OPEN

Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals

Daniel E. Leisman, BS[1]; Michael O. Harhay, PhD, MPH[2]; David J. Lederer, MD, MS[3,4]; Michael Abramson, MBBS, PhD[5]; Alex A. Adjei, MD, PhD[6]; Jan Bakker, MD, PhD, FCCM, FCCP[7]; Zuhair K. Ballas, MD[8]; Esther Barreiro, MD, PhD[9]; Scott C. Bell, MBBS, MD, FRACP[10]; Rinaldo Bellomo, MD, PhD[11]; Jonathan A. Bernstein, MD[12]; Richard D. Branson, MSc, RRT, FAARC, FCCM[13]; Vito Brusasco, MD[14]; James D. Chalmers, MD, PhD[15]; Sudhansu Chokroverty, MD, FRCP[16]; Giuseppe Citerio, MD[17]; Nancy A. Collop, MD[18]; Colin R. Cooke, MD, MS[19]; James D. Crapo, MD[20]; Gavin Donaldson, PhD[21]; Dominic A. Fitzgerald, MBBS, PhD, FRACP[22]; Emma Grainger, PhD[23]; Lauren Hale, PhD[24]; Felix J. Herth, MD, PhD[25]; Patrick M. Kochanek, MD[26]; Guy Marks, MBBS, PhD[27]; J. Randall Moorman, MD[28]; David E. Ost, MD, MPH[29]; Michael Schatz, MD, MS[30]; Aziz Sheikh, MD, MSc[31]; Alan R. Smyth, MA, MBBS MRCP, MD, FRCPCH[32]; Iain Stewart, PhD[33]; Paul W. Stewart, PhD[34]; Erik R. Swenson, MD[35]; Ronald Szymusiak, PhD[36]; Jean-Louis Teboul, MD, PhD[37]; Jean-Louis Vincent, MD, PhD[38]; Jadwiga A. Wedzicha, MD[39]; David M. Maslove, MD, MS[40]

[1]Icahn School of Medicine at Mount Sinai, New York, NY.

[2]Palliative and Advanced Illness Research (PAIR) Center and Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

[3]Regeneron Pharmaceutical, Inc., Tarrytown, NY.

[4]Columbia University Medical Center, New York, NY.

[5]Public Health and Preventive Medicine, Monash University, Melbourne, VIC, Australia.

[6]Department of Oncology, Mayo Clinic, Rochester, MN.

[7]Department of Pulmonology, Sleep Medicine, and Critical Care, New York University, New York, NY.

[8]Department of Internal Medicine, University of Iowa, Iowa City, IA.

[9]Research Institute of Hospital del Mar, Barcelona, Spain.

[10]Department of Thoracic Medicine, The Prince Charles Hospital, Brisbane, QLD, Australia.

[11]Department of Intensive Care Medicine, Austin Hospital and University of Melbourne, Melbourne, QLD, Australia.

[12]Department of Internal Medicine, University of Cincinnati College of Medicine, Cincinnati, OH.

[13]Department of Surgery, University of Cincinnati, Cincinnati, OH.

[14]Department of Internal Medicine, University of Genoa, Genoa, Italy.

DOI: 10.1097/CCM.0000000000004246

[15]University of Dundee, Dundee, Scotland.

[16]JFK Neuroscience Institute, Hackensack Meridian Health–JFK Medical Center, Hackensack, NJ.

[17]School of Medicine and Surgery, Università Milano Bicocca, Monza, Italy.

[18]Departments of Medicine and Neurology, Emory University School of Medicine, Atlanta, GA.

[19]Department of Medicine, University of Michigan, Ann Arbor, MI.

[20]Department of Medicine, National Jewish Hospital, Denver, CO.

[21]Asthma and COPD Group, National Heart and Lung Institute, Imperial College London, London, United Kingdom.

[22]The Children's Hospital at Westmead, Sydney Medical School, University of Sydney, Sydney, NSW, Australia.

[23]The Lancet, London, United Kingdom.

[24]Department of Population and Preventive Medicine, Stony Brook Medicine, Stony Brook, NY.

[25]Department of Pneumology and Critical Care Medicine, University of Heidelberg, Heidelberg, Germany.

[26]Department of Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA.

[27]Department of Medicine, University of Sydney School of Medicine, Sydney, NSW, Australia.

[28]Departments of Medicine, Physiology, Engineering, University of Virginia, Charlottesville, VA.

[29]Department of Pulmonary Medicine, University of Texas MD Anderson Cancer Center, Houston, TX.

[30]Department of Allergy, Kaiser Permanente Medical Center, San Diego, CA.

[31]Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, United Kingdom.

[32]Division of Child Health, Obstetrics & Gynecology, University of Nottingham, Nottingham, United Kingdom.

[33]Division of Respiratory Medicine, University of Nottingham, Nottingham, United Kingdom.

[34]Department of Biostatistics, University of North Carolina, Chapel Hill, NC.

[35]Department of Medicine, University of Washington, Seattle, WA.

[36]Departments of Medicine and Neurobiology, David Geffen School of Medicine at UCLA, Los Angeles, CA.

[37]CHU Bicêtre, Le Kremlin-Bicêtre, France.

[38]Université Libre de Bruxelles, Brussels, Belgium.

[39]National Heart and Lung Institute, Imperial College London, London, United Kingdom.

[40]Department of Critical Care Medicine, Queen's University, Kingston, ON, Canada.

Abstract: Prediction models aim to use available data to predict a health state or outcome that has not yet been observed. Prediction is primarily relevant to clinical practice, but is also used in research, and administration. While prediction modeling involves estimating the relationship between patient factors and outcomes, it is distinct from casual inference. Prediction modeling thus requires unique considerations for development, validation, and updating. This document represents an effort from editors at 31 respiratory, sleep, and critical care medicine journals to consolidate contemporary best practices and recommendations related to prediction study design, conduct, and reporting. Herein, we address issues commonly encountered in submissions to our various journals. Key topics include considerations for selecting predictor variables, operationalizing variables, dealing with missing data, the importance of appropriate validation, model performance measures and their interpretation, and good reporting practices. Supplemental discussion covers emerging topics such as model fairness, competing risks, pitfalls of "modifiable risk factors", measurement error, and risk for bias. This guidance is not meant to be overly prescriptive; we acknowledge that every study is different, and no set of rules will fit all cases. Additional best practices can be found in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines, to which we refer readers for further details. (Crit Care Med 2020; 48:623–633)

Key Words: critical care; pulmonary medicine; prediction models; sleep medicine

Prediction is the bedrock of clinical practice. Inherent in every diagnosis is a prediction about the course of illness, and every prescription invokes a prediction about a response to treatment. For the most part, clinical predictions are made on a case-by-case basis based on a combination of experience and evidence. More recently, the uptake of electronic health records (EHRs), adoption of genomics technologies, and the advent of data science and machine learning have accelerated the development and publication of data-driven prediction models throughout medicine. Respiratory, sleep, and critical care medicine are no exception; prediction modeling has strong foundations in these fields, and they continue to be influential in its refinement and uptake.

Journals are witnessing an increase in submissions related to prediction modeling. This stands to seed rapid advancement in research and practice but also comes at the risk of pursuing false leads. As statistical editors, associate editors, and editors-in-chief at leading pulmonary, sleep, and critical care journals (**Appendix 1**), we believe it is important to provide guidance on how to maximize the usefulness of prediction modeling to capitalize on the opportunities that modern statistics and data science afford our fields.

INTENDED PURPOSE

This document is intended for both readers and authors of studies that describe prediction models. It borrows from expert reviews (1–5) and the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines, a comprehensive set of recommendations on current best practices for publishing prediction models, to which we direct readers for more information (6, 7).

Our aim is to provide an accessible summary of best practices and recommendations on prediction modeling, rather than to prescribe editorial policy for participating journals. We hope this guidance will enhance the overall quality and scientific merit of submissions to our various journals, provide consistency and common ground, and support readers of these studies in their critical appraisal of the prediction literature. Although exhaustive discussion of all the salient aspects of prediction modeling are beyond our scope, we attempt to address specific issues that we most commonly encounter in the review process at our various journals. We also acknowledge that every study is different, and depending on the overall goals, some may not conform to the guidance herein. We further recognize that this guidance may require updates as the fast-moving field of prediction modeling evolves over time.

DEFINITIONS AND SCOPE

The goal of prediction is to use information currently available to forecast a future outcome (**Fig. 1**). A prediction model is any construct that uses known variables (often called independent variables, features, or inputs) to estimate the value of this outcome (often called the dependent variable, response, or output) before it is observed. This is distinct from causal inference modeling, which aims to determine how a dependent variable will change as a direct result of altering an independent variable (often called an exposure). Causal inference studies, which we discuss in more detail elsewhere (8), require careful consideration of confounding and other potential biases.

Some prediction models may include causal factors (e.g., smoking is both predictive and causative of lung cancer), but in the strictest sense, such causal relationships are not required (e.g., a rising creatinine may predict impending renal failure, but does not itself cause renal failure). Therefore, prediction models are developed using different methods and should not be used for drawing causal inferences. Doing so can lead to logical fallacies. For example, a study may find endotracheal intubation predicts mortality, but this certainly does not mean patients with a compromised airway should not be intubated. In fact, sometimes the

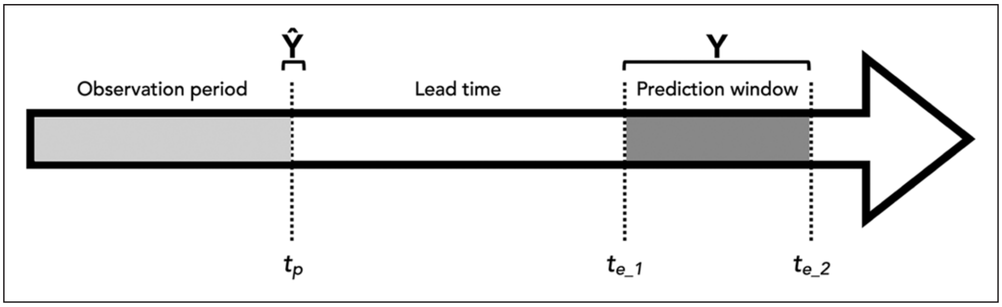

Figure 1. The chronology of information is critical in prediction. A prediction, $\hat{Y}$, is made at time, t_p, based on data collected up to and including that time, but no later. $\hat{Y}$ is the estimate of Y, which cannot be observed until a time in the future. The times at which we can observe Y (rather than just $\hat{Y}$) fall within a prediction window (t_{e_1} to t_{e_2}), which occurs after a certain amount of lead time has elapsed. The width of observation, lead time, and prediction intervals will influence the usefulness of any prediction model.

best predictors are interventions that counteract a causal process. Removing causal expectations means fewer restrictions on the variables a model can include, so long as the goal is properly understood to be an accurate and generalizable prediction, and not a deeper understanding of its biological significance (9).

A useful prediction model should, therefore, satisfy three core criteria:

1) It must provide a model whereby known variables estimate the value of the event of interest (e.g., for a binary outcome, it must have a classifier function),
2) The predictors must be known prior to knowing the outcome state, and
3) The model should retain accuracy when applied to new observations (i.e. it must be generalizable).

Some prediction models allow for clear identification of key predictors. Others may obscure the precise factors on which they most heavily rely, as well as the precise mechanisms by which they arrive at predictions. Although this is obviously of concern in causal inference exercises, it is sometimes less important in prediction modeling exercises, so long as the explanation underlying the prediction is felt to be unimportant.

Another important consideration is the intended purpose of any given prediction model. This generally falls into one of two categories. First are clinical prediction models intended for bedside use to inform the care of individual patients. For example, a rapid shallow breathing index calculated during a spontaneous breathing trial may predict whether a mechanically ventilated patient will be successfully extubated. Second are system prediction models intended for deployment across populations for research, benchmarking, and other administrative purposes. For example, an Acute Physiology and Chronic Health Evaluation IV score applied to a particular ICU cohort compares its overall predicted mortality risk with that of another cohort (10). The first type of predictive model is typically applied prospectively to forecast individual events, whereas the latter is typically deployed to characterize an overall population. Nonetheless, both should respect the temporal requirements necessary for an unbiased prediction. Though related in methodology, these two types of prediction models differ in how they should be evaluated and reported. Specifically, applying a model in clinical practice may require higher precision, making characteristics like positive predictive value more relevant than overall measures of discrimination, such as the area under the receiver operator characteristics curve (AUROC).

MODEL ARCHITECTURE

Considering Potential Predictor Variables

How researchers decide on variables to include in a prediction model (also referred to as "features") is equally if not more important than the specific variables themselves. The large datasets increasingly used in biomedical research, such as those captured from EHRs, administrative systems, and high-dimensional "omics" platforms, include features that may number in the hundreds to thousands.

Not all features available in modern datasets are practical or effective choices. There is a trade-off between the number of features included in a model and its capacity to generalize. This risk becomes particularly important when considering that associations between predictor and outcome may be idiosyncratic. A single center study may show bronchoscopies done on Tuesdays are predictive of lung cancer diagnosis, but this may simply reflect local practice in which the lung nodule clinic has access to the endoscopy suite on certain days.

As noted above, for clinical prediction modeling the only allowable predictor variables are those that will be known at the time the prediction is made. Consider a model to predict whether a chronic obstructive pulmonary disease (COPD) patient with pneumonia and respiratory failure in the emergency department will subsequently develop hypotension. Positive blood cultures may be highly predictive of this outcome, but blood cultures typically take hours to days to be reported. This variable cannot be included in a model intended to assist decision making for a patient's disposition from the emergency department because the information would never be available to decision-makers at the time they will use the prediction.

Authors should also consider how readily a variable can be obtained—including the cost, invasiveness, and risk of obtaining it—as well as how ubiquitously it is encountered in routine practice. The growing pervasiveness of smartphones and other devices, which can calculate complex scores automatically, has lessened the premium on "simple" scores with few variables, but clear trade-offs between parsimony and accuracy remain. For example, a prediction model for lung cancer that uses smoking history alone may be easy to use in any setting but may underperform compared with one based on smoking history, an exhaustive occupational history, and whole genome sequencing.

Modern datasets may lack representation from demographic groups historically under-represented in biomedical research. Prediction models based on such datasets may lead to bias in real world applications (11). Relatedly, including dimensions such as race as predictors inherently reflects assumptions of difference that are often tenuous, particularly when modeling a physiologic response. The complex issues and history surrounding prediction models creating or reinforcing biases are discussed elsewhere in-depth (11–15). Generally, we caution against arbitrarily including these variables in prediction models. Instead, we suggest careful consideration of what information they add in the context of specific study questions. We discourage their inclusion without reasonable suspicion that they contribute important predictive information.

Procedures for Predictor Selection

Usually, we approach prediction problems with prior knowledge about what features are likely to be predictive. In these cases, candidate variables can (and usually should) be preselected based on theory or prior evidence. The converse approach, selecting variables solely on strength of association, leads to problems. For example, suppose we select variables to predict the risk of future intubation. Lactate level may not initially appear associated with intubation in a population that includes many patients with asthma and COPD in whom lactate may be elevated on the basis of high-dose inhaled β-agonist therapy. Lactate may in fact be highly associated with intubation in patients presenting with pulmonary sepsis and shock. Here, β-agonists are effect modifiers: the predictive relationship between lactate and intubation depends on β-agonist use. This example is one of many reasons why experts have long recommended against using bivariable association to guide feature selection (16).

Another commonly used but problematic methodology is "stepwise" selection (17). Stepwise selection refers to procedures where decisions to include a predictor are based solely on *p* values associated with that predictor throughout multiple iterations of a model. For example, a study that enters 20 candidate variables into a model and continuously removes those meeting a threshold (e.g., $p > 0.10$ or >0.05) until all remaining terms are "statistically significant" uses backward-selection. By deciding to include variables based on *p* values, stepwise selection essentially amounts to multiple comparisons without appropriate correction. Further, unlike prespecified models, where all potential relationships are grounded in plausibility, stepwise models have no "prior" about what to include, making them highly prone to overfitting. Overfitting occurs when a model contains too many variables for the dataset to support. This results in a close fit to the data on which it was trained, but poor generalization to other datasets (18), undermining a key criterion of useful prediction. Any process that rigidly adheres to *p* value thresholds for variable selection poses similar risks of spurious conclusions (19, 20). In the vast majority of scenarios, *p* value based feature selection methods are strongly discouraged (21).

Underscoring the rationale to eschew stepwise and *p* value screening, modern statistical approaches to feature selection (e.g., penalized regression) avoid these pitfalls and can also inherently improve overall accuracy (22). For these reasons, TRIPOD guidelines explicitly recommend these alternative procedures for prediction studies (6). These methods have their own caveats and are comparatively more complex. Accordingly, they may not be amenable to all studies. Ultimately, the balance of prior knowledge about which factors are likely predictive and the need for data-driven discovery of novel predictors should guide a specific study's approach to feature selection. Fully prespecified theory-based feature selection may be appropriate when there is extensive prior knowledge, whereas penalization methods may be preferred when prior knowledge is lacking or in discovery exercises. **Table 1** presents an overview of feature selection techniques for statistical prediction models.

Finally, we note that the number of outcome events, not simply sample size, influences how many predictor terms can be included without overfitting. Large models (e.g., 50+ predictors) are often imposed on smaller datasets (e.g., $n = 300$, with 30 outcomes) that cannot support them. Methods to determine how many predictors a model can accommodate are discussed elsewhere, and we encourage authors to consider what their datasets can actually support. For sample size guidance, we direct readers to methodological papers based on the type of outcome being predicted—continuous (23, 24), binary (25, 26), or time-to-event (25)—and recommend authors clearly indicate how they determined whether their dataset could support the chosen model.

MODEL CONSTRUCTION

In addition to identifying predictor variables, prediction model development involves decisions about how to operationalize the predictors, define outcomes, handle missing data, and select a method to generate predictions.

Operationalizing Predictor Variables

One practice often employed in preprocessing data is to split continuous variables into dichotomous ones (27, 28). This practice risks discarding information and replacing it with assumptions that rarely have biological plausibility. For example, we might want to include respiratory rate (RR) in a model predicting the need for positive pressure ventilation. Consider four patients with RRs of 12, 29, 31, and 40, respectively. A logical interpretation is that the first patient has the lowest (RR-attributable) risk, the second and third patients have higher but similar risk, and the fourth has the highest risk. Suppose we instead split the population into two groups: RR greater than or equal to 30 and less than 30. This makes several illogical assumptions and discards useful predictive information, as described in **Figure 2**.

Investigators sometimes create categorical groupings because they suspect a nonlinear relationship (e.g., a threshold rather than a dose-response effect). Although using nonlinear terms in a prediction equation might better accommodate this relationship than categorizing continuous data, using a linear term to model a suspected "J-shaped" curve would also be inappropriate. These issues should underscore the importance of thinking about predictor relationships ahead of time when feasible.

Although avoiding categorization is generally preferred, thresholds may, at times, be useful to formulate prediction scores that can be easily calculated at the bedside. As with

TABLE 1. Approaches to Feature Selection in Prediction Models

Strategy	Mechanics	Pros	Cons
Traditional methods			
Full prespecification	Before constructing a model, authors prespecify exactly which predictors and interactions will be included. All terms retained regardless of relationship to the outcome.	All predictors grounded in plausibility (strong "priors"). Less vulnerable to overfitting and dataset-specific idiosyncrasies. Computationally simple.	Requires reliable prior knowledge about what factors are predictive. Ideal theoretical model may exceed number of dimensions the dataset can reasonably support. Models may be less accurate than those employing appropriate penalization or shrinkage methods.
Penalization methods			
LASSO (L1) regularization	Imposes a "penalty" to bias predictor coefficients toward zero. The penalty is a function of the coefficient's absolute magnitude and a weight, λ.	Can set coefficients to zero: useful to identify a small subset of most predictive features in datasets with many candidate predictors. Does not require prior knowledge about which features are predictive. If correctly weighted, can enhance prediction accuracy separate from use in feature selection.	May exclude features that are moderately predictive, discarding valuable information. May include implausible features and omit known predictive features. Individual predictors difficult to interpret. Performs poorly when predictors are highly correlated. Performs poorly when there are more predictors than observations. Requires internal cross-validation to estimate λ. λ unstable in smaller datasets. Computationally intensive.
Ridge (L2) regularization	Imposes a "penalty" to bias predictor coefficients toward zero. The penalty is a function of the coefficient's squared magnitude and a weight, λ.	Computationally less intensive vs LASSO/ Elastic-Net. Does not require prior knowledge about which features are predictive. If correctly weighted, can enhance prediction accuracy; outperforms LASSO if predictors are correlated.	Not a true means of predictor selection: penalization increasingly mild below 1.0 and unable to exclude features altogether. Individual predictors difficult to interpret. Requires internal cross-validation to estimate λ. λ unstable in smaller datasets.
Elastic-Net (Mixed L1-L2) regularization	Imposes a "penalty" to bias predictor coefficients toward zero. The penalty function is a mix of both LASSO and Ridge, with a weight between 0 (complete Ridge) and 1 (complete LASSO).	Can set coefficients to zero: can exclude nonpredictive features. Better handles correlated predictors vs LASSO. When more predictors than observations, better performance vs LASSO. If correctly weighted, can enhance prediction accuracy separate from use in feature selection. Does not require prior knowledge about which features are predictive.	May include implausible features and omit features already known to be predictive. Requires estimation of the optimal balance between L1/L2. Individual predictors difficult to interpret. Requires internal cross-validation to estimate λ. λ unstable in smaller datasets. Computationally intensive.

other choices in prediction modeling, this involves trade-offs between usability and accuracy. For example, imposing thresholds might be more reasonable if a score is intended as a simple and easily remembered tool but less so when deployed in an electronic risk calculator.

Specifying the Outcome

One of the most important steps in developing prediction models is to determine what precisely to predict. This sounds obvious, but confusion or misspecification here can lead to concerns about the model's applicability. For example, a model predicting the onset of sepsis should specify whether the labels "sepsis" and "non-sepsis" are assigned based on International Classification of Diseases (ICD) codes, Sequential Organ Failure Assessment scores derived from EHR data, expert opinion after manual chart reviews, or some other methodology. Transparency is important, especially when surrogate endpoints are used. For instance, a model might use a blood culture order and a prescription for IV antibiotics as a proxy for infection. Readers should recognize that the model therefore predicts the proxy state, rather than state of interest directly, and that overlap of the two may vary.

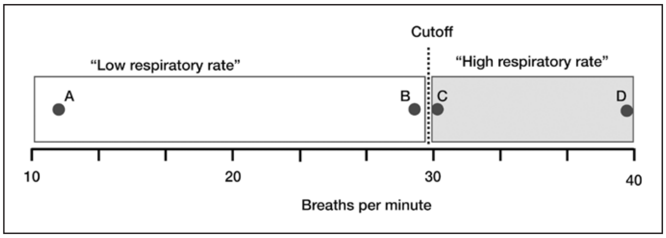

Figure 2. Problems with categorizing continuous variables. Consider the example of splitting respiratory rate (RR) values into "high" and "low" based on a cut-off of 30 breaths per minute. Note this makes several assumptions, namely: 1) that there is no difference between a RR of 12 and a RR of 29 (points *A* and *B*); 2) that there's no difference between a RR of 31 and a RR of 40 (points *C* and *D*); and 3) that RRs of 29 and 30 are categorically different (points *B* and *C*).

Authors should fully specify criteria used to adjudicate outcomes. Those criteria should consider ease of use, veracity, and consistency. For instance, ICD coding may be ubiquitous in some jurisdictions, but practices for coding any given condition may differ between institutions. Last, the timing of the outcome must be clearly specified. The outcome must be known only after all predictor variables have been collected, ideally with lag time that respects the realities of clinical practice. A model that predicts the onset of clinical deterioration 24 hours in advance is likely more useful than one that provides only 10 minutes of advanced warning.

Data Preprocessing

Missing data frequently complicate prediction modeling. This is especially true when models are built with existing datasets, rather than those collected specifically for the purpose of model development. The former, though convenient, may contain gaps where key variables were not routinely collected or reliably recorded. Missing data can introduce bias, especially if not missing completely at random. Suppose we want to predict bleeding risk during an inpatient stay, and choose to include the admission international normalized ratio (INR) as a predictor. Our dataset includes this measure for 65% of patients. Is the remaining 35% missing completely at random? Since clinical presentation likely dictates whether the INR is measured at admission, the mere presence of an INR value (regardless of its result) carries information (29). If we include only patients with an admission INR measured, we may select a group with higher bleeding risk; these patients are likely on anticoagulants, or deemed by the clinician at high enough risk that an INR was ordered. By omitting patients without an INR we may bias the prediction model.

Several strategies can address missing data. The approach using only patients with complete data (complete case analysis) can introduce bias and decreases sample size. This practice is therefore discouraged (30). Another strategy is multiple imputation of missing values using methods described elsewhere (31). The success of these approaches will depend on the methods used, the amount of missing data, and why the data are missing to begin with. Sometimes too much data are missing to support the proposed prediction modeling. Regardless of the methods chosen, the quantity of missing data, along with the methods used to deal with it, should be reported.

Other types of data preprocessing include the identification and removal of outliers, physiologically implausible values, or features that do not vary across patients and therefore contribute no information. The methods used in these preprocessing steps should be fully outlined (perhaps as a supplement), as they may introduce important biases in the ensuing models.

MODEL EVALUATION

Once developed, a prediction model must be evaluated to determine how useful it might be, and under what circumstances it might be used (32). This requires appropriate validation and quantification of model performance.

Model Validation

Evaluating a model's predictive performance can be helpful during derivation in the fine tuning of its variables. However, such evaluation does not constitute justification for the model's adoption. This is because models overly adapted to the idiosyncrasies of a particular dataset may perform well in that dataset, despite having poor accuracy for new observations (33). Many of the pitfalls described above will lead to overly optimistic performance models that fail to generalize.

Validation refers to the process of confirming whether a prediction model generalizes to data that were not used in its development. Internal validation involves determining whether model performance is reproducible in the same underlying population (as distinct from the same specific sample) used to derive it (33). External validation determines whether a model is transportable by evaluating its performance in a population that is somehow distinct from the one used for derivation (33).

Generally, prediction models perform worse in new datasets compared with the sample in which they were developed. Models that perform without large decrements in accuracy in new datasets are more likely to generalize to other contexts we might care about—for example, clinical practice. Conversely, substantial decrements in performance suggest the model is overfitted. Emerging frameworks evaluate both the magnitude of and reasons for performance degradation (34–36); such as the validation cohort being inherently different from the derivation cohort (36). However, determining how much of a decrease in performance during validation is too much proves difficult and will likely depend on the specific study (35).

The typical lifecycle of a prediction model thus involves progression through various stages of derivation and validation. Often, initial model descriptions may be based on validation in archival datasets. When entirely separate datasets are not available, a common compromise approach is to split a single dataset into two parts: a derivation cohort, and a separate validation cohort that is not used in developing the model itself. There are several strategies to handle a single dataset in this way, each with their own drawbacks and benefits (**Supplemental Table 1**, Supplemental Digital Content 1, http://links.lww.com/CCM/F343) (37). We acknowledge that the distinction between internal and external validation is not always concrete, and will depend on study context.

TABLE 2. Selected Measures to Evaluate Prediction Model Performance

Measure	Property	Meaning	Interpretation	Uses	Pitfalls/Misuses
Scaled Brier's Score	Explained variance (discrimination + calibration)	Squared difference of observed vs predicted outcomes, standardized to the score of a noninformative model.	Represents total variation in event the model cannot explain. Higher is better: 1 = all variation explained by model, 0 = no variation explained by model.	Highly informative summary statistic. Represents overall prediction performance.	Does not directly reflect predictive value of individual predictions. Influenced by performance in irrelevant regions of the model.
AUROC (*C*-statistic)	Discrimination	The probability a true event will have higher predicted probability than a nonevent across all possible threshold values (rank order test).	Reflects overall accuracy in discriminating those that experienced the outcome from those that did not. Higher is better: 1.0 = perfect discrimination; 0.5 = noninformative prediction.	May be useful to visually assess performance over range of possible thresholds. Potentially suitable for comparing benchmarking models that use prediction frameworks.	Not relevant for clinical decisions (retrospective). Biased when events occur infrequently. Only assesses discrimination; incomplete view of performance. Influenced by performance in irrelevant regions of the model.
Area under precision recall curve	Discrimination	The average probability that a positive prediction will be a true event across all possible sensitivities (i.e., the average precision).	Reflects overall probability that any given positive prediction will become a true event. Higher is better: 1.0 = perfect precision; the overall event frequency = noninformative prediction.	Avoids the rare event bias of AUROCs. Reflects positive predictive value and sensitivity. Like AUROC, useful to visualize performance over range of possible thresholds.	Only assesses discrimination; incomplete view of performance. Influenced by performance in irrelevant regions of the model. Interpretation not intuitive.
Hosmer–Lemeshow Test	Calibration	Observed probability vs predicted probability across deciles of prediction.	Statistical hypothesis test—the null hypothesis is that the model fits the data. At high test statistics (low *p* values), reject the null.	Easily represented and interpreted graphically. Graphing is itself useful to show specific regions of risk misspecification.	Arbitrary grouping may poorly reflect risk distributions. May not detect subtle mis-calibrations in smaller datasets. May be overly conservative in larger datasets. Only assesses calibration; incomplete view of performance.
Net benefit (decision curve analysis)	Utility	Number of true positives minus weighted number of false positives, divided by sample size, plotted over range of threshold probabilities. Weight is the ratio of harm to benefit.	Higher net benefit indicates more utility at a given probability threshold.	Allows (unequal) weighting of false positives vs false negatives. Reflects the usefulness of any individual model prediction.	How "usefulness" is operationally defined is subjective.

AUROC = area under receiver operator characteristics curve.

Nevertheless, a compelling validation must use the exact same model obtained from the derivation exercise. Retraining on external data will yield an entirely new model, potentially leading to the same risks of overfitting that validation is meant to overcome.

In the near future, we may see a new approach in which models are deliberately re-trained on local data (38). This strategy might improve accuracy in the specific venue where the method is applied and may become more feasible as more health systems accrue the large datasets and computing infrastructure needed to support such a strategy. Local implementations of any given method may differ in the features selected, and the weights assigned within each local model. In these cases, the prediction method must be shown to generalize, rather than the model itself (38). Nonetheless, validation of the specific method under study is still necessary.

While precise approaches can vary between applications, the most important element of any validation is that the model's performance is interrogated among observations that were not used in its development or fine tuning. An appropriate validation is the most critical component of assessing model performance. Without it, even models that appear highly accurate are simply quantitative hypotheses.

Performance Metrics

There are numerous ways to quantify a prediction model's performance (**Table 2**). These measures are reviewed in detail elsewhere (39). Generally, the appropriateness of any given metric will depend on how we intend to use a model and the nature of the data it describes. For binary outcomes, the often-used AUROC (also known as the *C*-statistic) measures discrimination—the ability to separate events from nonevents. Indeed prediction models that cannot discriminate are useless, but when discrimination is reasonable, usefulness may become dependent on calibration, which is the ability to specify the probability of the outcome correctly (40). Separating two patients with 1% versus 5% risk has different implications than separating patients with 5% versus 25% risk, yet both correspond to a five-fold risk difference.

When outcomes occur infrequently, measures of overall accuracy can have misleadingly strong discrimination if they prioritize specificity: if only 5% of patients experience an outcome, a model that predicts zero outcomes can attain 95% accuracy (41). Sensitivity and specificity, reflected by AUROC, are inherently retrospective properties (the number of correct predictions among cases versus noncases, respectively). This may be appropriate for benchmarking indices (e.g., severity of illness scoring), but bedside prediction models are better judged by properly contextualized positive and negative predictive values (how many correctly predicted outcomes among the positive and negative predictions, respectively). These measures also have caveats; predictive values vary with the underlying outcome prevalence and therefore can vary considerably between populations.

Priorities for prediction are context-specific but invariably involve trade-offs determined by the potential consequences of false positives (i.e., overtreatment), and false negatives (i.e., missed cases). For example, in predicting poor neurologic recovery following cardiac arrest, we may wish to know the performance at a false positive rate of 0; we never want to wrongly predict a poor outcome, as this is likely to lead to the withdrawal of life-sustaining therapies. In other cases, it may be more important to accurately predict as many positive outcomes as possible, with less regard to "false alarms."

TABLE 3. Key Reporting Metrics for Prediction Models (Adapted From the Transparent Reporting of a Multivariable Prediction Model For Individual Prognosis or Diagnosis Checklist) (5)

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/variables. 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.. area under receiver operator characteristics curve or area under precision recall curve), a measure of calibration (e.g.. Hosmer–Lemeshow, scaled Brier Score), and clinically relevant performance (e.g.. positive predictive value, negative predictive value) as indicated.

Interpreting Performance

It is very rare that a novel prediction model does not require comparison to an existing one. We should seek these comparisons because it is hard to interpret a model's usefulness in isolation. A prediction model that discriminates with 75% accuracy might be useful if currently used frameworks are little better than a coin flip. Another model with 80% accuracy might not be useful if clinical judgment is right 90% of the time. Similarly, a model might accurately predict the onset of a condition at a given time, but this will only be useful if the diagnosis has not already been made. Considerations beyond standard performance metrics may therefore be important. For example, a model that accurately predicts the onset of infection may have limited clinical usefulness if the majority of patients identified are already receiving antibiotics at the time of the alert.

In the era of big data, the distinction between clinical and statistical significance becomes particularly important. The width of CIs and the size of *p* values are both inversely proportional to the sample size of a dataset. As models developed from datasets with thousands of patients become more common, it is important to consider what a "significant" difference really means. In such datasets, a complex, 40-variable model with 85% accuracy may well be statistically distinguishable from a simple, two-variable model with 84% accuracy. We must then ask whether that difference is meaningful, and worth all the added complexity.

Additional Considerations

We encourage readers to review additional considerations related to assessing model bias, competing risks, measurement error, algorithm generalization, and so-called "modifiable risk factors" in the accompanying **Supplemental Materials** (Supplemental Digital Content 1, http://links.lww.com/CCM/F343).

GUIDANCE FOR REPORTING

Table 3 provides key components that must be present for readers to properly evaluate a prediction model. More detailed guidance for reporting prediction modeling studies is widely available. We encourage authors to refer to the TRIPOD checklist (https://www.tripod-statement.org) and ensure they include all recommended elements. Studies that leverage EHR derived datasets present unique considerations, and for these, we also encourage authors to refer to the Reporting of studies Conducted using Observational Routinely-collected Data (RECORD) checklist (https://www.record-statement.org). Including these checklists in submissions is highly recommended. In general, reporting should be as transparent as possible, and should include full specification of statistical models and their diagnostics. Though making data and statistical code available is not required at most journals, it is certainly encouraged.

FINAL COMMENTS–CONSIDERING IMPACT

Above, we outline what prediction models do, and offer recommendations for their use (summarized in **Table 4**). However, considering why a new prediction model is needed may supersede all these considerations. Many published prediction models will never be used. Therefore, our final guidance is to consider what unmet need a prediction model confronts. Perhaps a model has novelty, addressing diseases or outcomes where no data currently exist. Perhaps it shows clinical usefulness, flagging occultly high-risk patients or improving discrimination compared with current practice. Perhaps a model

TABLE 4. 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 Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) and Reporting of studies Conducted using Observational Routinely-collected Data (RECORD).	

facilitates enrolling enriched populations in clinical trials, or provides an administrative index to make institutional comparisons equitable. We challenge authors to reflect on how their models will benefit patients both when designing their studies and preparing their manuscripts. For example, do authors of a bedside tool now plan to use the model in their own practice?

Drs. Leisman and Maslove conceived of the concept for the article. The core writing group included Drs. Leisman, Harhay, Lederer, Maslove, with many others providing important revisions. All authors approved of the final version and appear as signatories for their journals.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccmjournal).

Dr. Harhay's institution received funding from National Institutes of Health (NIH)/National Heart, Lung, and Blood Institute under Award Number K99HL141678. Drs. Harhay, Cooke, and Moorman received support for article research from the NIH. Dr. Lederer received funding from Roche, Regeneron Pharmaceuticals, Boehringer Ingelheim, Sanofi Genzyme, Fibrogen, Global Blood Therapeutics, Galecto, Veracyte, Pliant therapeutics, BMS, and Galapagos. Dr. Abramson's institution received funding from Pfizer, Boehringer-Ingelheim, Sanofi, and GlaxoSmithKline (GSK) (speaker). Dr. Ballas received funding from Immune Deficiency Foundation. Dr. Bernstein received funding from INEOS (Medical Director of Immunosurveillance), Shire, CSL Behring, Pharming, AZ, Sanofi-Regeneron, Optinose, Kalvista, Biocryst, Bernstein Allergy Group (partner), Bernstein Clinical Research Center (partner), and the University of Cincinnati. His institution received funding from NIH UO1 and NIH R34. Dr. Collop's institution received funding from Jazz Pharmaceuticals, and she received funding from UptoDate. Dr. Donaldson received funding from Micom, AstraZeneca, and the Flanders Research Organisation. Dr. Hale's institution received funding from the NIH, and she received funding from National Sleep Foundation. Dr. Kochanek received funding from serving as an expert witness. Dr. Marks' institution received funding from AstraZeneca and GSK. Dr. Moorman received funding from Advanced Medical Predictive Devices, Diagnostics and Displays. Dr. Schatz's institution received funding from Merck, ALK, and Teva. Dr. Stewart's institution received funding from National Institute for Health Research Biomedical Research Centre. Dr. Teboul received funding from Pulsion/Getinge. Dr. Wedzicha received research grant funding from Astra Zeneca, Boehrnger, GSK, Chiesi, Novartis, and Johnson & Johnson. Dr. Maslove was supported by the Southeastern Ontario Academic Medical Organization. The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article: david.maslove@queensu.ca

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

1. Altman DG, Vergouwe Y, Royston P, et al: Prognosis and prognostic research: Validating a prognostic model. *BMJ* 2009; 338:b605
2. Moons KG, Altman DG, Vergouwe Y, et al: Prognosis and prognostic research: Application and impact of prognostic models in clinical practice. *BMJ* 2009; 338:b606
3. Moons KG, Royston P, Vergouwe Y, et al: Prognosis and prognostic research: What, why, and how? *BMJ* 2009; 338:b375
4. Royston P, Moons KG, Altman DG, et al: Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009; 338:b604
5. Steyerberg E: Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Second Edition. New York, Springer International Publishing, 2019
6. Collins GS, Reitsma JB, Altman DG, et al: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med* 2015; 162:735–736
7. Moons KG, Altman DG, Reitsma JB, et al: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; 162:W1–73
8. Lederer DJ, Bell SC, Branson RD, 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
9. Shmueli G: To explain or to predict? *Statist Sci* 2010; 25:289–310
10. Zimmerman JE, Kramer AA, McNair DS, et al: Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; 34:1297–1310
11. Goodman SN, Goel S, Cullen MR: Machine learning, health disparities, and causal reasoning. *Ann Intern Med* 2018; 169:883–884
12. Braun L: Race, ethnicity and lung function: A brief history. *Can J Respir Ther* 2015; 51:99–101
13. Eneanya ND, Yang W, Reese PP: Reconsidering the consequences of using race to estimate kidney function. *JAMA* 2019; 322:113–114
14. Gianfrancesco MA, Tamang S, Yazdany J, et al: Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018; 178:1544–1547
15. Rajkomar A, Hardt M, Howell MD, et al: Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018; 169:866–872
16. Sun GW, Shook TL, Kay GL: Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol* 1996; 49:907–916
17. Smith G: Step away from stepwise. *J Big Data* 2018; 5:32
18. Hawkins DM: The problem of overfitting. *J Chem Inf Comput Sci* 2004; 44:1–12
19. Mundry R, Nunn CL: Stepwise model fitting and statistical inference: Turning noise into signal pollution. *Am Nat* 2009; 173:119–123
20. Walter S, Tiemeier H: Variable selection: current practice in epidemiological studies. *Eur J Epidemiol* 2009; 24:733–736
21. Wasserstein RL, Lazar NA: The ASA's statement on p-Values: Context, process, and purpose. *Am Stat* 2016; 70:129–133
22. Steyerberg EW, Eijkemans MJC, Habbema JDF: Application of shrinkage techniques in logistic regression analysis: A case study. *Stat Neerl* 2001; 55:76–88
23. Riley RD, Snell KIE, Ensor J, et al: Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Stat Med* 2019; 38:1262–1275
24. Moons KGM, Wolff RF, Riley RD, et al: PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med* 2019; 170:W1–W33
25. Riley RD, Snell KI, Ensor J, et al: Minimum sample size for developing a multivariable prediction model: PART II - Binary and time-to-event outcomes. *Stat Med* 2019; 38:1276–1296
26. van Smeden M, Moons KG, de Groot JA, et al: Sample size for binary logistic prediction models: Beyond events per variable criteria. *Stat Methods Med Res.* 2019; 28:2455–2474
27. Senn S: Dichotomania: An obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. *In:* Proceedings of the International Statistical Institute, 55th Session, 2005, Sydney, Australia, April 6-12, 2005
28. Royston P, Altman DG, Sauerbrei W: Dichotomizing continuous predictors in multiple regression: A bad idea. *Stat Med* 2006; 25:127–141
29. Sharafoddini A, Dubin JA, Maslove DM, et al: A new insight into missing data in intensive care unit patient profiles: Observational study. *JMIR Med Inform* 2019; 7:e11605
30. Ware JH, Harrington D, Hunter DJ, et al: Missing data. *N Engl J Med* 2012; 367:1353–1354
31. Newgard CD, Haukoos JS: Advanced statistics: Missing data in clinical research–Part 2: Multiple imputation. *Acad Emerg Med* 2007; 14:669–678
32. Altman DG, Royston P: What do we mean by validating a prognostic model? *Stat Med* 2000; 19:453–473
33. Steyerberg EW, Vergouwe Y: Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur Heart J* 2014; 35:1925–1931

34. van Klaveren D, Gönen M, Steyerberg EW, et al: A new concordance measure for risk prediction models in external validation settings. *Stat Med* 2016; 35:4136–4152

35. Debray TP, Vergouwe Y, Koffijberg H, et al: A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015; 68:279–289

36. Vergouwe Y, Moons KG, Steyerberg EW: External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010; 172:971–980

37. Steyerberg EW, Harrell FE Jr, Borsboom GJ, et al: Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001; 54:774–781

38. Lee J, Maslove DM: Customization of a severity of illness score using local electronic medical record data. *J Intensive Care Med* 2017; 32:38–47

39. Steyerberg EW, Vickers AJ, Cook NR, et al: Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* 2010; 21:128–138

40. Alba AC, Agoritsas T, Walsh M, et al: Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *JAMA* 2017; 318:1377–1384

41. Leisman DE: Rare events in the ICU: An emerging challenge in classification and prediction. *Crit Care Med* 2018; 46:418–424

APPENDIX 1. CONTRIBUTING AUTHORS AND JOURNALS

Core writing group: Daniel E. Leisman (Editorial Board, *Critical Care Medicine*), Michael O. Harhay (Statistical Editor, *Annals of the American Thoracic Society*), David J. Lederer (Former Editor-in-Chief, *Annals of the American Thoracic Society*), David M. Maslove (Associate Editor, *Critical Care Medicine*). **Contributing authors**: Michael Abramson (Statistical Review Board, *Respirology*), Alex A. Adjei (Editor-in-Chief, *Journal of Thoracic Oncology*), Jan Bakker (Editor-in-Chief, *Journal of Critical Care*), Zuhair K. Ballas (Editor-in-Chief, *The Journal of Allergy & Clinical Immunology*), Esther Barreiro (Editor-in-Chief, *Archivos de Bronconeumologia*), Scott C. Bell (Editor-in-Chief, *Journal of Cystic Fibrosis*), Rinaldo Bellomo (Editor-in-Chief, *Critical Care & Resuscitation*), Jonathan A Bernstein (Editor-in-Chief, *Journal of Asthma*), Richard D. Branson (Editor-in-Chief, *Respiratory Care*), Vito Brusasco (Editor-in-Chief, *COPD: Journal of Chronic Obstructive Pulmonary Disease*), James D. Chalmers (Deputy Chief Editor, *European Respiratory Journal*), Sudhansu Chokroverty (Editor-in-Chief, *Sleep Medicine*), Giuseppe Citerio (Editor-in-Chief, *Intensive Care Medicine*), Nancy A. Collop (Editor-in-Chief, *Journal of Clinical Sleep Medicine*), Colin R. Cooke (Editor-in-Chief, *Annals of the American Thoracic Society*), James D. Crapo (Editor-in-Chief, *Journal of the COPD Foundation*), Gavin Donaldson (Associate Editor, *American Journal of Respiratory and Critical Care Medicine*), Dominic A. Fitzgerald (Editor-in-Chief, *Paediatric Respiratory Reviews*), Emma Grainger (Editor, *The Lancet Respiratory Medicine*), Lauren Hale (Editor-in-Chief, *Sleep Health*), Felix J. Herth (Editor-in-Chief, *Respiration*), Patrick M. Kochanek (Editor-in-Chief, *Pediatric Critical Care Medicine*), Guy Marks (Editor-in-Chief, *International Journal of Tuberculosis and Lung Disease*), J. Randall Moorman (Editor-in-Chief, *Physiological Measurement*), David E. Ost (Editor-in-Chief, *Journal of Bronchology and Interventional Pulmonology*), Michael Schatz (Editor-in-Chief, *The Journal of Allergy & Clinical Immunology: In Practice*), Aziz Sheikh (Editor-in-Chief, *NPJ: Primary Care Respiratory Medicine*), Alan R. Smyth (Joint Editor-in-Chief, *Thorax*), Iain Stewart (Statistical Editor, *Thorax*), Paul W. Stewart (Associate Editor, *Pediatric Pulmonology*), Erik R. Swenson (Editor-in-Chief, *High Altitude Medicine & Biology*), Ronald Szymusiak (Editor-in-Chief, *SLEEP*), Jean-Louis Teboul (Editor-in-Chief, *Annals of Intensive Care*), Jean-Louis Vincent (Editor-in-Chief, *Critical Care*), Jadwiga A. Wedzicha (Editor-in-Chief, *American Journal of Respiratory and Critical Care Medicine*).