Moving beyond AUC: decision curve analysis for quantifying net benefit of risk prediction models

Mohsen Sadatsafavi [1], Amin Adibi [1], Milo Puhan [2], Andrea Gershon [3], Shawn D. Aaron [4] and Don D. Sin [5]

[1]Respiratory Evaluation Sciences Program, Collaboration for Outcomes Research and Evaluation, Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver, BC, Canada. [2]Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland. [3]Institute of Clinical Evaluation Sciences, University of Toronto, Toronto, ON, Canada. [4]The Ottawa Hospital Research Institute, University of Ottawa, Ottawa, ON, Canada. [5]Centre for Heart Lung Innovation, University of British Columbia, Vancouver, BC, Canada.

Corresponding author: Mohsen Sadatsafavi (msafavi@mail.ubc.ca)

Check for updates

Shareable abstract (@ERSpublications)
Statistical constructs such as ROC/AUC do not answer the critical question of how much clinical utility a risk prediction model confers. This paper overviews decision curve analysis, a novel method for quantifying net benefit of a risk prediction model. https://bit.ly/3h1rraX

Cite this article as: Sadatsafavi M, Adibi A, Puhan M, *et al*. Moving beyond AUC: decision curve analysis for quantifying net benefit of risk prediction models. *Eur Respir J* 2021; 58: 2101186 [DOI: 10.1183/13993003.01186-2021].

Copyright ©The authors 2021. For reproduction rights and permissions contact permissions@ersnet.org

Received: 26 April 2021
Accepted: 31 Aug 2021

Background

Risk prediction models that quantify the risk of clinically important events based on multiple patient characteristics are cornerstones of precision medicine. There are many risk prediction models for respiratory diseases, though only a few have been widely adopted [1, 2]. With the increasing availability of data and accessibility to electronic health records that can automate risk prediction, the uptake of predictive analytics in respiratory medicine will increase.

Classically, a "good" risk prediction model is identified as one that has good calibration and high discrimination. The former is usually expressed through a calibration plot or slope, and the latter through a receiver operating characteristic (ROC) curve and its area under the curve (AUC). The AUC, in particular, is often touted as a singular measure of model performance [3]. However, AUC is a statistical index and thus cannot reflect the usefulness of a model in clinical practice [4].

Critically, neither model discrimination nor model calibration can answer the question of whether a risk prediction model can improve disease management. Direct assessment of such clinical utility requires randomised trials that compare the outcomes of using *versus* not using the risk model, which is not feasible in the early phases of risk model development. Prior to late-phase clinical trials, a decision curve analysis can be used to quantify the expected "net benefit" of a risk prediction model [5], based on the same data (predicted risks and observed outcomes) as those required to create ROC curves and calibration plots. Decision curve analysis can thus accompany statistical metrics when the performance of a risk prediction model is being evaluated. Here, we present an overview of this approach and provide a case example of how decision curves can be used to investigate the clinical utility of risk prediction models in COPD.

How does it work? A case study in prediction of acute exacerbations of COPD

Major COPD guidelines and management strategies, such as the Global Initiative for Chronic Obstructive Lung Disease (GOLD), recommend a stepwise approach to COPD pharmacotherapy that is based on the "frequent exacerbator" phenotype [6]. GOLD defines the frequent exacerbator status as having $\geqslant$2 moderate or $\geqslant$1 severe acute COPD exacerbations in the previous 12 months. This approach may be improved by combining patients' prior exacerbation history with other characteristics. One such algorithm is ACute COPD Exacerbation Prediction Tool (ACCEPT), which predicts acute exacerbation risk based on 13 predictors covering demographics, exacerbation history, lung function, smoking status, functional

capacity and medication history [7]. We applied ACCEPT to patient-level data from ECLIPSE (Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints), a prospective cohort study [8]. We based our evaluation on COPD patients with complete data (n=1819); we used information in the first follow-up year to predict the risk of moderate or severe acute exacerbations of COPD for the second year of follow-up. ECLIPSE has been used as an external validation dataset of ACCEPT and, as such, no refitting or recalibration was performed. The model was well calibrated in individuals with a high risk of acute COPD exacerbation, but overestimated the risk in patients at a low risk (*e.g.* those with a negative acute exacerbation history). The calculated AUC for predicting the 1-year risk of acute COPD exacerbation using ACCEPT was 0.78 (95% CI 0.76–0.80). In comparison, the AUC based only on the frequent exacerbator status was 0.68 (95% CI 0.66–0.70). However, does this level of calibration combined with the significantly improved AUC with ACCEPT translate into better care of COPD patients?

To address this question, we considered two common clinical scenarios in which step-up therapy may be warranted:

Scenario 1: Should patients who are on monotherapy with long-acting muscarinic antagonists (LAMA) be switched to dual bronchodilator therapy (LAMA plus long-acting beta-2 agonists (LABA))?

Scenario 2: Should patients who are on maximum inhaled therapy be provided with azithromycin as an add-on therapy?

Based on GOLD recommendations, patients should be considered for step-up therapy if they are a frequent exacerbator.

Treatment threshold at the core of net benefit calculation for risk predictors

Unlike the binary definition of a frequent exacerbator, a risk model generates continuous risk scores, which need to be dichotomised by specifying a "treatment threshold" (such that step-up therapy is considered if the predicted risk is above this threshold). The threshold varies for treatments depending on their benefit/harm profile. For example, given the relatively safe profile of LABA compared with azithromycin (which is associated with increased risk of cardiovascular events, hearing loss and antibacterial resistance [9]), it is likely that patients and physicians have a lower threshold for adding LABA than azithromycin to current treatment. For the purposes of illustration, let us assume that a guideline panel, following consultations with patients and considering all aspects of treatment benefit and harm, concludes that the threshold for adding LABA is at an acute COPD exacerbation risk of >20% per year, and that for azithromycin is >40%. Ideally, the optimal treatment threshold should be decided using a rigorous, quantitative evaluation of benefits and harms of therapies, which should also take into account patient preferences on how they weigh the benefits against harms. For example, preventing acute COPD exacerbations in patients with low lung function might be more critical; thus the treatment threshold might be lower in such patients. An appealing aspect of decision curves is that the clinical utility of a model can be evaluated at any threshold that is considered relevant, thus separating the task of determining clinical utility at a given threshold from that of determining optimal thresholds.

Returning to the azithromycin scenario, a threshold value of 40% means that we would not step-up therapy in patients whose predicted exacerbation risk is <40% per year, while we would consider adding azithromycin for those with a predicted risk of >40% per year. At a risk of exactly 40%, there would be ambivalence.

What does ambivalence mean? Consider a group of patients who all have a predicted risk of exactly 40%. If we give azithromycin to all such patients, we would appropriately treat the 40% who would have otherwise exacerbated, and inappropriately the 60% who would not have exacerbated. Our ambivalence means that we consider the overall utility of treating all patients in this cohort (all with acute exacerbation of COPD risk of 40%) to be equal to treating none of them. This can only happen when the benefit of treating 40 true-positives with azithromycin cancels out the harm of treating 60 false-positives. This indicates that, by using a threshold value of 40%, we have placed a weight of 40/60, or 2/3, on the harm of a treating a false-positive (and causing some side-effects) relative to the benefit of finding and then treating a true-positive patient (and thus reducing the exacerbation risk).

The critical concept behind decision curve analysis is that we can use this relative weight of 2/3 as an exchange rate between true and positive diagnosis to calculate an overall net benefit (benefits minus harms). In ECLIPSE, at a treatment threshold of 0.40, 53.9% of the sample were true-positives, and 36.0% were false-positives according to the ACCEPT prediction model. Because each false-positive diagnosis is

weighted as two-thirds of a true-positive diagnosis, the net benefit of using the ACCEPT prediction model at a threshold of 40% is 0.539–0.360×2/3=0.299.

Generally, for any risk threshold r, the exchange rate between true- and false-positives can be calculated as $r/(1-r)$. That is, one true-positive diagnosis is equal to (but in opposed direction of) $r/(1-r)$ true-negative diagnoses. Net benefit associated with any threshold can thus be calculated as follows:

$$Net\ benefit = \frac{Number\ of\ true\ positives}{Sample\ size} - \left(\frac{Number\ of\ false\ positives}{Sample\ size} \times \frac{Threshold}{1 - Threshold}\right)$$

The first term in this equation is the benefit of using the treatment for patients who will otherwise experience the event (true-positive), and the term in brackets is the harm associated with treating false-positive patients, converted to the (negative) true-positive units by using the exchange rate. We note that this net benefit is in "true-positive" units, and is mostly interpretable in comparison with the net benefit of using alternative models.

In practice, we calculate the net benefit across an entire range of thresholds (0 to 1) to evaluate other thresholds that might be of interest to decision-makers. The resulting graph is the decision curve, as shown in figure 1 for our case study. The figure depicts the net benefit of the ACCEPT model *versus* the use of acute COPD exacerbation history alone (by employing the frequent exacerbator definition), and two default decisions that should always be present in a decision curve: treating none (with a net benefit of zero) and treating all.

This curve now provides an opportunity to revisit our specific questions. For the decision to add azithromycin (a threshold value of 40%), ACCEPT clearly provides net benefit above acute exacerbation of COPD history alone, as well as the two default strategies of treating all or none. Interestingly, at this threshold, the use of acute COPD exacerbation history does not provide any extra benefit over the decision

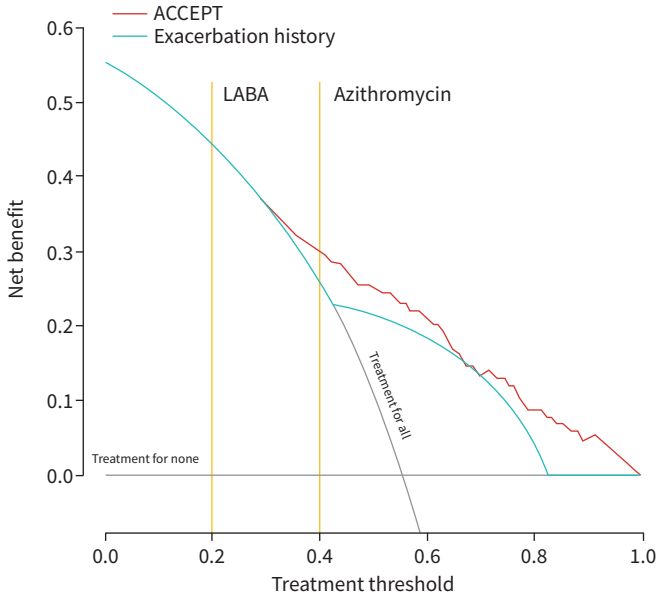

FIGURE 1 The decision curve for the ACCEPT (Acute COPD Exacerbation Prediction Tool) model (red) compared with frequent exacerbator status (blue). The vertical lines highlight the two treatment thresholds (long-acting beta-2 agonist (LABA) and azithromycin) discussed in the text. From this curve, it is apparent that the clinical utility of both the ACCEPT model and frequent exacerbator definition depends on the treatment threshold. At threshold values less than 0.28, there is no point in using any prediction tool; instead, in this range of treatment threshold, all individuals should be treated. At the threshold values over 0.28, ACCEPT provides more net benefit compared with the frequent exacerbator definition. The frequent exacerbator definition is tied with ACCEPT within a narrow range around the threshold value of 0.7, but otherwise does not provide clinical utility above ACCEPT.

to treat all patients. On the other hand, for the LABA therapy scenario (which has a treatment threshold of 20%), the best decision is to provide treatment to all patients, as neither ACCEPT nor acute COPD exacerbation history provides any extra utility. An overall interpretation of this decision curve is provided in the caption for figure 1.

A common misconception about decision curve analysis is that it can be used to determine the threshold for treatment. For example, the fact that the net benefit of using azithromycin is higher at a risk threshold of 30% compared with 40% (figure 1) does not mean that one should move the treatment threshold to 30%. Because at each threshold, a different weight is assigned to false-positive classifications, net benefits at different thresholds cannot be compared with each other. The optimal threshold should always be decided by considering benefits and harms of treatments and the preferences of patients. This threshold can then be applied to the decision curve to determine which treatment strategy provides the highest clinical utility.

The main goal of this overview was to familiarise the reader with the concept behind decision curves. Interested readers are referred elsewhere [10] for detailed methodology and its application in survival analysis, adjustments for overfitting, and incorporation of the burden of testing into net benefit calculations [10]. Simple-to-use software, along with tutorials and example data, are available at http://decisioncurveanalysis.org/. There are also methods for calculating confidence intervals around decision curves [11]. However, in decision theory, it is the "expected" benefit of a decision that matters; as such, uncertainty in decision curves is not as relevant as uncertainty in statistical metrics of model performance [10].

Current state of the art in respiratory medicine

Decision curve analysis has been considered a "breakthrough" in predictive analytics [12], as it directly quantifies the clinical utility of a risk prediction algorithm. Decision curve analysis is rapidly gaining popularity in medical literature. The original 2006 publication [5] on this topic has received >2100 citations as of September 2021, and major medical journals are routinely recommending its use for evaluating prediction models [10]. However, decision curves have only been sporadically used in respiratory research [13, 14].

How is it likely to be used in future?

With the increasing availability of research data to develop prediction models and easier integration of such models at point of care *via* electronic health records, it is likely that multivariable risk prediction will eventually replace simple risk stratification schemes such as the frequent exacerbator phenotype (which was recently shown to be an unstable classifier [15]). However, as our case study demonstrated, the clinical utility of a prediction model varies depending on the treatment threshold. As such, summary indices such as AUC and calibration slope (that are not dependent on treatment threshold) are insufficient in determining whether or not the use of risk prediction is clinically beneficial in a given treatment context. Decision curves can provide direct information on the net benefits of a management approach (*e.g.* the use of one model *versus* another) that can guide therapeutic choices. In addition to researchers, who can use decision curves to investigate the clinical utility of competing risk prediction models, guideline developers can apply this methodology to investigate whether the benefits of a treatment outweigh its harms at a plausible range of treatment thresholds. We invite the respiratory research community to fully utilise decision curves in exploring the clinical utility of risk prediction models.

Conflict of interest: M. Sadatsafavi reports grants and personal fees from Boehringer Ingelheim and AstraZeneca, personal fees from GlaxoSmithKline and Teva, outside the submitted work. A. Adibi has nothing to disclose. M. Puhan has nothing to disclose. A. Gershon has nothing to disclose. S.D. Aaron has nothing to disclose. D.D. Sin reports grants and personal fees for advisory board and educational work from AstraZeneca, personal fees for lectures from Boehringer Ingelheim, personal fees for advisory board work from Grifols, outside the submitted work.

Support statement: This work was supported by the BC SUPPORT Unit Methods Cluster Project Award, grant HESM 205. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Bellou V, Belbasis L, Konstantinidis AK, *et al.* Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. *BMJ* 2019; 367: l5358.

2 Bridge J, Blakey JD, Bonnett LJ. A systematic review of methodology used in the development of prediction models for future asthma exacerbation. *BMC Med Res Methodol* 2020; 20: 22.

3 Van Calster B, McLernon DJ, van Smeden M, *et al.* Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019; 17: 230.
4 Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; 115: 928–935.
5 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; 26: 565–574.
6 Singh D, Agusti A, Anzueto A, *et al.* Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease: the GOLD science committee report 2019. *Eur Respir J* 2019; 53: 1900164.
7 Adibi A, Sin DD, Safari A, *et al.* The Acute COPD Exacerbation Prediction Tool (ACCEPT): a modelling study. *Lancet Respir Med* 2020; 8: 1013–1021.
8 Vestbo J, Anderson W, Coxson HO, *et al.* Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *Eur Respir J* 2008; 31: 869–873.
9 Taylor SP, Sellers E, Taylor BT. Azithromycin for the prevention of COPD exacerbations: the good, bad, and ugly. *Am J Med* 2015; 128: 1362.e1–1362.e6.
10 Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019; 3: 18.
11 Sande SZ, Li J, D'Agostino R, *et al.* Statistical inference for decision curve analysis, with applications to cataract diagnosis. *Stat Med* 2020; 39: 2980–3002.
12 Kerr KF, Marsh TL, Janes H. The importance of uncertainty and opt-in v. opt-out: best practices for decision curve analysis. *Stat Med* 2019; 39: 491–492.
13 Puhan MA, Hansel NN, Sobradillo P, *et al.* Large-scale international validation of the ADO index in subjects with COPD: an individual subject data analysis of 10 cohorts. *BMJ Open* 2012; 2: e002152.
14 Xu L, Su H, She Y, *et al.* Which N descriptor is more predictive of prognosis in resected non-small cell lung cancer: the number of involved nodal stations or the location-based pathological N stage? *Chest* 2021; 159: 2458–2469.
15 Sadatsafavi M, McCormack J, Petkau J, *et al.* Should the number of acute exacerbations in the previous year be used to guide treatments in COPD? *Eur Respir J* 2021; 57: 2002122.