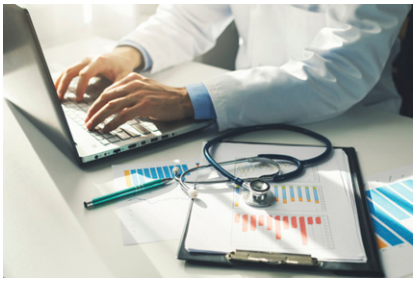
Methodologic Guidance and Expectations for the Development and Reporting of Prediction Models and Causal Inference Studies

Michael O. Harhay, Ph.D.[1,2*], David H. Au, M.D., M.S.[3,4*], Sharon D. Dell, M.D.[5*], Michael K. Gould, M.D., M.S.[6*], Susan Redline, M.D.[7,8*], Christopher J. Ryerson, M.D.[9,10*], and Colin R. Cooke, M.D., M.S.[11*]

[1]Palliative and Advanced Illness Research (PAIR) Center and [2]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania; [3]Health Services Research and Development Service, Center of Innovation for Veteran-Centered and Value-Driven Care, Veterans Affairs Puget Sound Health Care System, Seattle, Washington; [4]Division of Pulmonary, Critical Care, and Sleep Medicine, University of Washington, Seattle, Washington; [5]Department of Pediatrics, Hospital for Sick Children, Toronto, Ontario, Canada; [6]Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, California; [7]Beth Israel Deaconess Medical Center, and [8]Division of Sleep and Circadian Disorders, Departments of Medicine and Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; [9]Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada; and [10]Centre for Heart Lung Innovation, St. Paul's Hospital, Vancouver, British Columbia, Canada; and [11]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan

ORCID IDs: 0000-0002-0553-674X (M.O.H.); 0000-0001-8207-4975 (D.H.A.); 0000-0003-2169-9407 (S.D.D.); 0000-0001-6749-8315 (M.K.G.); 0000-0001-9713-5371 (C.R.C.).



The May 2020 issue of *Critical Care Medicine* includes a new consensus document for developing, validating, updating, and reporting prediction models (1). This document was co-created and is co-signed by statistical editors, associate editors, and editors-in-chief at 31 pulmonary, sleep, and critical care journals, including *AnnalsATS*. This prediction model guidance document is the second guidance document produced by this large group of editors following the guidance on causal inference studies published in *AnnalsATS* in January 2019 (2).

Authors submitting their work to *AnnalsATS* will often be directed to these (and potentially future) guidance documents, as well as the *AnnalsATS* detailed instructions to authors. Our goal in the present editorial is to provide the rationale and vision for how these documents and the recommendations within them can best be used by authors and reviewers.

What Kind of Studies Should Use the Guidance Documents?

AnnalsATS receives a diverse collection of manuscripts that ask a wide variety of scientific questions. Many studies are motivated by causal questions (3, 4). Some examples of such questions include: Does air pollution cause asthma? Does adherence to guidelines improve patient outcomes? Are outcomes better with a double-lung than with a single-lung transplant?

The goal of the first guidance document was to provide an accessible contemporary summary and reference guide for authors to use to explore such questions using causal inference methods. Causal inference methods offer powerful and recommended conceptual and empirical tools to design studies, develop and refine statistical models, and estimate and report effect estimates (5, 6). Though causal inference methods can be used to improve the design and analysis of randomized trials, they are especially useful in guiding observational studies that seek to examine relationships between an exposure and an outcome using nonrandomized data sources.

In contrast to causal inference studies, prediction modeling studies aim to develop, validate, or update a mathematical equation that calculates a specific probability or risk of a condition or future event for an individual (7). To clarify how these two study designs differ, consider the setting of lung cancer. An observational causal inference study might seek to provide an estimate of the average increase in the risk of lung cancer for each year of smoking among subjects who smoke compared with subjects without a smoking history. That is, it seeks to estimate a relationship between an exposure and an outcome, such that we can consider what would have happened if a patient had smoked less or not at all. In hopes of achieving an informative effect estimate while minimizing bias, a researcher would use causal inference methods to identify key confounders and the appropriate statistical model to generate an effect estimate for the association of

*M.O.H. is Statistical Editor; D.H.A., S.D.D., M.K.G., S.R., and C.J.R. are Deputy Editors; and C.R.C. is Editor-in-Chief of *AnnalsATS*. Their participation complies with American Thoracic Society requirements for recusal from review and decisions for authored works.

Supported by National Institutes of Health grant R00 HL141678 (M.O.H.).

DOI: 10.1513/AnnalsATS.202002-141ED

smoking history with lung cancer. For a prediction model, a potential goal would be to assign an individual probability of developing lung cancer for each patient at a specific time in the future. Though the equation developed to accomplish this "prediction" goal may include variables identified from causal inference studies, such as smoking history, it does not have to include such variables, nor are effect estimates and confounding of primary interest. The primary goal of a prediction model is instead to identify the combination of variables that most accurately predicts individual outcomes, regardless of whether these variables fall along causal pathways. Prediction models familiar to readers might include the Acute Physiology and Chronic Health Evaluation IV score (8) to predict mortality at the time of intensive care unit admission or the Framingham Risk Score for general cardiovascular disease risk (9). The goal of this new document published in *Critical Care Medicine* (1) was to provide a unique and accessible summary of the statistical literature on the best practices for developing, validating, and reporting prediction model studies.

Why Have These Documents Been Produced?

Prediction modeling and causal inference studies are abundant in the medical literature. Although these have long been accepted research pursuits, these study designs are also frequently undertaken using methods and approaches that are both prone to bias and no longer recommended by the statistical community. Accordingly, there is logic in assembling guidance documents that summarize the leading perspectives and recommended approaches for these study designs that can be used by authors, reviewers, and readers alike. Doing so across journals provides several benefits. First, there is a general desire among editorial teams to provide greater clarity to authors regarding the statistical analysis expectations at their respective journals. Second, promoting the use of accepted and preferred methodology helps elevate the rigor and quality of research in our respective disciplines. Third, copromotion across journals helps authors design studies and draft manuscripts that are broadly acceptable for peer review at multiple venues. Finally, communicating a common set of criteria for evaluating the methodological rigor of statistical analysis approaches helps reviewers and editors avoid redundancy by providing generalized responses to common issues such as underreporting and methodological concerns. This improves the efficiency of the editorial process for everyone.

How Should Authors and Reviewers Approach These Documents?

First, we want to emphasize that these guidance documents are not a set of prescriptive rules that must be followed without deviation. Furthermore, they are not a simple recipe that, if followed, will lead to publication. They should be interpreted as strong guidance representing contemporary views and consensus recommendations on best practices for common study types, and not as immutable editorial policy. A principal goal of these documents was to push authors toward greater conceptual and reporting clarity in their manuscript submissions. Authors retain their full discretion to pursue the study design they deem most appropriate for their research questions. We invite authors to innovate; yet, when they do, they should justify the validity of their approach. Our goal is to improve the overall quality of research; it is not to stifle development or use of novel and innovative approaches.

Table 1. Key reporting metrics for prediction models

Domain	Key Reporting Elements
Data source	Were data collected prospectively for this purpose or repurposed from an archival dataset? Wherever possible, the data used should be made available to readers.
Participants	Which patients were included in the study? Were separate populations used for model derivation and validation? How many patients were included in each of these groups? A "Table 1" describing relevant clinical features is useful.
Outcome	Specific details on how the outcome was defined.
Predictors	A specific accounting of the predictor variables included in the final model, along with the method by which these variables were selected.
Missing data	How much data were missing from the predictors and from the outcome? How was missing data handled?
Model specification	What sort of model was used (e.g., linear regression, random forest)? The final model itself should be reported with as much detail as possible, including specific equations/parameters. Whenever possible (particularly in the case of machine learning models), the code used should be provided in full such that others can replicate the analyses.
Model structure	The full model equation should be reported when applicable (e.g., statistical models), along with equations required to interpret results (e.g., the baseline hazard function in a time-to-event model).
Validation	How was the model validated (internal vs. external)? If internal validation only was performed, how was the dataset split?
Model performance	Performance measures should be tailored to the intended purpose of the model but generally should include a measure of discrimination (e.g., AUROC or AUPRC), a measure of calibration (e.g., Hosmer-Lemeshow, scaled Brier score), and clinically relevant performance (e.g., PPV, NPV) as indicated.

Definition of abbreviations: AUPRC = area under the precision recall curve; AUROC = area under the receiver operating characteristic curve; NPV = negative predictive value; PPV = positive predictive value. Reprinted by permission from Reference 1, adapted from the (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) checklist (7).

Table 2. Summary of guidance for prediction models

Recommended Practices	Cautions
Consider competing priorities of precision, parsimony, and transparency when approaching a prediction task.	Prediction frameworks should not be used to make causal inferences.
Think carefully about the prediction's intended purpose and prioritize feature selection elements as appropriate.	Using *P* values from bivariable comparisons or stepwise procedures to select predictors leads to bias and overfitting.
Report the prevalence and handling of missing data; consider steps other than case exclusion to address missing data.	The size of a dataset, as well as the number of outcomes it contains, limit the number of predictor variables that the model can accommodate.
Consider the expected nature of the relationships between predictors and the outcome (e.g., linear, exponential, etc.).	Categorizing continuous variables can lead to loss of information.
Conduct external validation to demonstrate a model can generalize to new observations.	External validation should use the same model used to report the internal performance; avoid retraining on the external dataset.
Seek reasonable comparators other than "no model" when evaluating model performance.	Relying on the area under receiver operator characteristics curve alone can lead to an incomplete understanding of a model's performance.
Follow appropriate reporting guidelines such as TRIPOD and RECORD.	

Definition of abbreviations: RECORD = Reporting of Studies Conducted Using Observational Routinely-collected Data; TRIPOD = Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis.
Reprinted by permission from Reference 1.

What Is Contained in the New Prediction Recommendations?

For ease, we have reproduced two tables that summarize key elements to consider reporting in prediction model manuscripts (Table 1) and recommended prediction model practices (Table 2). We highlight a few specific topics that the editorial team at *AnnalsATS* will be keenly assessing as we move forward.

First, the guidance builds heavily on the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) checklists for prediction model development and validation (www.equator-network.org/reporting-guidelines/tripod-statement/). Since 2015, many leading journals have required adherence to the TRIPOD checklists. Submissions to *AnnalsATS* for all prediction model studies should have completed TRIPOD checklists appropriate to the study.

Second, we are interested in receiving work that goes beyond the development and internal validation of prediction tools. Articles validating a prediction tool in novel cohorts or articles showing that a prediction tool leads to changes in practice, clinical decision making, or outcomes will be prioritized. Although model development and validation for important topics is essential work that we will continue to publish, we challenge authors to send us work with real-world benefit, not just work with higher c-statistics. The medical literature is full of prediction models that perform well but are never used or fail to change clinical care. Rather than expand the number of unused models, *AnnalsATS* aspires to publish studies that prioritize added value to patient care.

Third, we are particularly interested in the topic of prediction model "bias" or "fairness" (10). Prediction models that originate from datasets that lack representation of demographic groups traditionally overlooked in biomedical research may lead to bias in real-world applications (11, 12). As a result, these prediction models may create or reinforce biases and health disparities (11–16). For example, inclusion of black race in a prediction model could suggest that being black is associated with poor outcomes in patients with chronic obstructive pulmonary disease or low continuous positive airway pressure therapy adherence in patients with sleep apnea. However, if the poor outcomes observed in the dataset used to generate the prediction model reflect surmountable access to care issues or the outcomes of a small total number of black patients (i.e., unrelated to ancestry-linked biological factors), then perpetuating this effect in a published prediction model could negatively impact future care decisions for black patients based on this model. As a general recommendation, inclusion of such factors in prediction models should be carefully considered and justified by what information they add in the context of specific study questions (17).

In closing, we reiterate that these guidance documents are better thought of as strong recommendations and not as unwavering rules and requirements for publication. Our intent is to help authors improve the rigor of their studies rather than to discourage submission. Ultimately, each submitted manuscript is evaluated on its own merits. Innovations in statistical methods, the allure of machine learning, and expanding data size and richness have led to an increase in the number of submissions related to prediction modeling as well as observational research studies. Our goal is very simply to provide tools that promote high-quality contributions to medical research in these areas. We look forward to reviewing your contributions. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

Acknowledgment: The authors thank Drs. Dan Leisman and Edward Kennedy for their comments on an earlier draft of the manuscript.

References

1 Leisman DE, Harhay MO, Lederer DJ, Abramson M, Adjei AA, Bakker J, *et al*. Development and reporting of prediction models: guidance for authors from editors of respiratory, sleep, and critical care journals. *Crit Care Med* 2020;48:623–633.

2 Lederer DJ, Bell SC, Branson RD, Chalmers JD, Marshall R, Maslove DM, *et al*. Control of confounding and reporting of results in causal inference studies: guidance for authors from editors of respiratory, sleep, and critical care journals. *Ann Am Thorac Soc* 2019;16:22–28.

3 Hernán M. The C-word: the more we discuss it, the less dirty it sounds [letter]. *Am J Public Health* 2018;108:625–626.
4 Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. *Am J Public Health* 2018;108: 616–619.
5 Hernán MA, Robins JM. Causal inference: what if. Boca Raton, FL: Chapman & Hall/CRC; 2020.
6 Pearl J, Mackenzie D. The book of why: the new science of cause and effect. New York: Basic Books; 2018.
7 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* 2015; 162:55–63.
8 Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006;34: 1297–1310.
9 D'Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, *et al*. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008;117:743–753.
10 Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA* 2019;322:2377–2378.
11 Goodman SN, Goel S, Cullen MR. Machine learning, health disparities, and causal reasoning [editorial]. *Ann Intern Med* 2018;169: 883–884.
12 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–453.
13 Braun L. Race, ethnicity and lung function: a brief history. *Can J Respir Ther* 2015;51:99–101.
14 Eneanya ND, Yang W, Reese PP. Reconsidering the consequences of using race to estimate kidney function. *JAMA* 2019;322: 113–114.
15 Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018;178:1544–1547.
16 Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–872.
17 Kaplan JB, Bennett T. Use of race and ethnicity in biomedical publication. *JAMA* 2003;289:2709–2716.

Prognosis for Mechanically Ventilated Patients: A Moving Target

Sandra E. Zaeh, M.D., M.S.[1], and Anuj B. Mehta, M.D.[2,3]

[1]Division of Pulmonary and Critical Care Medicine, Johns Hopkins University, Baltimore, Maryland; [2]Division of Pulmonary, Critical Care, and Sleep Medicine, Department of Medicine, National Jewish Health, Denver, Colorado; and [3]Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine, University of Colorado School of Medicine, Aurora, Colorado

Invasive mechanical ventilation remains fundamental to the management of critically ill patients in the ICU. As the worldwide population ages and develops an increasing number of medical comorbidities, rates of mechanical ventilation are also rising (1, 2).

Within this context, prognostic information regarding mechanically ventilated patients is increasingly important for patients and their surrogate decision-makers. In the ICU, shared decision-making, or medical decisions made through a partnership among physicians, patients, and their loved ones, is the recommended standard (3). However, patients, their surrogates, and physicians often have different expectations regarding prognosis, with prior data showing a >50% discordance between surrogates and physicians (4). A key pillar of shared decision-making is being able to provide patients and their surrogates with reliable expectations. Previous studies have attempted to predict mechanical ventilation outcomes at specific time points (Day 1, Day 14, and Day 21) with the assumption of static prognoses (5–7). However, patients receiving mechanical ventilation change from day to day and static prognoses at predetermined time points may not be appropriate for an ever-changing population.

In this issue of *AnnalsATS*, Ruan and colleagues (pp. 729–735) used data from 162,200 episodes of respiratory failure included in Taiwan's National Health insurance database to investigate dynamic changes in mechanical ventilation prognoses based on each additional day of mechanical ventilation needed (8). The authors identified adults who received mechanical ventilation for two consecutive days, and calculated the cumulative probabilities of weaning success and death in the subsequent 90 days. Their results showed that >90% of successful weaning occurred in the initial 30 days after mechanical ventilation, with a decreasing trend over time. In contrast, deaths initially increased after mechanical ventilation, but then decreased after the 19th day on the ventilator, with the probability of death surpassing the probability of weaning success on the 28th ventilator day. The authors' findings were consistent across multiple subgroups.

Based on their results, the authors created an online inquiry system to provide tailored prognostic information based on ventilator day, age, and sex (http://mvp.nhri.org.tw/NHIA-NHRI2017/count.html). They believe that this information may provide patients and surrogates with more dynamic information regarding evolving prognoses that may impact decision-making in the ICU.

DOI: 10.1513/AnnalsATS.202003-242ED