

Meta-analysis under the spotlight: Focused on a meta-analysis of ventilator weaning*

Martin J. Tobin, MD; Amal Jubran, MD

Objective: Because the results of a meta-analysis are used to formulate the highest level recommendation in clinical practice guidelines, clinicians should be mindful of problems inherent in this technique. Rather than reviewing meta-analysis in abstract, general terms, we believe readers can gain a more concrete understanding of the problems through a detailed examination of one meta-analysis. The meta-analysis on which we focus is that conducted by an American College of Chest Physicians/American Association for Respiratory Care/American College of Critical Care Medicine Task Force on ventilator weaning.

Data Source: Two authors extracted data from all studies included in the Task Force's meta-analysis.

Data Synthesis and Overview: The major obstacle to reliable internal validity and, thus, reliable external validity (generalizability) in biological research is systematic error, not random error. If systematic errors are present, averaging (as with a meta-analysis) does not decrease them—instead, it reinforces them, pro-

ducing artifact. The Task Force's meta-analysis commits several examples of the three main types of systematic error: selection bias (test-referral bias, spectrum bias), misclassification bias (categorizing reintubation as weaning failure, etc.), and confounding (pressure support treated as unassisted breathing). Several additional interpretative errors are present.

Conclusions: An increase in study size, as achieved through the pooling of data in a meta-analysis, is mistakenly thought to increase external validity. On the contrary, combining heterogeneous studies poses considerable risk of systematic error, which impairs internal validity and, thus, external validity. The strength of recommendations in clinical practice guidelines is based on a misperception of the relative importance of systematic vs. random error in science. (Crit Care Med 2008; 36:1–7)

KEY WORDS: meta-analysis; clinical practice guidelines; mechanical ventilation; weaning; systematic error; random error

The term *meta-analysis* was coined from the Greek prefix *meta*, meaning after or transcending, and the root *analysis*. In contrast with the original data analysis, meta-analysis constitutes a second-order analysis of analyses (1). The promise of meta-analysis is attractive: a diligent tracking down of every original study on a topic, tabulation of major results, a meticulous listing of inclusion and exclusion criteria, and a uniform ap-

proach to writing that enables a formulaic presentation (2). Few meta-analysts stop at this stage. Most calculate a summary statistic, based on pooling of data from multiple studies (3). The summary statistic is portrayed as a synthesis of all that is known about a topic, so providing a generalizable conclusion.

An underlying assumption is that, in an area of controversy, a statistical averaging technique can extract the unbiased, "true" conclusion hidden beneath a mass of conflicting data (4). The conclusion has considerable persuasive power. Meta-analysis is graded at the highest level of scientific quality by the evidence-based medicine movement (5). Accordingly, the conclusion leads automatically to the highest-level recommendation in clinical practice guidelines (5). Given the power of meta-analysis to influence patient care, practicing clinicians should be mindful of the problems inherent in this methodology.

Many authors have reviewed problems encountered in meta-analyses (6–9). The articles, however, discuss problems in abstract generalities. We believe readers can gain a more concrete understanding of

the problems through examination of a single meta-analysis in depth. The selected meta-analysis is one conducted by an American College of Chest Physicians/American Association for Respiratory Care/American College of Critical Care Medicine Task Force on ventilator weaning (10, 11). This report has several attractions. It was published >5 yrs ago, and, during the intervening time, serious reservations have not been published. The report received the imprimatur of three professional societies. The report contains several high-level recommendations for patient management. The authors are among the most experienced in the methodology of meta-analysis, evidence-based medicine, and formulation of clinical practice guidelines.

Task Force Findings

The Task Force focused predominantly on frequency-to-tidal volume ratio (f/V_T), a measure of rapid shallow breathing (12, 13); an f/V_T value of <100 suggests that a patient is likely to tolerate ventilator discontinuation (12). From 22 studies of f/V_T (Table 1) (12, 14–34), the

*See also pp 328 and 329.

From the Division of Pulmonary and Critical Care Medicine, Edward Hines Jr. Veterans Affairs Hospital, Hines, IL; and Loyola University of Chicago Stritch School of Medicine, Hines, IL.

Dr. Tobin receives royalties for two books on critical care published by McGraw-Hill, New York, NY. Dr. Jubran has not disclosed any potential conflicts of interest.

Supported, in part, by a Merit Review grant from the Veterans Administration Research Service and by grant R01 NR008782 from the National Institutes of Health.

For information regarding this article, E-mail: mtobin2@lumc.edu

Copyright © 2007 by the Society of Critical Care Medicine and Lippincott Williams & Wilkins

DOI: 10.1097/01.CCM.0000297883.04634.11

Table 1. Studies of frequency-to-tidal volume ratio (f/V_T) in the meta-analysis of the American College of Chest Physicians/American Association for Respiratory Care/American College of Critical Care Medicine Task Force

Authors	No. of Patients	Outcome End Point	Pretest Probability of Success	Sensitivity	Specificity	Data Available to Primary Physician
Yang and Tobin (12)	64	WF or EF	0.56	0.97	0.64	No
Gandia and Blanco (30)	40	WF or EF	0.7	0.89	0.83	No
Sassoon and Mahutte (15)	45	WF or EF	0.78	0.97	0.40	No
Yang (19)	31	WF or EF	0.52	0.94	0.73	No
Mohsenifar et al. (33)	29	WF or EF	0.62	1.00	0.27	Not clear
Lee et al. (23)	52	EF only	0.83	0.72	0.11	Yes
Capdevila et al. (29)	67	EF only	0.82	0.73	0.75	Yes
Epstein (20)	94	EF only	0.81	0.92	0.22	Yes
Chatila et al. (16)	100	WF or EF	0.63	0.89	0.41	Yes
Chatila et al. (16)	100	WF or EF	0.63	0.98	0.59	Yes
Dojat et al. (17)	38	WF or EF	0.45	0.94	0.81	No
Leitch et al. (22)	163	EF only	0.982	0.96	0.00	Yes
Mergoni et al. (24)	75	WF or EF	0.49	0.65	0.58	Yes
Epstein and Ciubotaru (21)	218	EF only	0.84	NR	NR	Yes
Khan et al. (27)	208	EF only	0.84	NR	NR	No
Baumeister et al. (28)	47 Ped	EF only	0.81	0.79	0.78	No
Гологорский et al. (26)	127	Not Defined	NR	0.84	0.83	Not clear
Jacob et al. (31)	183	WF or EF	0.92	0.97	0.33	Yes
Jacob et al. (31)	183	WF or EF	0.92	0.96	0.31	Yes
Krieger et al. (25)	49	WF	0.78	0.74	0.73	Yes
Krieger et al. (25)	49	WF	0.78	0.93	0.89	Yes
Del Rosario et al. (14)	49	WF or EF	0.78	NR	NR	Not clear
Farias et al. (34)	84 Ped	WF	0.75	0.48	0.86	No
Vallverdu et al. (32)	217	WF or EF	0.58	0.90	0.36	Yes
Afessa et al. (18)	118	WF only	0.57	NR	NR	Yes

WF, weaning failure; EF, extubation failure; NR, not reported; Ped, pediatric.

The listed studies are those that reported data on the accuracy of f/V_T as a predictor of weaning outcome. Three groups of investigators, Chatila et al. (16), Jacob et al. (31), and Krieger et al. (25), report data under two different conditions in their articles; both sets of data are presented. Pretest probability of success in a study is the fraction of patients with a successful outcome out of the total population (both success and failure patients) included in the study.

Task Force calculated pooled likelihood ratios. They concluded that all predictors have low power; therefore, physicians should bypass their measurement (10, 11).

Before discussing the specific problems in this meta-analysis, we discuss issues fundamental to study design in general.

Study Size, Averaging, Validity

The first goal in a research study is to make accurate measurements, to estimate the variable of interest with the least possible error (Fig. 1). Investigators encounter two broad types of error: random and systematic (35).

Random errors occur in an unpredictable manner; they both overestimate and underestimate the true value (36). The primary means to decrease random error is to increase study size; this leads to greater confidence that the average represents the true value (35).

Systematic error consists of “any trend in the collection, analysis, interpretation, publication, or review of data that can lead to conclusions that are systematically different from the truth” (37).

Systematic errors consistently underestimate or consistently overestimate the true value (36). Unlike random error, an increase in study size does not decrease systematic error (36).

Even when measured data are accurate, systematic errors may arise secondary to selection bias, misclassification, or confounding (35) (Fig. 1). Systematic errors are the major source of error in biological research (38). If systematic errors are present, averaging (as with a meta-analysis) does not decrease them; instead, it reinforces them, producing artifact (38). Accordingly, averaging plays a much smaller role in enhancing valid scientific inference than commonly suspected.

Well-grounded internal validity (validity of inferences from the subjects in a study) depends on minimizing systematic error (Fig. 1). Internal validity is a prerequisite for legitimate external validity (also termed generalizability). Valid generalizability enables legitimate inferences about future patients in general, rather than (only) for the actual patients within a study (39); generalizability is the ultimate goal of research (39).

An increase in study size, as achieved through data pooling, is thought to lead to better generalizability. This is a misconception. Rothman (35) attributes the misconception to the fallacy of statistical representativeness: “that generalizing from an epidemiologic study involves a mechanical process of making an inference about a target population of which the study population is considered a sample.” Such a generalization does apply to survey sampling (opinion polls) (35), but it differs from generalization in science.

Scientific generalizability—formulating a general law through valid extrapolation based on a particular experiment—is not a statistical process (40). Rather, it is a process of causal inference. Generalizability is “based more on scientific knowledge, insight, and even conjecture about nature than on the statistical representativeness of the actual study participants” (35). As such, study size is not the primary concern for external validity: large numbers do not increase the ability of a study result to form a valid general law. The major obstacle to internal validity and, thus, valid generalizability

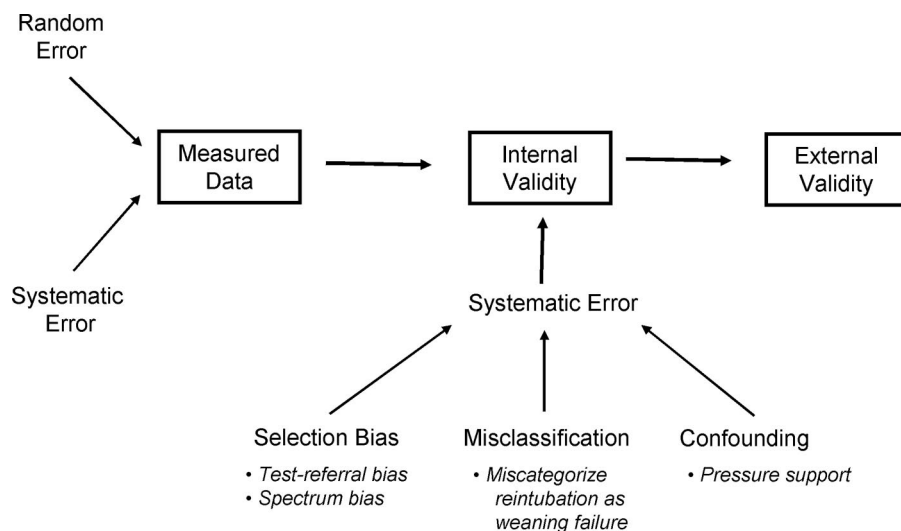


Figure 1. All data measured (in a research study) contain some random and systematic error. In a study of frequency-to-tidal volume ratio (f/V_T), random error occurs when a patient's f/V_T is based on an insufficient number of breaths; systematic error occurs if the spirometer is calibrated incorrectly. In addition to problems related to data collection in an individual patient, systematic errors may arise in a group of patients as a result of selection bias, misclassification bias, and confounding (examples discussed later in text). Designs that minimize such systematic errors enhance internal validity. External validity, also termed generalizability, is the abstract extrapolation of what the study results mean to future patients; specifically, generalizability is not based on statistical analysis, nor is it simply linked to the representativeness of the subjects in a particular study (35).

ity is systematic error, not random error (38). The failure to grasp this distinction leads to the heightened, and inappropriate, expectation that meta-analysis can uncover hidden truth (4).

Systematic Error

The meta-analysis conducted by the Task Force contains several systematic errors (10). These are grouped under three broad subheadings: selection bias, misclassification bias, and confounding (these categories can overlap with one another) (35).

Selection Bias. Selection bias is a systematic error or distortion that arises from the procedures used to select subjects for a study that leads "to an effect estimate among subjects included in the study different from the estimate obtainable from the entire population theoretically targeted for study" (40). Weaning involves undertaking three diagnostic tests in sequence: measurement of weaning predictor tests, followed by a weaning trial, followed by a trial of extubation. Undertaking of three tests in sequence poses an enormous risk for the occurrence of test-referral bias (41, 42).

Test-referral bias arises when a test under evaluation (weaning predictor test) influences which patients will undergo either of the two subsequent tests. If

tolerance of extubation is used as the reference standard for evaluating the reliability of the weaning predictor test, the requirement to pass a weaning trial before extubation will necessarily exclude all patients who fail a weaning trial. This step will have two effects on the study population: there will be fewer patients with negative results of the weaning predictor test and relatively more patients with positive results (41, 42).

The first consequence will produce a decrease in the specificity (true-negative rate) of the weaning predictor test in this population as compared with the population in which the test was originally developed. The second consequence will increase the sensitivity (true-positive rate) of the test. The lower mean specificity of studies in the Task-Force meta-analysis, 0.51 ± 0.27 (SD) (15–17, 19, 20, 22–25, 28–33), as contrasted with 0.64 in the original report (12), is consistent with occurrence of test-referral bias. Sensitivity of f/V_T in the original report was 0.97 (12). Because sensitivity has a ceiling of 1.00, a value of 0.97 does not leave much room to detect a further increase in sensitivity. Eleven (of 18, 61%) studies (15–17, 19, 20, 22, 25, 31–33) reveal sensitivities for f/V_T of ≥ 0.90 , a finding consistent with test-referral bias. The occurrence of test-referral bias and spec-

trum bias, in which a study population is skewed toward less-ill patients, in the Task-Force's meta-analysis is discussed in greater detail in another report (43).

Theoretically, sensitivity and specificity are constant properties of a test (41). The development of test-referral and spectrum bias, however, leads to major changes in sensitivity, specificity, and likelihood ratio (the latter is simply the combination of sensitivity and specificity) (41). The major conclusion of the Task Force is based on a summary statistic of pooled likelihood ratios (10, 11). This summary likelihood ratio, calculated by the Task Force, however, is fundamentally flawed because of the varying degree of test-referral and spectrum bias among the studies from which they derived it (41).

Test-referral and spectrum bias resulted in significant heterogeneity in the prevalence of successful outcome among the studies in the Task Force's meta-analysis ($p < .00001$) (43). Whenever significant heterogeneity is found among a group of studies, methodologists emphasize that it is invalid to calculate a summary statistic (44, 45). Greenland (46), who otherwise supports the use of meta-analysis, is harsh in his condemnation of this practice: "Synthetic meta-analyses that ignore heterogeneity should indeed be banned from publication, if only because they violate sound statistical and scientific practice."

Misclassification Bias. Misclassification bias is a systematic error arising because erroneous information causes the placing of study subjects in an incorrect category (35).

To calculate the aggregated likelihood ratios, the Task Force divided the original studies into three *mutually exclusive categories*: first, predicting outcome of an unassisted breathing trial (weaning success, weaning failure); second, predicting outcome of a trial of extubation; and third, predicting combined outcome of an unassisted breathing trial followed by extubation. In each category, they misclassified studies.

The first category included the studies of Sassoon and Mahutte (15) and Chatila et al. (16), although the investigators stated that 50% and 16.2% of their failure groups, respectively, required reintubation (thus, belonging to the third category). The Task Force included the study of del Rosario et al. (14), although the authors state that 70% of this study population were also included in the report of Sassoon and Mahutte (15). Thus, three

of four studies in the first category do not belong there.

Among the studies in the second category are those of Yang (19) and Mergoni et al. (24), both of whom listed weaning failure as an outcome (thus, belonging to the third category). Krieger et al. (25) did not state that any patient was extubated (thus, belonging to the first category). Leitch et al. (22) stated that 20.9% of their patients required noninvasive ventilatory assistance after extubation. Some investigators (47–49) consider such patients to represent extubation failure; the Task Force did not mention this consideration.

Included in the third category is a study by Capdevila et al. (29), although the authors stated that they did not measure predictors until after the patients had already tolerated 20 mins of a T-tube trial (by which time many weaning-failure patients will have already declared themselves). Moreover, these authors used an f/V_T threshold of 60, and their data were merged with studies using thresholds of 100 or 105.

In addition, misclassification bias arose from merging of 1) studies in which the authors did not state criteria for deciding weaning success or weaning failure (26), 2) studies varying from 60 mins (50) to 3–7 days (18) between the measurement of f/V_T and ventilator discontinuation, 3) studies in which patients in one report were included in a second report (14, 50), and 4) studies in infants were merged with studies in adults (27, 28).

One study (26) lists sensitivities and specificities of weaning predictor tests according to three diagnostic categories (pneumonia, cardiac dysfunction, polyorgan dysfunction). The authors, however, do not specify the proportion of patients in each category that were weaning successes or failures. Thirty-three patients were reintubated for repeat surgery rather than respiratory distress; they do not specify whether these patients were included in the weaning-failure category (26). The Task Force included this study in their meta-analysis, but do not indicate how they handled the unclear classification and major inconsistencies in the primary data.

Confounding. Confounding is a situation in which a noncausal association between an exposure and an outcome is observed as a result of a third variable (the confounder) (51). In the case of a weaning predictor test, a confounding variable is one that distorts the numerical

values generated by the test, altering their ability to predict weaning outcome (41).

The Task Force writes: “We include as trials of ‘unassisted’ breathing those trials completed on a low level of pressure support to overcome the additional work of breathing through a ventilator circuit” (10). It is an oxymoron to label pressure support as unassisted breathing. The word *support* is an antonym of *unassisted*. The problem is not simply semantic: f/V_T is 23–52% higher during unassisted breathing than with pressure support of 5 cm H₂O and is 46–82% higher during unassisted breathing than with pressure support of 10 cm H₂O (50, 52–54).

The Task Force categorized the data of Mohsenifar et al. (33) as “unassisted breathing,” although the investigators measured frequency and V_T at a pressure support of “about 7 to 8 cm H₂O.” The Task Force, however, categorized a study by Rivera and Weissman (55) as “mechanical support,” although the investigators report data at a pressure support of 5 cm H₂O.

The Task Force asserts that pressure support is simply overcoming the resistance posed by an endotracheal tube. This assertion fails to recognize the higher-than-normal resistance of the upper airway after removal of an endotracheal tube. Straus et al. (56) found work of breathing equivalent after extubation to what it had been in patients breathing through an endotracheal tube.

Errors of Interpretation

Interpretation of research data requires rigorous application of logic (57). The Task Force commits the formal deductive fallacy of *affirming the consequent*. Consider the following *antecedent, consequent, and conclusion*: a) if f/V_T is a poor predictor, then pooled likelihood ratios will be weak; b) the pooled likelihood ratios are weak; c) therefore, f/V_T is a poor predictor. The conclusion is a *non sequitur*, it is not a logical deduction from the premises. This fallacy sometimes arises because it resembles the premise of a *modus ponens*, a valid argument affirming the antecedent.

Expertise in a Research Field

Before embarking on research, individuals need to have mastery of that particular area of biology (58). The following are a few examples of where the Task Force failed to appreciate long-estab-

lished principles of diagnostic testing and physiology.

Weaning predictor tests constitute a form of diagnostic testing. The clinical worth of a diagnostic test depends on the magnitude of change between pretest probability (clinical gestalt) and posttest probability (41). Bayes’ theorem determines this magnitude for every diagnostic test. The fulcrum around which Bayes’ theorem revolves is pretest probability (42). However, the Task Force failed to take this factor into account. When we entered the data from studies in the Task Force’s meta-analysis into a Bayesian model, with pretest probability as the operating point, the reported posttest probabilities were significantly correlated with the values predicted by the original report on f/V_T ($r = .86$ and 0.82 ; $p < .0001$) (Fig. 2) (43).

Turning to physiology, the Task Force writes: “We include as trials of ‘unassisted’ breathing . . . those completed on a low level of continuous positive airway pressure to offset the loss of physiologic continuous positive airway pressure caused by the presence of an endotracheal tube” (10). This claim of physiologic positive end-expiratory pressure (PEEP) does not square with longstanding knowledge of pulmonary physiology. Lung volume at end expiration generally approximates the relaxation volume of the respiratory system: the volume determined by the static balance between the opposing elastic recoil of the lung and chest wall (59–61). The static recoil pressure of the respiratory system is thus zero at end expiration in a healthy person. Addition of PEEP is of consequence in ventilated patients: Smith and Marini (62) found that PEEP of 5 cm H₂O decreased work of inspiration by 19% in patients with exacerbated air flow obstruction.

Two, they regard normal airway resistance to be as high as 15 cm H₂O·L⁻¹·sec⁻¹ (11). The normal value is <5 cm H₂O·L⁻¹·sec⁻¹ (63–65).

Three, they consider rapid shallow breathing as “advantageous from an energetics perspective” (11). A low tidal volume (characteristic of weaning-failure patients), however, leads to a relative increase in dead-space ventilation (13). Compensation through an increase in frequency will produce an increase in work of breathing (per minute). The increase in work of breathing with rapid shallow breathing is not simply linear. It is expo-

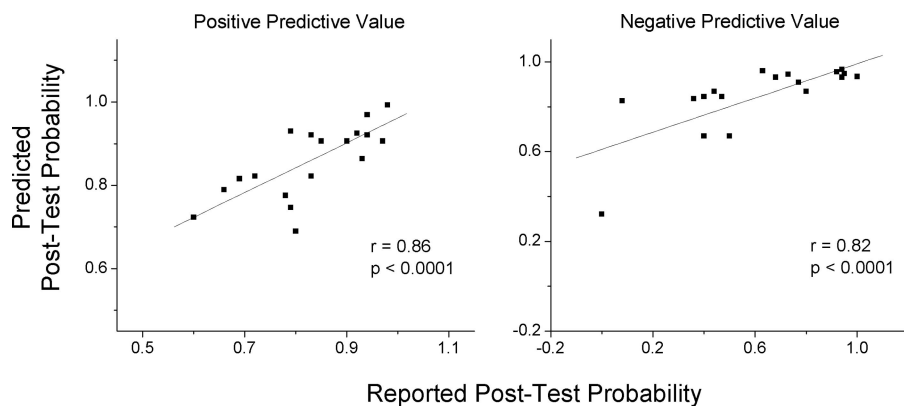


Figure 2. The relationship between the reported posttest probabilities for frequency-to-tidal volume ratio (f/V_T) (among the studies included in the meta-analysis of the American College of Chest Physicians/American Association for Respiratory Care/American College of Critical Care Medicine Task Force) and the values predicted by observed pretest probability together with the sensitivity and specificity originally reported by Yang and Tobin (12). The weighted Pearson's correlation for positive predictive value was $r = .86$, $p < .0001$ (left) and for negative-predictive value was $r = .82$, $p < .0001$ (right). Reproduced with permission from Tobin and Jubran (43).

nential: work = $4035e^{(0.0017 \cdot f/V_T)}$, $r = .90$ (66, 67). No advantage in that.

Systematic Reviews. "A meta-analysis is a type of systematic review that uses statistical methods to combine and summarize the results of several primary studies" (67). Promoters of systematic reviews write: "The concepts and techniques involved, including that of meta-analysis, are at least as subtle and complex as many of those currently used in molecular biology. . . . The task has already been likened in scope and importance to the Human Genome Project" (4). Promoters claim that a "systematic review faithfully summarizes the evidence from all relevant studies on the topic of interest, and it does so concisely and transparently" (68). And, "As their name implies, systematic reviews—not satisfied with finding part of 'the truth'—look for 'the whole truth'" (4).

How well does the Task Force's meta-analysis match these claims? In the methods section of their report, they state they evaluated reports to see whether investigators enrolled a representative sample of patients. They concluded (10): "The reporting of patient selection in individual studies was not detailed. For the vast majority of studies, selection bias was not evident." On the contrary, most investigators reported criteria for enrolling patients. As discussed above, there is striking evidence for selection bias (in the form of test-referral bias). For example, the Task Force included in their meta-analysis a report in which the investigators explicitly state they used frequency of >35 breaths/min

and V_T of <5 mL/kg as exclusion criteria when enrolling patients (32); for an average 70-kg patient, these criteria result in f/V_T of <100 . The meta-analysis contains additional systematic biases (misclassification, confounding).

A major difference between systematic reviews and traditional review articles is money. Granting agencies will not fund authors who wish to write a traditional review article. Systematic reviews secure considerable funding. The meta-analysis on weaning received \$248,356 from the Agency for Health Care Policy and Research, a subsidiary of the U.S. Department of Health and Human Services (CG Williams, personal communication).

Prescriptive Connotation

The results of a meta-analysis are automatically accorded level 1 status in the evidence-based medicine hierarchy of evidence and, thus, serve as the warrant for a grade A recommendation in clinical practice guidelines (5).

The Evidence-Based Medicine Working Group base their "approach to classifying strength of recommendations" on the (unproven) claim that systematic reviews are more precise and less biased than traditional review articles (69). They write: "Recently, the health sciences community has reduced the bias and imprecision of traditional literature summaries and their associated recommendations through the development of rigorous criteria for both literature overviews and practice guidelines" (5, 69).

The Working Group specifies that the strength of recommendations revolves around sample size and random error: "the greater the sample size, the more precise our estimates of intervention effects, the narrower the confidence interval (CI) around our estimate of those effects, and the greater our ability to make strong recommendations" (69). Statistical precision and confidence interval relate to random error (35), but the main obstacle to making a valid scientific generalization (clinical recommendation) is systematic error, not random error (38, 40); moreover, sample size is not relevant to systematic error or external validity (39, 40). The Task Force's meta-analysis provides a catalog of systematic errors. Thus, the system used to grade clinical recommendation rests on a misunderstanding of the difference between science and statistics.

Physicians need to be aware that the grading used for strength of clinical recommendations is based on a misperception of the relative importance of systematic vs. random error in science. Unaware, physicians risk placing patients in harm's way.

Conclusion

Science and medicine have traditionally assumed that clinical researchers with first-hand experience of the limitations in the methodology of a field are best placed to judge how new research might be applied to patient care. A meta-analysis, however, is conducted at one or more removes from the original data. The greater the number of removes between the primary researcher and the meta-analysts, the greater the likelihood of error. Commonly, meta-analysts have limited hands-on experience with the instrumentation (and study design) used for recording the primary data and thus fail to detect systematic errors in the primary studies (6–8).

It is disappointing that physicians have accepted uncritically the technique of meta-analysis. If authors were to submit a manuscript based on an original research study that contained the systematic errors included in the Task Force's meta-analysis (selection bias, misclassification bias, and confounding), a conscientious reviewer would instantly recommend rejection. However, the same systematic errors are ignored when included in a meta-analysis. It is as if mathematical pooling can act in the manner of

a fractionating column, cleansing the data of error and yielding a purified distillate.

The quantitative allure of extensive pooling and aggregation bestows on the calculations a false sense of statistical rigor. The false halo surrounding a bottom line, summary statistic breeds a sense of finality and “discourages innovative investigators from seeking new approaches to elusive problems” (70).

In their founding article, the evidence-based medicine movement stressed the importance of applying rules of evidence to determine the validity of a study (71). Former Supreme Court Justice Felix Frankfurter noted, “The validity and moral authority of a conclusion largely depends on the mode by which it was reached.” The mode whereby the summary statistic was calculated by the Task Force, without control for numerous systematic errors, raises serious doubts about the “validity and moral authority” of the conclusion. We welcome criticisms from members of the Task Force regarding our critique.

REFERENCES

1. Glass GV: Primary, secondary, and meta-analysis of research. *Educ Res* 1976; 5:3–8
2. Irwig L, Tosteson AN, Gatsonis C, et al: Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994; 120:667–676
3. Bailar JC III: The promise and problems of meta-analysis. *N Engl J Med* 1997; 337:559–561
4. Mulrow CD, Cook DJ, Davidoff F: Systematic reviews: Critical links in the great chain of evidence. *Ann Intern Med* 1997; 126:389–391
5. Cook DJ, Guyatt GH, Laupacis A, et al: Clinical recommendations using levels of evidence for antithrombotic agents. *Chest* 1995; 108(4 Suppl):227S–230S
6. Feinstein AR: Meta-analysis: Statistical alchemy for the 21st century. *J Clin Epidemiol* 1995; 48:71–79
7. Shapiro S: Is meta-analysis a valid approach to the evaluation of small effects in observational studies? *J Clin Epidemiol* 1997; 50:223–229
8. Bailar JC III: The practice of meta-analysis. *J Clin Epidemiol* 1995; 48:149–157
9. Bailar JC III: Passive smoking, coronary heart disease, and meta-analysis. *N Engl J Med* 1999; 340:958–959
10. Meade M, Guyatt G, Cook D, et al: Predicting success in weaning from mechanical ventilation. *Chest* 2001; 120(6 Suppl):400S–424S
11. MacIntyre NR, Cook DJ, Ely EW Jr, et al: Evidence-based guidelines for weaning and discontinuing ventilatory support: A collective task force facilitated by the American College of Chest Physicians; the American Association for Respiratory Care; and the American College of Critical Care Medicine. *Chest* 2001; 120(6 Suppl):375S–395S
12. Yang KL, Tobin MJ: A prospective study of indexes predicting the outcome of trials of weaning from mechanical ventilation. *N Engl J Med* 1991; 324:1445–1450
13. Tobin MJ, Perez W, Guenther SM, et al: The pattern of breathing during successful and unsuccessful trials of weaning from mechanical ventilation. *Am Rev Respir Dis* 1986; 134:1111–1118
14. del Rosario N, Sassoon CS, Chetty KG, et al: Breathing pattern during acute respiratory failure and recovery. *Eur Respir J* 1997; 10:2560–2565
15. Sassoon CS, Mahutte CK: Airway occlusion pressure and breathing pattern as predictors of weaning outcome. *Am Rev Respir Dis* 1993; 148(4 Pt 1):860–866
16. Chatila W, Jacob B, Guaglianone D, et al: The unassisted respiratory rate-tidal volume ratio accurately predicts weaning outcome. *Am J Med* 1996; 101:61–67
17. Dojat M, Harf A, Touchard D, et al: Evaluation of a knowledge-based system providing ventilatory management and decision for extubation. *Am J Respir Crit Care Med* 1996; 153:997–1004
18. Afessa B, Hogans L, Murphy R: Predicting 3-day and 7-day outcomes of weaning from mechanical ventilation. *Chest* 1999; 116:456–461
19. Yang KL: Inspiratory pressure/maximal inspiratory pressure ratio: A predictive index of weaning outcome. *Intensive Care Med* 1993; 19:204–208
20. Epstein SK: Etiology of extubation failure and the predictive value of the rapid shallow breathing index. *Am J Respir Crit Care Med* 1995; 152:545–549
21. Epstein SK, Ciubotaru RL: Influence of gender and endotracheal tube size on preextubation breathing pattern. *Am J Respir Crit Care Med* 1996; 154(6 Pt 1):1647–1652
22. Leitch EA, Moran JL, Grealby B: Weaning and extubation in the intensive care unit: Clinical or index-driven approach? *Intensive Care Med* 1996; 22:752–759
23. Lee KH, Hui KP, Chan TB, et al: Rapid shallow breathing (frequency-tidal volume ratio) did not predict extubation outcome. *Chest* 1994; 105:540–543
24. Mergoni M, Costa A, Primavera S, et al: Valutazione di alcuni nuovi parametri predittivi dell'esito dello svezzamento dalla ventilazione meccanica. *Minerva Anestesiol* 1996; 62:153–164
25. Krieger BP, Isber J, Breitenbucher A, et al: Serial measurements of the rapid-shallow-breathing index as a predictor of weaning outcome in elderly medical patients. *Chest* 1997; 112:1029–1034
26. ●●●
27. Khan N, Brown A, Venkataraman ST: Predictors of extubation success and failure in mechanically ventilated infants and children. *Crit Care Med* 1996; 24:1568–1579
28. Baumeister BL, el Khatib M, Smith PG, et al: Evaluation of predictors of weaning from mechanical ventilation in pediatric patients. *Pediatr Pulmonol* 1997; 24:344–352
29. Capdevila XJ, Perrigault PF, Perey PJ, et al: Occlusion pressure and its ratio to maximum inspiratory pressure are useful predictors for successful extubation following T-piece weaning trial. *Chest* 1995; 108:482–489
30. Gandia F, Blanco J: Evaluation of indexes predicting the outcome of ventilator weaning and value of adding supplemental inspiratory load. *Intensive Care Med* 1992; 18:327–333
31. Jacob B, Chatila W, Manthous CA: The unassisted respiratory rate/tidal volume ratio accurately predicts weaning outcome in post-operative patients. *Crit Care Med* 1997; 25:253–257
32. Vallverdu I, Calaf N, Subirana M, et al: Clinical characteristics, respiratory functional parameters, and outcome of a two-hour T-piece trial in patients weaning from mechanical ventilation. *Am J Respir Crit Care Med* 1998; 158:1855–1862
33. Mohsenifar Z, Hay A, Hay J, et al: Gastric intramural pH as a predictor of success or failure in weaning patients from mechanical ventilation. *Ann Intern Med* 1993; 119:794–798
34. Farias JA, Alia I, Esteban A, et al: Weaning from mechanical ventilation in pediatric intensive care patients. *Intensive Care Med* 1998; 24:1070–1075
35. Rothman KJ: *Epidemiology: An Introduction*. New York, Oxford University Press, 2002
36. Chatburn RL: Principles of measurement. In: *Principles and Practice of Intensive Care Monitoring*. Tobin MJ (Ed). New York, McGraw-Hill, 1998, pp 45–62
37. Kulkarni AV: The challenges of evidence-based medicine: A philosophical perspective. *Med Health Care Philos* 2005; 8:255–260
38. Charlton BG: The scope and nature of epidemiology. *J Clin Epidemiol* 1996; 49:623–626
39. Horton R: Common sense and figures: The rhetoric of validity in medicine (Bradford Hill Memorial Lecture 1999). *Stat Med* 2000; 19:3149–3164
40. Rothman KJ: *Modern Epidemiology*. Boston, Little, Brown, 1986
41. Feinstein AR: *Clinical Epidemiology: The Architecture of Clinical Research*. Philadelphia, WB Saunders, 1985
42. Sox HC Jr, Clatt MA, Higgins MC, et al: *Medical Decision Making*. Boston, Butterworths, 1988
43. Tobin MJ, Jubran A: Variable performance of weaning-predictor tests: Role of Bayes' theorem and spectrum and test-referral bias. *Intensive Care Med* 2006; 32:2002–2012
44. Brand R, Kragt H: Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat Med* 1992; 11:2077–2082
45. Schmid CH, Lau J, McIntosh MW, et al: An

- empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med* 1998; 17:1923–1942
46. Greenland S: Can meta-analysis be salvaged? *Am J Epidemiol* 1994; 140:783–787
 47. Maldonado A, Bauer TT, Ferrer M, et al: Capnometric recirculation gas tonometry and weaning from mechanical ventilation. *Am J Respir Crit Care Med* 2000; 161:171–176
 48. Jiang JR, Tsai TH, Jerng JS, et al: Ultrasonographic evaluation of liver/spleen movements and extubation outcome. *Chest* 2004; 126:179–185
 49. Haberthur C, Mols G, Elsasser S, et al: Extubation after breathing trials with automatic tube compensation, T-tube, or pressure support ventilation. *Acta Anaesthesiol Scand* 2002; 46:973–979
 50. Sassoon CS, Light RW, Lodia R, et al: Pressure-time product during continuous positive airway pressure, pressure support ventilation, and T-piece during weaning from mechanical ventilation. *Am Rev Respir Dis* 1991; 143:469–475
 51. Szklo M, Nieto FJ: *Epidemiology: Beyond the Basics*. Sudbury, MA, Jones and Bartlett Publishers, 2007, pp 151–181
 52. Brochard L, Pluskwa F, Lemaire F: Improved efficacy of spontaneous breathing with inspiratory pressure support. *Am Rev Respir Dis* 1987; 136:411–415
 53. Tokioka H, Saito S, Kosaka F: Effect of pressure support ventilation on breathing patterns and respiratory work. *Intensive Care Med* 1989; 15:491–494
 54. Jubran A, Van de Graaff WB, Tobin MJ: Variability of patient-ventilator interaction with pressure support ventilation in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 1995; 152:129–136
 55. Rivera L, Weissman C: Dynamic ventilatory characteristics during weaning in postoperative critically ill patients. *Anesth Analg* 1997; 84:1250–1255
 56. Straus C, Louis B, Isabey D, et al: Contribution of the endotracheal tube and the upper airway to breathing workload. *Am J Respir Crit Care Med* 1998; 157:23–30
 57. Ziman J: *An Introduction to Science Studies: The Philosophical and Social Aspects of Science and Technology*. Cambridge, Cambridge University Press, 1984
 58. Miettinen OS: Evidence in medicine: Invited commentary. *CMAJ* 1998; 158:215–221
 59. Rahn H, Otis AB, Chadwick LE, et al: The pressure-volume diagram of the thorax and lung. *Am J Physiol* 1946; 146:161–178
 60. Agostoni E, Mead J: Statics of the respiratory system. In: *Handbook of Physiology*. Fenn WO, Rahn H (Eds). Washington, DC, American Physiological Society, 1964, pp 387–409
 61. Vinegar A, Sinnett EE, Leith DE: Dynamic mechanisms determine functional residual capacity in mice, *Mus musculus*. *J Appl Physiol* 1979; 46:867–871
 62. Smith TC, Marini JJ: Impact of PEEP on lung mechanics and work of breathing in severe airflow obstruction. *J Appl Physiol* 1988; 65:1488–1499
 63. Mead J, Whittenberger JL: Evaluation of airway interruption technique as a method for measuring pulmonary airflow resistance. *J Appl Physiol* 1954; 6:408–416
 64. Frank NR, Mead J, Ferris BG Jr: The mechanical behavior of the lungs in healthy elderly persons. *J Clin Invest* 1957; 36:1680–1687
 65. West JB: *Pulmonary Pathophysiology: The Essentials*. Baltimore, Williams and Wilkins, 1978
 66. Otis AB, Fenn WO, Rahn H: Mechanics of breathing in man. *J Appl Physiol* 1950; 2:592–607
 67. Tobin MJ: Noninvasive monitoring of ventilation. In: *Principles and Practice of Intensive Care Monitoring*. Tobin MJ (Ed). New York, McGraw-Hill, 1998, pp 465–495
 68. Cook DJ, Mulrow CD, Haynes RB: Systematic reviews: Synthesis of best evidence for clinical decisions. *Ann Intern Med* 1997; 126:376–380
 69. Guyatt GH, Sackett DL, Sinclair JC, et al: Users' guides to the medical literature: IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. *JAMA* 1995; 274:1800–1804
 70. Ify L: Randomized clinical trials. *N Engl J Med* 1991; 325:1514
 71. Evidence based medicine: A new approach to teaching the practice of medicine. Evidence-Based Medicine Working Group. *JAMA* 1992; 268:2420–2425