included 23,291 patients, of whom 4,851 (20.8%) were treated with statins. We found a similar pattern of decision making (figure 4) after the

guidelines were released, with the model having similar performance (Brier score 0.148). The popular vote decision threshold was 3.9% and the majority vote decision threshold was 23.7%, which are 0.3% and 0.7% more than the thresholds from the pre-2013 cohort, respectively.

Discussion

We find that an analysis of past, aggregate physician behavior can translate the combination of revealed patient preferences, clinical judgement, 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 of learning empirical decision thresholds by examining past clinician behavior of ordering an intervention. For the PCEs, our data-derived thresholds demonstrate how collective physician behavior concurred with the decision thresholds determined by an expert panel before they were released as a guideline. In addition, the popular vote decision threshold we found for statin prescribing prior to 2013 was located near the decision threshold later suggested by a well-done cost-effectiveness analysis,[13] reassuring us that the assumptions we used from utility theory to build our equations reasonably captured important decision-making information.

The cholesterol treatment guidelines in effect before 2013[14] used a strategy based on tiered cholesterol targets to recommend medical therapy. When the PCEs were released with a new risk-based framework, our method could have been used to translate the existing practice into reference decision thresholds for the new risk-stratification model. Such context could be useful to communicate with physicians who were weary of how to reconcile these new risk scores with their standard practice.[15] One particular benefit of the approach we describe is that it can be used even when a risk-stratification algorithm did not previously exist.

The proposed method can be also be used to monitor the pattern of decision-making after the deployment of a risk-stratification model to identify changes and biases in clinician behavior. For example, we saw only minor changes in learned decision thresholds after 2013, suggesting physician behavior in practice was similar even when a new algorithm to calculate cardiovascular risk became standard. The method could also be used to assess how decision-making differs across relevant subgroups after controlling for the risk of the outcome. Observing a difference in the learned thresholds may be evidence of a differential standard of care and could generate hypotheses for further auditing to identify disparities in care processes and model derivation.[16]

The decision-making equation we present uncovers two decision thresholds. It may be helpful to consider these different thresholds as reducing either errors in omission or commission of care[17] when deployed. For example, an operating point near a popular vote decision threshold could be used to flag deviances in care where patients with risk scores lower than that threshold are prescribed interventions, exposing them to potential side effects and unnecessary costs. On the other hand, an operating point set near the majority vote threshold could be used to alert a provider who has not yet prescribed a high-risk patient a potentially beneficial treatment, avoiding an omission in care. In contrast, setting a single decision threshold between these ranges attempting to reduce both omission and commission errors may trigger improper alerts, contributing to alert fatigue[18] and the significant override rates seen for clinical decision support tools.[19]

Our approach differs from other methods to evaluate the choice of decision thresholds by allowing flexibility for the multiple components of decision making that are not often explicitly measured but are still significant to the final decision, such as patient risk tolerance and individualized estimates of harms and benefits. In contrast, medical decision analysis requires an upfront measurement of utility values to complete the decision modeling process[3]. A different approach, decision curve analysis,[5] is useful to highlight the range of decision thresholds where 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, which our method directly observes.

One important choice in our presented method involves determining the best mathematical equation to fit treatment decisions. We found a small difference in the linear and logarithmic equations in the example of statin prescribing. Clinical problems with fundamentally different consequences for harms and benefits may require different non-linear transformations. For example, prospect theory suggest that utilities may take an exponential form and that harms may be more negatively valued than benefits by decision makers in real life.[20] Additionally, we used the sigmoid function in our equations to link net utility to the decision to treat, but other functions could be used. Our approach is flexible enough to accommodate different assumptions as long as they can be expressed as functions of predicted risk, and an implementer may consider using the equation that best fits decision-making measured by an objective scoring function such as the Brier score.

Overall, a learned decision threshold can provide useful empirical information to evaluate the context of potential operating points when using a risk-stratification model to guide care.

Acknowledgements

Author contributions

Dr. Patel 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.
Concept and design: Patel
Acquisition, analysis, or interpretation of data: All authors
Drafting of the manuscript: Patel
Critical revision of the manuscript for important intellectual content: All authors
Statistical analysis: Patel
Administrative, technical, or material support: Shah
Supervision: Shah

Disclosure of potential conflicts of interest

Dr. Shah 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. Dr. Patel, Mr. Steinberg, and Mr. Pfohl have no reported conflict of interest disclosures.

Funding

This work was supported by the NHLBI under award 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, 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.

References

1. Schwartz J, Moy A, Rossetti S, Elhadad N, Cato K. Clinician involvement in research on machine learning–based predictive clinical decision support for the hospital setting: A scoping review. *J Am Med Inform Assoc*. 2020;ocaa296. doi.org:10.1093/jamia/ocaa296
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. doi:10.2196/15182
3. Sox H, Higgins M, Owens D. *Medical Decision Making*. John Wiley & Sons, Ltd; 2013. doi:10.1002/9781118341544
4. Pauker S, Kassirer J. The threshold approach to clinical decision making. *N Engl J Med*. 1980 May 15;302(20):1109-17. doi:10.1056/NEJM198005153022003
5. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-574. doi:10.1177/0272989X06295361
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):128-138. doi:10.1097/EDE.0b013e3181c30fb2
7. Longhurst CA, Harrington RA, Shah NH. A 'green button' for using aggregate patient data at the point of care. *Health Aff*. 2014 Jul;33(7):1229-35. doi:10.1377/hlthaff.2014.0099
8. Poole S, Schroeder LF, Shah N. An unsupervised learning method to identify reference intervals from a clinical database. *J Biomed Inform*. 2016;59:276-284. doi:10.1016/j.jbi.2015.12.010
9. Goff DC, Lloyd-Jones DM, Bennett G, et al. 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-S73. doi:10.1161/01.cir.0000437741.48606.98
10. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*. 2019;140(11):e596-e646. doi:10.1161/CIR.0000000000000678
11. Habbema JD. Clinical decision theory: the threshold concept. *Neth J Med*. 1995 Dec;47(6):302-7. doi: 10.1016/0300-2977(95)00035-6.
12. Bernoulli D. Exposition of a New Theory on the Measurement of Risk. *Econometrica*. 1954;22(1), 23-36. doi:10.2307/1909829
13. Pandya A, Sy S, Cho S, Weinstein MC, Gaziano TA. Cost-effectiveness of 10-Year Risk Thresholds for Initiation of Statin Therapy for Primary Prevention of Cardiovascular Disease. *JAMA*. 2015;314(2):142-150. doi:10.1001/jama.2015.6822
14. 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 Dec 17;106(25):3143-421. doi:10.1161/circ.106.25.3143
15. Virani SS, Pokharel Y, Steinberg L, Chan W, Akeroyd JM, Gowani SA, Kalra A, Polsani V, Miedema MD, Jones PH, Nambi V, Petersen LA, Ballantyne CM. Provider understanding of the 2013 ACC/AHA cholesterol guideline. J Clin Lipidol. 2016 May-Jun;10(3):497-504.e4. doi: 10.1016/j.jacl.2015.11.002
16. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019 Oct 25;366(6464):447-453. doi:10.1126/science.aax2342
17. Leape LL. Error in Medicine. *JAMA*. 1994;272(23):1851–1857. doi:10.1001/jama.1994.03520230061039

18. Co Z, Holmgren AJ, Classen DC, Newmark L, Seger DL, Danforth M, Bates DW. 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 Aug 1;27(8):1252-1258. doi: 10.1093/jamia/ocaa098
19. Poly TN, Islam MM, Yang HC, Li YJ. Appropriateness of Overridden Alerts in Computerized Physician Order Entry: Systematic Review. *JMIR Med Inform*. 2020 Jul 20;8(7):e15653. doi: 10.2196/15653
20. Attema AE, Brouwer WB, I'Haridon O. Prospect theory in the health domain: a quantitative assessment. *J Health Econ*. 2013 Dec;32(6):1057-65. doi: 10.1016/j.jhealeco.2013.08.006

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, from observational data as a function of the predictions from a risk-stratification model, shown here on the x-axis.

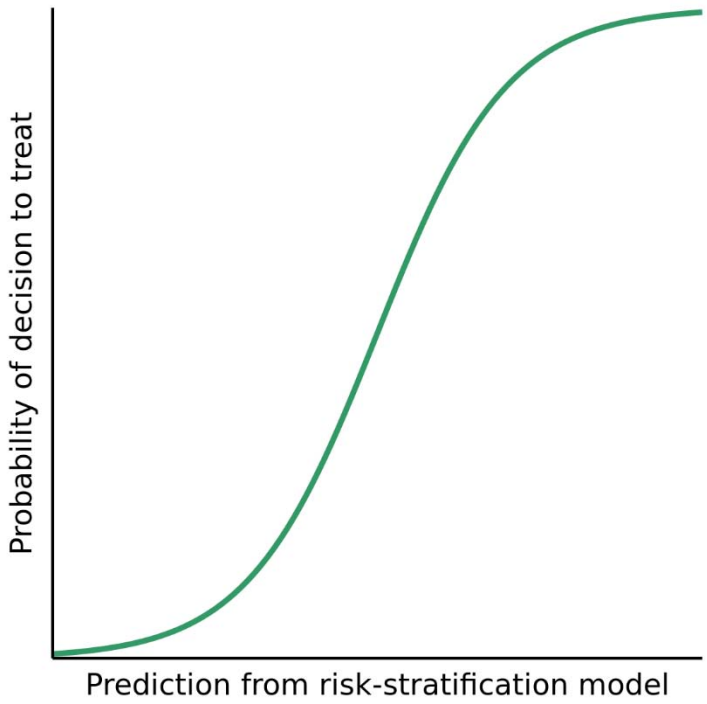


Figure 2

Decision Thresholds Derived from a Regression Model

The probability of being prescribed a statin generally increases as a function of the 10-year atherosclerotic risk score estimated by the PCEs (light grey circles are treatment rates for patients binned to the nearest whole number risk score percent). The fitted decision-making equation (dark grey line) captures the relationship between the risk score and the decision to prescribe a statin. The average treatment rate with a statin in the cohort (lower dashed line) corresponds to the popular vote decision threshold of 3.6% 10-year risk (blue line), while the fifty percent probability of statin treatment (upper dashed line) corresponds to the majority vote decision threshold of 23.0% 10-year risk (red line).

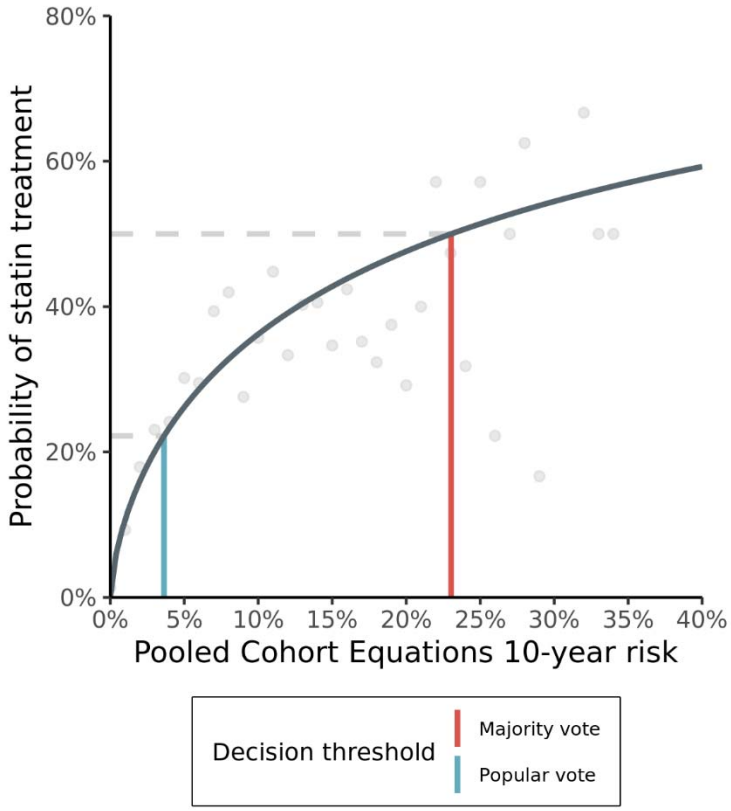


Figure 3

Operating Points Located on the Receiver Operating Characteristic Curve

The ROC curve demonstrates the relationship between sensitivity and specificity at various potential thresholds. To make decisions based on the output of the PCEs, a continuous risk score is converted to a dichotomous recommendation by setting an operating point on the ROC curve. Published clinical guidelines set operating points at 5%, 7.5%, and 20% 10-year risk (white circles). The popular vote operating point (blue circle) seen in practice in the observational data is located near the beginning of the borderline risk category identified by clinical guidelines, while the majority vote operating point (red circle) at the fifty percent treatment probability is located near the high-risk category identified by clinical guidelines.

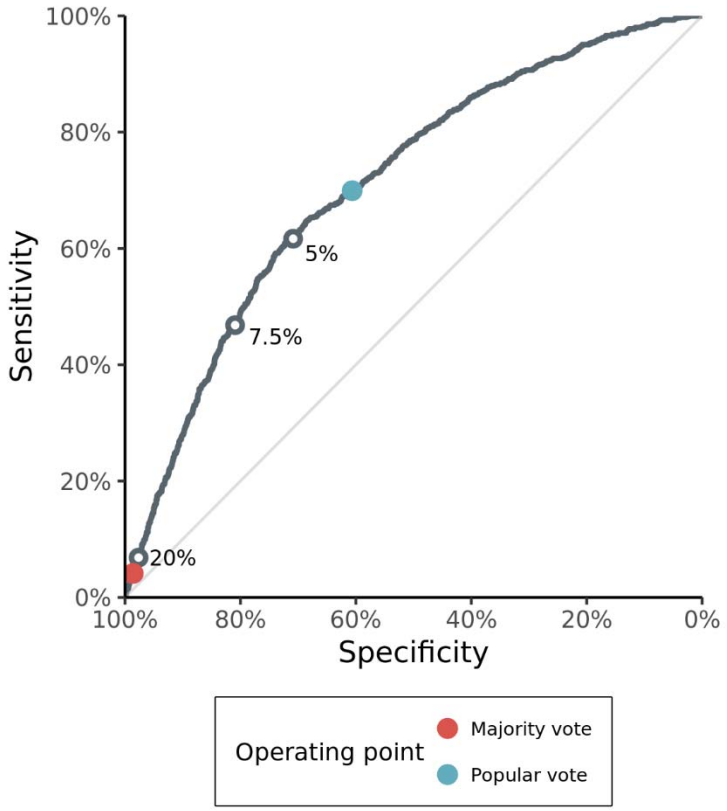
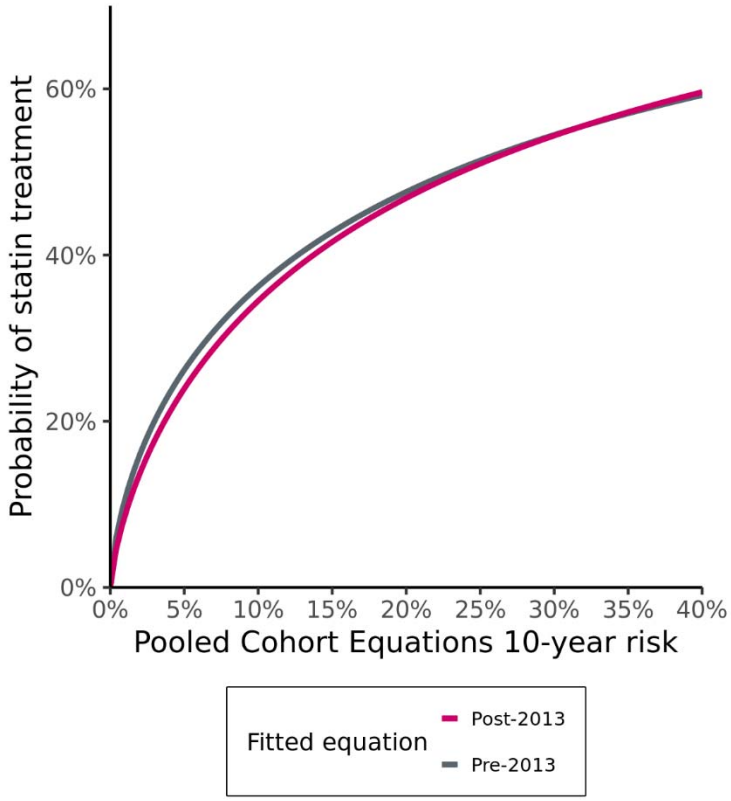


Figure 4

Comparison of Fitted Decision-Making Equations for the Pre- and Post-2013 Cohorts

The fitted decision-making equation for the pre-2013 (grey line) and post-2013 cohorts (magenta line) closely overlap over the range of observed PCE risk scores, which result in similar derived decision thresholds before and after publication of risk prediction guidelines in 2013.