The Network Relative Model Accuracy (Nerma) Score Can Quantify the Relative Accuracy of Prediction Models in Concurrent External Validations

Carl Walraven (carlv@ohri.ca)
Ottawa Hospital
Meltem Tuna
Ottawa Hospital

Research Article

Keywords:

Posted Date: April 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1521400/v1

Abstract

Background: Network meta-analysis (NMA) quantifies the relative efficacy of 3 or more interventions from trials evaluating some, but usually not all, treatments. This study applied the analytical approach of NMA to quantify the relative accuracy of prediction models with distinct applicability that are evaluated on the same population ('concurrent external validation').

Methods: We simulated binary events in 5000 patients using a known risk function. We biased the risk function and modified its precision by pre-specified amounts to create 15 prediction models with varying accuracy and distinct patient applicability. Prediction model accuracy was measured using the Scaled Brier Score (SBS). Overall prediction model accuracy was measured using fixed-effects methods accounting for model applicability patterns. Prediction model accuracy was summarized as the Network Relative Model Accuracy (NeRMA) Score which increases as models become more accurate and ranges from <0 (model less accurate than random guessing) through 0 (accuracy of random guessing) to 1 (most accurate model in concurrent external validation).

Results: The unbiased prediction model had the highest SBS. The NeRMA score correctly ranked all simulated prediction models by the extent of bias from the known risk function. A SAS macro and R-function was created and available to implement the NeRMA Score.

Conclusions: The NeRMA Score makes it possible to quantify the relative accuracy of binomial prediction models with distinct applicability in a concurrent external validation.

What Is New?

Key Findings: The Network Relative Model Accuracy (NeRMA) Score makes it possible to quantify the relative accuracy of binomial prediction models with distinct applicability in a concurrent external validation.

What This Adds to What is Known: This analysis makes it possible to measure the accuracy of multiple prediction models having distinct applicability on a common cohort of patients.

What is the Implication: When multiple prediction models are available, their relative accuracy can now be more precisely measured by conducting a concurrent external validation.

Introduction

Measuring the accuracy of prediction models in samples not used for their creation is termed 'external validation'. External validation of prediction models is essential for clinicians and researchers to confidently apply them to their patients for diagnostic, therapeutic, and prognostic decision making.[1]

Increasingly, several validated models predicting the same outcome have been published. In such situations, one will want to use the most accurate model. To identify the most accurate model, it would be

optimal to measure their accuracies in a common cohort of patients ("concurrent external validation"). Concurrent external validations ensure a 'level playing field' by comparing the performance of all models on the same group of patients.

However, analyzing a concurrent external validation is problematic when prediction models have distinct applicability patterns, with each model applying to a unique portion of patients in the concurrent external validation patients that overlaps with other models. Patients might not apply to a particular prediction model for several reasons. First, a model might require information unavailable for that patient such as undocumented patient history or results of investigations not ordered for that patient. Second, a patient may not strictly meet a model's inclusion criteria but the model should apply regardless. Such inclusion criteria are usually based on the original study's design and have no biological basis. For example, a model by Leibovici et. al.[2] returns bloodstream infection probability for internal medicine patients should apply to those admitted under other hospital services.

Because of these issues, probability estimates for some prediction models may be missing in concurrent external validation patients. These 'applicability patterns' will vary between patients. Each option for comparing model accuracy with varying applicability patterns has potential problems. Limiting analysis to patients meeting inclusion criteria of all models creates a very select cohort with questionable generalizability. Repeating relative accuracy measures within patient subgroups based on model applicability patterns might return conflicting model accuracy measures. The optimal method to measure the relative accuracy of prediction models in a concurrent external validation with distinct patient inclusion criteria is unclear.

The problem of missing probability estimates due to distinct prediction model applicability reflects the data structure of a network meta-analysis (NMA). NMA can quantify the relative efficacy of three or more interventions from trials measuring outcomes for some, but not all, interventions. In NMA, data are missing for interventions that were not studied in particular trials. In our situation, probability estimates are missing in patients to whom a model could not be applied. This study determined whether we could use analytical methods for NMA to measure the relative accuracy of binomial prediction models having distinct inclusion criteria.

Methods

This analysis used simulated data having a binary outcome. We started by generating for 5000 simulated patients an **event probability** by randomly selecting an **event log-odds** value from a normal distribution having a mean of 0 and standard deviation of 0.75. This value was converted to an event probability using the formula: $e^{log-odds}/(1 + e^{log-odds})$. As a result, the overall event probability in the patient cohort was 50%. Each person was then deemed to have had an event if a uniformly distributed random number ranging from 0 and 1 was below the event probability.

Model-based event probabilities were then generated by systematically modifying each person's event log-odds. Fifteen distinct model-based event probabilities – reflecting probability estimates from 15 different models with varying accuracies - were created for each patient in three steps:

i) a *bias parameter* (-0.8, -0.2, 0, 0.2, or 0.8) was added to the event log-odds;

ii) to the output from i) we then added a random number from a normal distribution having a mean of 0 and a standard deviation determined by the *precision parameter* (0, 0.25, or 0.50);
iii) the output from ii) was used to generate a model-based expected event probability using the formula: $e^{\text{log-odds'}}/(1 + e^{\text{log-odds'}})$.

These 3 steps generated for each simulated person a total of 15 model-based expected event probabilities having varying bias and precision. Notably:

i) One model (that with a bias parameter of 0 and a precision parameter of 0) returned a model-based event probability that equaled the true event probability used to generate the observed events. This model should be the most accurate.

ii) We had model pairs having the same precision but equivalent bias both above and below the true event probability. The accuracy of these model pairs should be the same.

Model applicability was then determined for all 15 models. Models were deemed applicable to a particular patient if a randomly selected number from a uniform distribution between 0 and 1 was less than 0.9. This process aimed to reflect distinct model applicability.

These processes were repeated 1000 times (i.e. 1000 **iterations** of 5000 patients each).

Analysis

Within distinct model applicability patterns of each iteration, the accuracy of each model was quantified with the Scaled Brier Score (SBS):

$$[1] SBS = 1 - \frac{\frac{1}{n} * \sum_1^n (p - y)^2}{\frac{1}{n} * \sum_1^n \left(\bar{y} - y\right)^2}$$

Where: n = number of people with that applicability pattern; p = model-based event probability; y = event status ('1' if event occurred and '0' if event did not occur); and $\bar{\mathbf{y}}$ = event rate in the n patients. The SBS ranges from -∞ to 1. The SBS increases with model accuracy. Models having an SBS of 0 are as accurate as predicting the average event rate for all people. SBS was used instead of the Brier Score because values of the latter vary with event rates[3] (and we expected varying event rates within different

applicability patterns). We calculated Brier Score variance using methods from Bradley et. al.[4]; this was divided by the denominator in the second term of the SBS (see equation [1]) to return SBS variance.

Excepting simulated patients to whom all prediction models applied, at least one of the 15 SBS values were missing. This data structure is like that seen in network meta-analysis (NMA) in which studies include a subset of therapeutic interventions. We therefore applied fixed-effects methods described by Loveman et. al.[5] to conduct NMA having a continuous outcome. Fixed-effects methods, rather than random-effects methods, were used because data heterogeneity was not an issue since prediction models were all simulated using the same methods. The fixed-effects model (PROC GENMOD, SAS 9.4, Cary NC) predicted SBS as a function of 2 class variables: i) the model applicability patterns; and ii) the prediction model. Observations in the fixed-effects analysis were weighted by the inverse of SBS variance. The analysis excluded applicability patterns having 5 or less patients and those having an observed event prevalence of 0 or 100%; the latter condition was implemented because BS variance is undefined in the absence of events.[4]

In the fixed-effects model, we used the unbiased and precise prediction model as the reference. Therefore, parameter estimates for all other prediction models were the absolute difference in SBS between it and the unbiased, precise model (after accounting for the different applicability patterns). The fixed effects model was repeated within each of the 1000 iterations. The mean model SBS was then calculated with 95% confidence intervals created using the percentile method.[6]

Similar to the Surface Under the Cumulative RAnking (SUCRA) score in network meta-analysis,[7] we quantified each prediction model's relative accuracy using the **NEtwork Relative Model Accuracy (NeRMA)** score as:

$$[2]\, ModelNeRMAscore = 1 - \frac{RelativeModelSBS}{RandomRiskPredictionSBS}$$

Where: Relative Model SBS is the prediction model's SBS relative to the most accurate prediction model (i.e. the model with the greatest parameter estimate values in the fixed effects model); and Random Risk Prediction SBS is the SBS for a randomly generated prediction model. The random risk prediction SBS was measured by adding to the analytical dataset a risk estimate for each person which was randomly selected from a uniformly distributed number between 0 and 1. Therefore, the model NeRMA score could range from 1 (the most accurate prediction model) through 0 (a model as accurate as randomly estimating risk) to < 0 (a model less accurate than randomly estimating risk).

Anchoring the NeRMA score at 0 to indicate the accuracy of a random model required the addition of data to the analytical dataset (namely, randomly generated event probabilities). These additional observations could narrow confidence intervals around parameter estimates in the fixed effects model. To account for this, we first measured 95% confidence interval width for each prediction model's parameter estimates in the fixed-effects model without random risk predictions. We then calculated each model's confidence interval width relative to the parameter estimate's value:

$$[3] 0.5 * ABS\left(\frac{Upper95\%CI - Lower95\%CI}{PE}\right)$$

Where: ABS is the absolute value function; Upper95%CI is the value for the upper confidence interval; Lower95%CI is the value of the lower confidence interval; and PE is the model parameter estimate. In the final fixed effects model (with random model data), this value was multiplied by parameter estimates of the final fixed-effects model and then subtracted from and added to those parameter estimates of to create 95% confidence intervals.

Results

The final analytical dataset included 1000 iterations of 5000 people each. Within each iteration, patients had a mean (95%CI) true event probability of 50.0% (49.96%-50.03%) with 50.0% (48.59%-51.34%) of people having an event.

Fifteen models were created by modifying the true event risk function, generating distinct model-based event probabilities (Fig. 1). Unbiased models returned mean model-based event probabilities of 50% (Table 1A). Models that were biased down slightly and biased down extensively had mean model-based event probabilities of 45% and 31%, respectively; models that were biased up slightly and biased up extensively had mean model-based event probabilities of 55% and 69%, respectively. Distributions of model-based event probabilities became progressively less kurtotic as precision decreased (Fig. 1). 95% confidence intervals for mean event probability widened as model precision decreased (Table 1A). The random prediction model had a mean expected event probability of 50% (3%-98%).

Because each model had a 10% probability of being inapplicable to each person in the simulation, only 20.6% of patients (20.6%-20.6%) had an expected event probability from all 15 models. Within 5000 simulated patients, there was an average of 727.5 (95%CI 726.5-728.6) distinct applicability patterns.

Model accuracy, as measured by the mean Scaled Brier Score [SBS]), was greatest in the unbiased, precise model (mean SBS 0.111, 95%CI 0.094–0.128)(Table 1B). Model accuracy decreased as bias increased. Within a given amount of bias, mean SBSs increased by the same amount whether the model was biased up or down; for example, the mean SBS for precise models that were biased slightly up or down was the same at 0.103 (Table 1B). Within each amount of bias, model accuracy decreased as precision decreased. The random prediction model had the lowest SBS of -0.335 (95%CI -0.368- -0.301).

An average of 18.4% of observations (95%CI 17.3–19.4) were excluded from the fixed-effects model because SBS variance could not be calculated (due to an event rate of 0% or 100%) or because 5 or less people had that applicability pattern.

The fixed-effects model correctly identified the relative accuracies of the prediction models (Table 1C). Compared to the unbiased, precise prediction model, all other prediction models were significantly less accurate with negative adjusted SBS values. Within the same precision level, adjusted SBS values

decreased as bias increased. In addition, the extent that adjusted SBS values decreased was the same regardless of whether the model was biased up or down. Within the same bias level, adjusted SBS values decreased as precision decreased.

Results of the fixed-effects model are clarified with the NeRMA Scores (Table 1D). This is because the NeRMA Score is anchored between the random prediction model (NeRMA score = 0) and the unbiased, precise prediction model (NeRMA score = 1). The least accurate prediction models (extensively biased and imprecise) had NeRMA scores of 0.62 (95%CI 0.54–0.70). NeRMA scores decreased as bias increased with changes in NeRMA scores being equivalent regardless of the bias direction (after controlling for the precision level). After controlling for the extent of bias in prediction models, NeRMA scores decreased as precision decreased.

Discussion

This simulation study illustrates how to measure the relative accuracy of several prediction models even when many patients in the study cohort are inapplicable to one or more of those models. Using analytical methods similar to those used in network meta-analysis, a fixed-effects model can account for variable model inclusion criteria and return relative model accuracies. The output of this fixed-effects model is summarized by the Network Relative Model Accuracy (NeRMA) score where values of 0 and 1 represent predication accuracies of random and the most accurate model, respectively.

We believe that several issues regarding our analysis are noteworthy. **First**, we believe the NeRMA score will permit concurrent external validations to identify the best binomial prediction models. If multiple validated prediction models have been published (such as, for example, with bloodstream infections[8]), the NeRMA score will let clinicians and researchers identify the most accurate model when methodological factors[9] or performance in disparate validation cohorts cannot. Applicable models with the highest NeRMA score from a concurrent external validation maximize the likelihood of returning the most accurate predictions. **Second**, our results highlight the distinction between what logistic models return (i.e. an expected probability of an event) and what concerns us (i.e. whether or not an event actually happened). The stochastic nature of the latter was reflected by notable prediction error – as measured by the Scaled Brier Score (SBS) – even when a model perfectly recreated the true risk function underlying the event. This explains why our precise and unbiased model – which perfectly recreated the risk function used to determine event status – had a mean SBS of 0.111 (95%CI 0.094–0.128). Given that the least precise and most biased models had mean SBS approximating − 0.040, this observation highlights the extent of the predictive error of binomial prediction models due to the stochastic nature of the event. **Third**, because these models deal with a binomial outcome, all outputs of these models range between 0 and 1. This results in very small SBS values. As a result, small changes in the SBS can indicate large changes in model accuracy. **Finally**, our simulation focused on models predicting a binary outcome. However, our process could be modified for models predicting a continuous outcome (perhaps by using the mean squared error as the accuracy statistic).

Several factors should be kept in mind when interpreting our results. **First**, our simulation assumed that the log-odds for the true event probability function was normally distributed. This may not necessarily be the case. Therefore, further work should be done to test this method in a range of risk distributions. **Second**, our simulation assumed that an individual model's applicability was independent of all other models (i.e. each model had a 10% probability of being inapplicable to each simulated patient). This resulted in a very large average of 727.5 (95%CI 726.5-728.6) distinct applicability patterns in each 5000 simulated patients; many of these patterns had few patients and were excluded from the fixed-effects model (mean of 18.4% [95%CI 7.3–19.4]). With real data, it is quite likely that inclusion criteria of individual models will be correlated. This would result in fewer applicability patterns with a smaller proportion of patients excluded from the fixed-effects model. **Third**, our study design ensured we knew the rank order of prediction model accuracy. The NeRMA scores reflected these rankings and returned very similar or identical NeRMA scores for simulated models of equivalent accuracy. However, we could not know or predict the amount of difference in model accuracy. We therefore cannot be certain that the NeRMA score properly calibrates accuracy. **Fourth**, NeRMA scores will vary by the prediction models included in the analysis since the model accuracy associated with a NeRMA score of 1 is determined by the most accurate model being analyzed. Therefore, while NeRMA scores are helpful to compare the accuracy of models included a particular concurrent external validation, caution will be required when comparing NeRMA scores from different analyses having distinct prediction models. **Finally**, in addition to gauging model accuracy, the SBS is influenced by the shape of the underlying true event probability risk function. The SBS for the same prediction model will increase when its risk function of true event probability risk function becomes less kurtotic (i.e. the risk function becomes 'flatter'). For example, when the simulation is repeated with the underlying true event probability function having a log-odds standard deviation of 1.5 (instead of 0.75 in the study), the SBS for the unbiased precise model increases from 0.111 to 0.295. This occurs because increasing kurtosis of the true event probability risk function brings all probability estimates closer to the mean probability of an event, thereby bringing the numerator of the SBS equation (see above) closer to the denominator and the SBS closer to 0. Therefore, the SBS can compare the accuracies of models measured in the same cohort (wherein all models are assessed on the same true event probability risk function). However, one must be cautious when SBSs measured in different cohorts are compared since the true underlying event probability risk function may differ between those cohorts. This will modify the SBS values independent of model accuracy.

In summary, the NeRMA score permits one to compare the accuracy of several prediction models even if they have distinct inclusion criteria. We hope this will facilitate the conduct of concurrent external validations to identify most accurate models and improve medical decision making.

Declarations

Ethics approval and consent to participate: This was a simulation study making ethics approval unnecessary. No human subjects were involved in the study.

Consent for publication: Both authors consent to the study's publication.

Availability of data and materials: The simulated dataset is provided.

Competing interests: The authors have no competing interests.

Funding: No funding was required for the study.

Authors' contributions: Both authors contributed to the study's conception, analysis, and writing.

Acknowledgements: This study was presented in abstract form at the International Conference on Epidemiology and Medical Statistics, December 2021 (https://waset.org/epidemiology-and-medical-statistics-conference-in-december-2021-in-rome).

References

1. Bleeker SE, Moll HA, Steyerberg EW et al. External validation is necessary in prediction research:: A clinical example. J Clin Epidemiol 2003;56:826–832.
2. Leibovici L, Greenshtain S, Cohen O, Mor F, Wysenbeek AJ. Bacteremia in Febrile Patients: A Clinical Model for Diagnosis. Arch Intern Med 1991;151:1801–1806.
3. Steyerberg EW. Evaluation of Performance. *Clinical Prediction Models: A practical approach to development, validation, and updating*. 2nd ed. New York: Springer; 2019;255–280.
4. Bradley AA, Schwartz SS, Hashino T. Sampling Uncertainty and Confidence Intervals for the Brier Score and Brier Skill Score. Weather and Forecasting 2008;23:992–1006.
5. Loveman E, Copley VR, Colquitt J et al. The clinical effectiveness and cost-effectiveness of treatments for idiopathic pulmonary fibrosis: a systematic review and economic evaluation. Health Technology Assessment 2015;19:1–364.
6. Efron B, Tibshirani RJ. Confidence intervals based on bootstrap percentiles. *An introduction to the bootstrap*. New York: Chapman&Hall; 1994;168–177.
7. Watt J, Tricco AC, Straus S, Veroniki AA, Naglie G, Drucker AM. Research Techniques Made Simple: Network Meta-Analysis. Journal of Investigative Dermatology 2019;139:4–12.
8. Eliakim-Raz N, Bates DW, Leibovici L. Predicting bacteraemia in validated models - a systematic review. Clinical Microbiology and Infection 2015;21:295–301.
9. Wolff RF, Moons KGM, Riley RD et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 2019;170:51–58.

Table

TABLE 1: Simulated prediction models and their accuracies

MODEL TRAITS RELATIVE TO TRUE RISK FUNCTION		A - Mean Expected Event Probability (95% CI)	B - Mean Scaled Brier Score† (95% CI)	C – Adjusted Scaled Brier Score Relative to Unbiased Precise Model†† (95% CI)	D – Network Relative Model Accuracy (NeRMA) Score (95% CI)
BIAS DIRECT-ION	Δ IN PRE-CISION				
↓↓	0	0.31 (0.09, 0.66)	-0.009 (-0.035, 0.016)	-0.112 (-0.138, -0.086)	0.69 (0.62, 0.76)
	↓	0.31 (0.09, 0.68)	-0.017 (-0.041, 0.008)	-0.118 (-0.144, -0.092)	0.68 (0.61, 0.75)
	↓↓	0.31 (0.07, 0.72)	-0.039 (-0.067, -0.014)	-0.137 (-0.166, -0.109)	0.62 (0.55, 0.70)
↓	0	0.45 (0.16, 0.78)	0.103 (0.086, 0.120)	-0.007 (-0.013, -0.001)	0.98 (0.96, 1.00)
	↓	0.45 (0.15, 0.79)	0.092 (0.073, 0.111)	-0.016 (-0.026, -0.007)	0.96 (0.93, 0.98)
	↓↓	0.45 (0.12, 0.83)	0.058 (0.038, 0.078)	-0.045 (-0.062, -0.028)	0.88 (0.83, 0.92)
0	0	0.50 (0.19, 0.81)	0.111 (0.094, 0.128)	0 (REF)	1
	↓	0.50 (0.18, 0.82)	0.099 (0.080, 0.118)	-0.010 (-0.018, -0.003)	0.97 (0.95, 0.99)
	↓↓	0.50 (0.15, 0.85)	0.065 (0.043, 0.085)	-0.039 (-0.054, -0.024)	0.89 (0.85, 0.93)
↑	0	0.55 (0.22, 0.84)	0.103 (0.085, 0.120)	-0.007 (-0.013, 0.000)	0.98 (0.96, 1.00)
	↓	0.55 (0.21, 0.85)	0.091 (0.071, 0.109)	-0.017 (-0.028, -0.007)	0.95 (0.92, 0.98)
	↓↓	0.55 (0.17, 0.88)	0.058 (0.036, 0.079)	-0.046 (-0.063, -0.029)	0.87 (0.83, 0.92)
↑↑	0	0.69 (0.34, 0.91)	-0.009 (-0.035, 0.015)	-0.112 (-0.137, -0.088)	0.69 (0.62, 0.76)
	↓	0.69 (0.32, 0.91)	-0.017 (-0.045, 0.008)	-0.119 (-0.145, -0.093)	0.67 (0.60, 0.75)
	↓↓	0.69 (0.28, 0.93)	-0.040 (-0.069, -0.015)	-0.138 (-0.169, -0.108)	0.62 (0.54, 0.70)
RANDOM		0.50 (0.03, 0.98)	-0.335 (-0.368, -0.301)	-0.365	0

Models were simulated having 5 distinct bias setting (biased down extensively, biased down slightly, unbiased, biased up slightly, biased up extensively) and 3 distinct precision settings (precise, precision decreased slightly, precision decreased extensively) for a total of 15 models. For each model the mean predicted probability of the event and the Scaled Brier Score (calculated as:

$$1 - \frac{\frac{1}{n} * \sum_1^n (p - y)^2}{\frac{1}{n} * \sum_1^n (\bar{y} - y)^2}$$

Where: ***n*** = number of people with that applicability pattern; ***p*** = model-based event probability; ***y*** = event status [‘1’ if event occurred and ‘0’ if event did not occur]; and = event rate in the ***n*** patients) are presented. Models with greater Scaled Brier Scores are more accurate. 95% confidence intervals were created using the percentile method from the 1000 simulated iterations. Results for randomly generated event probabilities are also presented.

†Ignores clustering of patients within applicability patterns

†† From fixed effects model; accounts for clustering of patients within applicability patterns.

††† Quantifies model accuracy from 0 (least accurate) to 1 (most accurate).

Figures

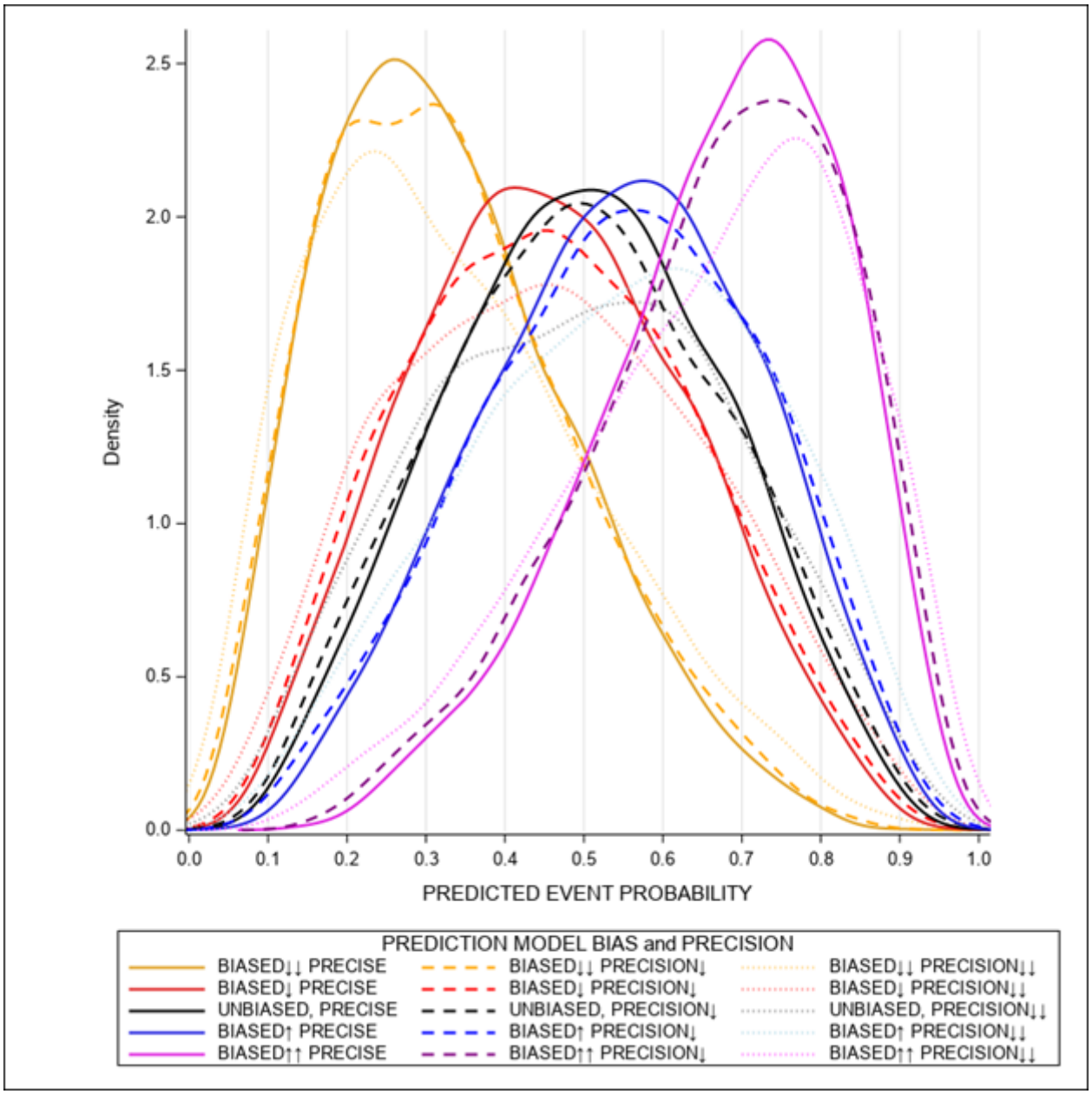

Figure 1

Distribution of patient-level expected event probabilities from 15 prediction models

Distributions of event probabilities (horizontal axis) for 5000 patients with 15 prediction models (legend) are presented. The true probability distribution (used to randomly generate actual events) is presented in black, solid line (unbiased, precise). Event probabilities for the other models are biased down (red and yellow models) or biased up (blue and violet models). Models with the same bias but that are as precise, less precise, and much less precise are represented with solid, dashed, and dotted lines, respectively.