Perspective

The number needed to benefit: estimating the value of predictive analytics in healthcare

Vincent X Liu¹, David W Bates², Jenna Wiens³, and Nigam H Shah⁴

¹Division of Research, Kaiser Permanente, Oakland, California, USA, ²Division of General Internal Medicine, Brigham and Women's Hospital, and Harvard Medical School, Boston, Massachusetts, USA, ³Division of Computer Science and Engineering, College of Engineering, University of Michigan, Ann Arbor, Michigan, USA, and ⁴Division of Biomedical Informatics Research, Stanford University, Stanford, California, USA

Corresponding Author: Vincent X Liu, MD, MS, Kaiser Permanente Division of Research, 2000 Broadway, Oakland, CA 94612, USA (vincent.x.liu@kp.org)

Received 11 February 2019; Revised 10 May 2019; Editorial Decision 13 May 2019; Accepted 17 May 2019

ABSTRACT

Predictive analytics in health care has generated increasing enthusiasm recently, as reflected in a rapidly growing body of predictive models reported in literature and in real-time embedded models using electronic health record data. However, estimating the benefit of applying any single model to a specific clinical problem remains challenging today. Developing a shared framework for estimating model value is therefore critical to facilitate the effective, safe, and sustainable use of predictive tools into the future. We highlight key concepts within the prediction-action dyad that together are expected to impact model benefit. These include factors relevant to model prediction (including the number needed to screen) as well as those relevant to the subsequent action (number needed to treat). In the simplest terms, a number needed to benefit contextualizes the numbers needed to screen and treat, offering an opportunity to estimate the value of a clinical predictive model in action.

Key words: predictive analytics, EHR, implementation science, cost-benefit analysis

Predictive analytics in healthcare has generated growing enthusiasm over the past several years, fueled by increases in the availability of high-quality patient data, flexible machine learning algorithms, and powerful computing platforms.^{1–7} This excitement is reflected in both investment in the area and in the rapidly growing body of clinical predictive model literature, particularly using electronic health record (EHR) data.^{8–11} Studies routinely highlight gains in the predictive performance of models—frequently reported based on the area under the receiver operating characteristic curve (AUC-ROC)—driven by the use of increasingly complex data and learning algorithms.^{12–14} In clinical practice today, some predictive models are also now embedded within EHR software platforms and available in real time at the point of care.

However, while predictive analytics has great potential in health care, our ability to estimate the benefits of applying a single predictive model to any specific clinical problem remains limited. One

major issue is getting providers to use the results of models, and to make the resulting actions fit within their workflows.^{15–21} But another fundamental problem is understanding a model's value. This is the tradeoff between the resources used and benefits gained by developing and deploying a specific model. The ability to assess value is critical to facilitate the effective, safe, and sustainable use of predictive tools. Without a framework for consistently estimating benefits, making decisions about predictive model development and deployment remains challenging.

For example, questions such as “What is the incremental value of a modest improvement in AUC-ROC on expected outcomes following model implementation? How do we set an optimal risk alert threshold to obtain better outcomes given our current capital and human resource constraints? How many patients with alerts are expected to benefit from an intervention?” are encountered routinely but left unanswered.

Approaches such as Net Benefit and decision analysis, which primarily quantify predictive model performance by comparing benefits and harms,^{22,23} are valuable, but may yield a partial understanding that may not hold in real-world implementation, particularly when health system resource constraints play a major role in deployment decision making. Ad hoc assessments fail to provide a framework for a rigorous and transparent comparison between alternative predictive models or those targeting different parts of a clinical workflow.²⁴ We believe that a shared framework for estimating predictive model value is needed for enhancing collaboration between data scientists, clinicians, patients, information technologists, and health system leaders.

We propose that the value of a predictive model and its resulting actions can be conceptually divided into 2 main components. The first can be represented as a number needed to screen (NNS) and arises from the characteristics of the model. NNS is the reciprocal of the positive predictive value (PPV) (also known as precision) and estimates the number of alerted patients the models must flag to identify 1 “true positive.”²⁵ The second can be represented as a number needed to treat (NNT) and arises from the effectiveness of the treatment. NNT estimates the number of true positive patients one must treat for 1 patient to benefit.^{26,27} This captures the fact that health interventions are rarely universally beneficial and can show considerable heterogeneity of treatment effects across a population.^{28–30}

The product of NNS and NNT or the number needed to benefit (NNB), contextualized with the costs of screening and treatment, can help highlight the costs and benefits of actions that result from responding to a model’s predictions. In practice, few patients who trigger a risk alert will benefit from an intervention.²² False positives will lead to diagnostic workup or treatment of at-risk patients who never would have gone on to have the adverse outcome, particularly when the prevalence of the outcome is low. And, as with all health interventions, not all true positives will experience benefits from additional testing or treatment.

For example, if a model has an NNS of 5 (corresponding to a precision or PPV of 0.20), and is followed by an intervention whose NNT is 10, then the estimated NNB from guiding actions with the model’s predictions would be 50. In other words, for 50 patients whose risk scores crossed an alerting threshold and received the intervention, only 1 patient would be expected to benefit. Accurately estimating the true value of a predictive model is significantly more complex than this simple NNB. However, even such basic estimates are missing from most current discussions.

A more comprehensive approach to evaluating a model’s utility should be based on several elements from each component of the prediction-action dyad (Table 1). Contextualizing the NNS requires an understanding of the prevalence of the outcome as well as model characteristics including the AUC-ROC, the area under the precision-recall curve, calibration metrics, and the specific risk thresholds at which action is taken.^{9,24,31,32} Together, these elements provide a detailed understanding of a model’s performance with respect to accuracy and applicability across a range of predicted risk thresholds. NNS focuses on the PPV of a model which should always be considered in the context of model sensitivity as well as the overall prevalence of the outcome. The area under the precision-recall curve displays the tradeoff between PPV and sensitivity across a range of thresholds. Considering how model performance metrics vary across risk thresholds is critical because the thresholds that initiate actions vary widely across different clinical scenarios.

Evaluating the NNT can be significantly more challenging because it is based on an understanding of a health system’s capacity for action, the costs of instantiating and maintaining the predictive model system, the presumed costs and benefits of the intervention, as well as the potential harms associated with intervention. Note that the manufacturing cost of the predictive model is another factor to consider separately. Further, patient and clinician preferences, particularly in the context of shared decision making related to health interventions, are also important components of this evaluation process.

Although many predictive models have been developed recently, relatively few prediction-action dyads have been evaluated using robust causal study designs like randomized controlled trials.³³ In some cases, the results of a handful of trials have been widely disparate. For example, the range of risk reduction reported in randomized studies of sepsis electronic alert and predictive tools ranges from no effect to a 58% reduction in inpatient mortality.^{34–36} This tremendous variability also reflects the fact that unlike in some other interventions (eg, hydrochlorothiazide for treating hypertension), interventions in response to the output of a predictive model applied to a similar population can differ depending on the environment in which they were designed and subsequently instantiated.^{37–39} This makes it challenging to assess and standardize the costs related to model use. In the case of newer EHR-based risk models, the quality of the underlying data can also vary widely across different health systems and settings. However, even with a paucity of firm data today, approximations of NNT elements with point estimates and confidence intervals could provide insight before empirical outcomes data become available.

We use an example here to highlight how having a utility framework can improve evaluation of predictive models used in clinical care. Familial hypercholesterolemia (FH) affects 0.4% of the population and results in a 20-fold higher risk of coronary artery disease.⁴⁰ Given that over 90% of the U.S. population with FH is undiagnosed, there is a critical need for better screening tools.⁴¹ In a lipid clinic, it is estimated that around 1 in 70 patients with high cholesterol have FH (ie, equivalent to an NNS of 70). Let us assume that current screening calls for a \$1000 diagnostic genetic test. Screening 70 patients leads to identifying a single true positive and requires \$70 000 in laboratory testing. An EHR-based screening model can flag patients at risk for FH with a PPV of 0.80 (ie, equating to an NNS of 1.25).⁴² As a result, the diagnostic testing cost to find 1 case would drop to \$1250. Given that efficacious interventions (ie, a PCSK9 inhibitor) are available (suggesting an NNT of 10⁴³) comparing the NNB in both situations— 700×10 without EHR-based screening versus 12.5×10 with the predictive model—would have alerted us to a significant potential benefit from using such a model to screen patients for FH.

We can take this a step further by considering costs. Assuming that the annual cost of intervention is \$6000, we can further evaluate the 2 scenarios. Without EHR-based screening, our cost to successfully treat 1 case would \$706 000 in year 1, and then \$6000 annually. With EHR-based screening, our cost to treat 1 case would be \$15 625 in year 1 and then \$6000 annually. As a result, the predictive model is associated with a dramatic cost reduction, primarily arising from the high costs of screening, in the first year.

A similar approach could lead to more nuanced discussions about the use of early warning risk prediction scores. For example, clinical deterioration, resulting in intensive care unit transfer or unexpected death, occurs in 3%–4% of hospitalized patients and is associated with substantially increased mortality and morbidity.^{44,45}

Table 1. Key elements for evaluating the utility of clinical predictive model development and deployment focusing on the prediction: action dyad

Term	Reason for inclusion
Prediction	
Receiver-operating characteristic curve	The most commonly used measure of overall model discrimination, or its ability to correctly rank pairs of observations with and without the binary outcome. The value measures the tradeoff between model sensitivity and specificity.
Precision-recall curve	A measure of model performance that is particularly useful when an outcome is highly skewed, a situation leading to a high number of false positive predictions. The value measures the tradeoff between model sensitivity and PPV.
PPV	The number of true positives, divided by the number of true and false positives. One divided by PPV is sometimes known as the NNS to detect 1 true positive or the workup-to-detection ratio. The prevalence of the outcome can strongly modify the PPV given a model with fixed operating characteristics.
Calibration metrics	A set of measures that together can be used to quantify how close the model's predicted risk values correspond with actual risk across a range of risk values.
Clinical risk thresholds	The predicted risk value at which an action or intervention is triggered. Risk thresholds vary widely between clinical scenarios, thus, overall measures of model accuracy may have only limited relevance depending on where the specific risk threshold is set for a clinical application.
Action	
System capacity	Integrating risk predictions into clinical workflows is essential for making prediction models effective and sustainable. End users, including clinicians, and health system leaders may be particularly sensitive to the limited capacity of existing teams and providers to respond to a new alerting system, particularly when the volume of alerts will significantly alter existing workflows.
Cost of intervention	The response to a risk alert will vary widely within specific clinical scenarios, but they all will entail some cost often related to provider time, closer observation or triage of patients, additional diagnostic testing, or specific treatments. There are often also costs involved with model instantiation and maintenance within EHR platforms.
Harms associated with intervention	Because only a few patients who trigger a risk alert will benefit from intervention, many patients may needlessly be subjected to an intervention which may be associated with harms. Harms can be diverse including, for example, side effects of treatment or patient anxiety about increased risk. Understanding both the costs and harms of action is essential to evaluating the downstream effects of model deployment.
Presumed outcome benefit	The costs and harms of action should be compared against the estimated outcome benefit resulting from that action. Varying the parameters of the presumed benefit, while keeping other parameters fixed, may help simulate the scenarios in which model utility can be maximized.
Cost of model instantiation	There are often substantial costs involved with model instantiation and maintenance within EHR platforms, which may not also account for resources devoted to model development and refinement.

In the simplest terms, prediction can be distilled into an NNS and action into a number needed to treat. When contextualized within this framework, the product of NNS and number needed to treat results in a number needed to benefit. The table outlines key variables in this framework that will alter the estimated number needed to benefit across different modeling and implementation scenarios.

EHR: electronic health record; NNS: number needed to screen; PPV: positive predictive value

A traditional point-based scoring system to identify high-risk inpatients, like the National Early Warning Score, demonstrates a PPV of 0.07 at a risk threshold of 8 points (corresponding to an NNS of 14.3) and a sensitivity of 0.20.²⁵ In contrast, a more complex early warning score, the Advance Alert Monitor (AAM), has a PPV of 0.08 (NNS = 12.5) and a higher sensitivity of 0.25 at a threshold of 7.5% predicted risk. If the scores otherwise identified the same patients at risk and resulted in the same interventions effective in 1 in every 7 patients, the NNB would be 100 with the simpler NEWS versus 88 with AAM score. In this highly simplified scenario, use of the AAM score could be justified if the incremental costs were <14% higher than the current costs, without including the cost of model manufacturing. Additional NNT scenarios, for example, in which a health system only could increase resource allocation for this effort by 5% or in which there were rare but important harms associated with model alerting, could further inform critical decisions about model use.

To be sure, the NNB is a simple proposal and only provides a rough estimate. Our core argument is that deployment and evaluation of predictive models in health care must consider both the properties of the model and the properties of the resulting action in the health system. Once the characteristics of the model, and its

responding action are in hand, existing approaches for quantifying the utility of healthcare interventions, like decision analysis or convex optimization,^{23,46} can offer considerably more nuanced quantitative guidance beyond the NNB. Future work is needed to accurately characterize NNB values and will require the input of diverse stakeholders with expertise in data science, clinical practice, operations, and technology. And, as more and more predictive models are developed and put in place, we believe that reporting NNB values can improve the effective and sustainable use of predictive analytics.

There is great interest in utilizing predictive analytics, but we believe that to go from hype to value, it will be essential to establish a shared framework for estimating predictive model benefit. Otherwise, we run the risk of falling prey to hype, paying high development and maintenance costs for predictive models, and failing to realize benefits in patient care and outcomes. We would also be playing right into the existing narrative of health information technology: overpromising, but under delivering. With the increasing use of predictive models in clinical practice, better approaches to assess benefits will help diverse stakeholders, from health system leaders to clinicians, patients and families understand and demonstrate the value of these promising tools.

FUNDING

This work was supported by National Institutes of Health grants R35GM128672 (to VXL) and R01LM011369 (to NHS) as well as the Permanente Medical Group (to VXL) and the Department of Medicine at Stanford (to VXL).

CONTRIBUTIONS

All authors contributed to the conception of the work; drafting or revising it critically for important intellectual content; providing final approval of the version to be published; and agreeing 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.

CONFLICT OF INTEREST STATEMENT

NHS is a co-founder of Cardinal Analytix Inc, which use advanced machine learning to better predict and responsibly contain health care costs. DWB consults for EarlySense, which makes patient safety monitoring systems. He receives cash compensation from CDI (Negev), Ltd, which is a not-for-profit incubator for health IT start-ups. He receives equity from ValeraHealth which makes software to help patients with chronic diseases. He receives equity from Clew which makes software to support clinical decision making in intensive care. He receives equity from MDClone which takes clinical data and produces deidentified versions of it. Dr Bates' financial interests have been reviewed by Brigham and Women's Hospital and Partners HealthCare in accordance with their institutional policies. VXL and JW have no competing interests to declare.

REFERENCES

- Parikh RB, Kakad M, Bates DW. Integrating predictive analytics into high-value care: the dawn of precision delivery. *JAMA* 2016; 315 (7): 651–2.
- Parikh RB, Schwartz JS, Navathe AS. Beyond genes and molecules—a precision delivery initiative for precision medicine. *N Engl J Med* 2017; 376 (17): 1609–12.
- Schneeweiss S. Learning from big health care data. *N Engl J Med* 2014; 370 (23): 2161–3.
- Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA* 2013; 309 (13): 1351–2.
- Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 2016; 375 (13): 1216–9.
- Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl J Med* 2017; 376 (26): 2507–9.
- Ohno-Machado L. Using health information technology for clinical decision support and predictive analytics. *J Am Med Inform Assoc* 2017; 24 (1): 1.
- Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017; 24 (1): 198–208.
- Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc* 2018; 25 (8): 969–75.
- Paige E, Barrett J, Stevens D. Landmark models for optimizing the use of repeated measurements of risk factors in electronic health records to predict future disease risk. *Am J Epidemiol* 2018; 187 (7): 1530–8.
- Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff (Millwood)* 2014; 33 (7): 1123–31.
- Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 2009; 302 (21): 2345–52.
- Pencina MJ, D'Agostino RB Sr. Evaluating discrimination of risk prediction models: the C statistic. *JAMA* 2015; 314 (10): 1063–4.
- Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15 (4): 361–87.
- Dummet BA, Adams C, Scruth E, Liu V, Guo M, Escobar GJ. Incorporating an early detection system into routine clinical practice in two community hospitals. *J Hosp Med* 2016; 11 (Suppl 1): S25–31.
- Bates DW, Kuperman GJ, Wang S, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc* 2003; 10 (6): 523–30.
- Miller K, Mosby D, Capan M, et al. Interface, information, interaction: a narrative review of design and functional requirements for clinical decision support. *J Am Med Inform Assoc* 2018; 25 (5): 585–92.
- Tso GJ, Tu SW, Oshiro C, et al. Automating guidelines for clinical decision support: knowledge engineering and implementation. *AMIA Annu Symp Proc* 2016; 2016: 1189–98.
- Middleton B, Sittig DF, Wright A. Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearb Med Inform* 2016; Suppl 1: S103–16.
- Sittig DF, Wright A, Osheroff JA, et al. Grand challenges in clinical decision support. *J Biomed Inform* 2008; 41 (2): 387–92.
- Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 2005; 330 (7494): 765.
- Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; 352: i6.
- Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; 26 (6): 565–74.
- Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; 350: g7594.
- Kipnis P, Turk BJ, Wulf DA, et al. Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. *J Biomed Inform* 2016; 64: 10–9.
- Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988; 318 (26): 1728–33.
- Nuovo J, Melnikow J, Chang D. Reporting number needed to treat and absolute risk reduction in randomized controlled trials. *JAMA* 2002; 287 (21): 2813–4.
- Rothwell PM, Warlow CP. Prediction of benefit from carotid endarterectomy in individual patients: a risk-modelling study. European Carotid Surgery Trialists' Collaborative Group. *Lancet* 1999; 353 (9170): 2105–10.
- Longford NT. Selection bias and treatment heterogeneity in clinical trials. *Stat Med* 1999; 18 (12): 1467–74.
- Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA* 2007; 298 (10): 1209–12.
- Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA* 2017; 318 (14): 1377–84.
- Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016; 74: 167–76.
- Wallace E, Uijen MJ, Clyne B, et al. Impact analysis studies of clinical prediction rules relevant to primary care: a systematic review. *BMJ Open* 2016; 6 (3): e009957.
- Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient

- survival and hospital length of stay: a randomised clinical trial. *BMJ Open Res* 2017; 4 (1): e000234.
35. Semler MW, Weavind L, Hooper MH, *et al.* An electronic tool for the evaluation and treatment of sepsis in the ICU: a randomized controlled trial. *Crit Care Med* 2015; 43 (8): 1595–602.
 36. Downing NL, Rolnick J, Poole SF, *et al.* Electronic health record-based clinical decision support alert for severe sepsis: a randomised evaluation. *BMJ Qual Saf* 2019 Mar 14 [E-pub ahead of print. doi:10.1136/bmjqs-2018-008765. PMID: 30872387.
 37. Oh J, Makar M, Fusco C, *et al.* A generalizable, data-driven approach to predict daily risk of clostridium difficile infection at two large academic health centers. *Infect Control Hosp Epidemiol* 2018; 39 (4): 425–33.
 38. Marafino BJ, Park M, Davies JM, *et al.* Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. *JAMA Netw Open* 2018; 1 (8): e185097.
 39. Rajkomar A, Oren E, Chen K. Scalable and accurate deep learning with electronic health records. *Digit Med* 2018; 1: 18.
 40. de Ferranti SD, Rodday AM, Mendelson MM, Wong JB, Leslie, LK, Sheldrick RC. Prevalence of Familial Hypercholesterolemia in the 1999 to 2012 United States National Health and Nutrition Examination Surveys (NHANES). *Circulation* 2016; 133 (11): 1067–72.
 41. Knowles JW, Rader DJ, Khourey MJ. Cascade screening for familial hypercholesterolemia and the use of genetic testing. *JAMA* 2017; 318 (4): 381–2.
 42. Banda JM, Sarraju A, Abbasi F, *et al.* Finding missed cases of familial hypercholesterolemia in health systems using machine learning. *NPJ Digit Med* 2019; 2 (1): 23.
 43. Behr PEB, Moriguchi EH, Castro I, *et al.* Indications of PCSK9 inhibitors for patients at high and very high cardiovascular risk. *Arq Bras Cardiol* 2018; 111 (1): 104–8.
 44. Escarce JJ, Kelley MA. Admission source to the medical intensive care unit predicts hospital death independent of APACHE II score. *JAMA* 1990; 264 (18): 2389–94.
 45. Escobar GJ, Greene JD, Gardner MN, Marelich GP, Quick B, Kipnis P. Intra-hospital transfers to a higher level of care: contribution to total hospital and intensive care unit (ICU) mortality and length of stay (LOS). *J Hosp Med* 2011; 6 (2): 74–80.
 46. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. *Med Decis Making* 1988; 8 (4): 279–89.