Clinical Knowledge from Observational Studies

Everything You Wanted to Know but Were Afraid to Ask

Andrea S. Gershon[1]*, S. Reza Jafarzadeh[2]*, Kevin C. Wilson[3], and Allan J. Walkey[3]

[1]Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada; and [2]Clinical Epidemiology Research and Training Unit and [3]Department of Medicine, Boston University School of Medicine, Boston, Massachusetts

ORCID IDs: 0000-0002-1099-9175 (S.R.J.); 0000-0003-4429-2263 (K.C.W.); 0000-0003-4685-6894 (A.J.W.).

> *The statements of science are not of what is true and what is not true, but statements of what is known with different degrees of certainty.*
>
> —Richard Feynman, physicist

An explosion of new data sources has increased focus on the use of "real-world" information to inform patient-centered care. Enthusiasm for such a "learning health care system" is shared by many organizations, including the Institute of Medicine (1), the U.S. Food and Drug Administration (2), and Agency for Healthcare Research and Quality (3). How can we bridge the divide between a learning health care system's need for research that infers information from routine clinical care and the current hierarchies that place the results from tightly structured randomized controlled trials (RCTs) at the apex of evidence?

RCTs provide accurate estimates of average treatment effects for groups that differ only with respect to the intervention of interest when adherence is perfect to blinded and randomly allocated interventions across large numbers of patients (4). However, inferences derived from RCTs may not consistently be readily transferable to clinical practice. Strict inclusion and exclusion criteria can limit the group of people being tested to a degree that does not allow evaluation of the risks or benefits of the intervention in the population that will ultimately receive the intervention. In contrast, observational studies can compare outcomes among individuals who have been exposed to an intervention to outcomes among individuals who have not been exposed (or correlate different exposure levels to outcomes) and enable inferences regarding patients for whom we care over long periods of time.

Valid and reliable observational studies provide an attractive complement to RCTs in many circumstances. These include when the outcomes of interest are rare (e.g., safety outcomes, outcomes in rare diseases) or occur over a long time frame (5), or when randomization is considered unethical, prohibitively difficult (e.g., health services, policy, or quality-improvement interventions), or alters the study population. Despite the potential for observational studies to yield important information, clinicians tend to be reluctant to apply the results of observational studies into clinical practice. Methods of observational studies tend to be difficult to understand, and there is a common misconception that the validity of a study is determined entirely by the choice of study design (i.e., RCTs are always more "trustworthy" than observational studies) (Figure 1) rather than a comprehensive evaluation of the quality of individual studies. Herein, we explore situations that arise when comparing the results of randomized and observational studies of similar clinical research questions; we also develop an approach to applying research from both types of studies into clinical practice.

Agreement between RCTs and Observational Studies

Clinical interventions are rarely supported by strong certainty or high-quality evidence in the traditional hierarchy (Figure 1); rather, evidence is usually uncertain. For example, three critical care–focused American Thoracic Society clinical practice guidelines published in 2017 made 27 clinical recommendations, but only one (4% of recommendations) was a "strong" recommendation supported by "high-quality evidence" (a recommendation to use noninvasive ventilation in hypercapnic chronic obstructive pulmonary disease [COPD] exacerbations) (6–8). A common reason that evidence is considered uncertain is that studies evaluating similar research questions yield different results (i.e., inconsistency). RCTs may disagree with other RCTs (9, 10), and RCTs may disagree with observational studies. Conflicting evidence poses difficulties for guideline panels and complicates decision-making; however, it provides a myriad of learning opportunities.

(Received in original form January 18, 2018; accepted in final form May 4, 2018)

*Co–first authors.

Author Contributions: Conception, design, and drafting the manuscript: all authors.

Correspondence and requests for reprints should be addressed to Kevin C. Wilson, M.D., The Pulmonary Center, R-304, 80 East Concord Street, Boston, MA 02118. E-mail: kwilson@thoracic.org.

Am J Respir Crit Care Med Vol 198, Iss 7, pp 859–867, Oct 1, 2018

Originally Published in Press as DOI: 10.1164/rccm.201801-0118PP on May 7, 2018
Internet address: www.atsjournals.org

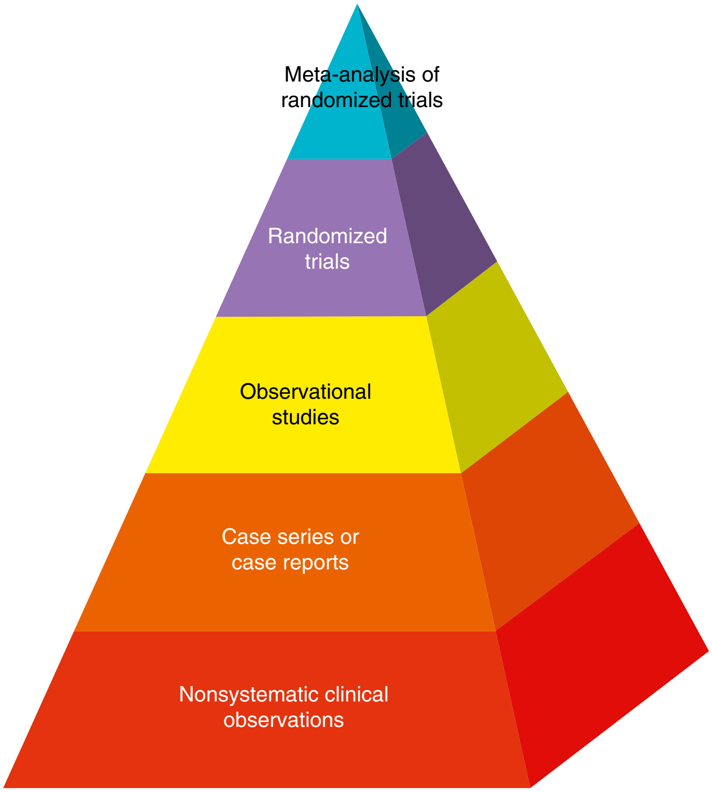
Meta-analysis of randomized trials
Randomized trials
Observational studies
Case series or case reports
Nonsystematic clinical observations

Figure 1. The hierarchy of evidence.

How Often Do Observational Studies Disagree with RCTs?

Different findings from observational studies and RCTs that evaluate similar research questions rarely result in the dismissal of the randomized trial design as a valid method of scientific inference; however, conflicts between RCTs and observational results often result in commentary regarding the times when observational studies "got it wrong" (11). Such discussions give the false impression that disagreement between RCTs and observational research is common; studies suggest the contrary (12–14).

A meta-analysis of 14 meta-analyses compared more than 1,000 pairs of observational studies and RCTs across 228 medical conditions and found that effect estimates from observational studies were not significantly different from those of RCTs (ratio of odds ratios, 1.08; 95% confidence interval [CI], 0.96–1.22) (12). The majority (71%) of the meta-analyses found minimal disagreement between RCTs and observational studies (12), indicating that on average observational studies and RCT results agree (15). Studies have concluded that discrepancies occur as frequently among observational study–randomized trial pairs as among randomized trial–randomized trial pairs (14).

When Observational Studies Do Not Agree with RCTs, Should I Always "Believe" RCTs over the Observational Studies?

Although evidence shows that observational studies and RCTs examining the same research questions usually agree, history has provided us with notable examples of times when they have not. Well-known examples include studies of hormone replacement therapy and cardiovascular disease for postmenopausal women (16, 17), statins and sepsis/acute respiratory distress syndrome (18, 19), procalcitonin to decrease antibiotic use in sepsis and COPD exacerbations (20–28), and activated protein C (29, 30). However, factors other than study design need to be considered when exploring reasons for a lack of agreement between RCTs and observational studies (12). When discrepancies exist, the observational study is not always wrong.

For example, initial observational studies and RCTs of normal saline and balanced fluid solutions in sepsis reached different conclusions. An observational study of 6,730 patients with sepsis concluded that balanced crystalloids were associated with a lower risk of mortality (relative risk [RR], 0.86; 95% CI, 0.78–0.94). However, a subsequent double-blind, randomized, cluster-crossover trial of 2,278 critically ill patients published a year later showed similar mortality estimates (RR, 0.88; 95% CI, 0.67–1.17) and concluded that larger studies were needed to evaluate the effects of crystalloid type on mortality (31). In 2018, a larger cluster-crossover trial of normal saline versus balanced crystalloid among critically ill patients showed a nearly identical effect estimate for mortality and renal dysfunction as the prior observational study and randomized trial among the subgroup of 2,336 patients with sepsis (odds ratio converted RR, 0.87; 95% CI, 0.78–0.97) (32). When faced with seemingly different results of observational studies and RCTs, the RCT cannot be uncritically relied on; instead, reasons that the results differed should be explored (in this case, the difference in study size).

What Can We Learn When Observational Studies and RCTs Disagree?

Disagreement between observational studies and RCTs can alert readers to flaws in internal validity and limitations in external validity (i.e., the degree to which results can be generalized to a target population, practice setting, intervention, comparator, or outcome beyond that which was specifically studied) of the studies.

Internal validity. Both observational studies and RCTs occasionally get things wrong. Table 1 outlines threats to the internal validity of observational studies. We focus on deficiencies of observational studies (unmeasured confounding variables, measurement error, prevalent user bias, and immortal person-time), because those of RCTs are extensively discussed elsewhere (33).

Confounding. When confounding occurs, there is systematic bias or an error in the measure of association between the exposure and the outcome because of the effect of another factor or confounder (34). Residual unmeasured confounding may exist even after adjustment of measured confounders. For example, nearly all observational studies of activated protein C identified lower mortality rates among patients who received activated protein C

Table 1. Factors to Consider When Observational Studies and Randomized Controlled Trials Disagree

Did the observational studies and RCTs ask the same question?

- Included the same population?
- Used the same intervention?
- Performed the same comparison?
- Measured outcomes in the same way?

If the answer is "No" to any of these questions, the studies did not ask the same research question and may complement, but not contradict, each other if both studies have high internal validity.

Was the internal validity of the observational studies optimized?

- Measured confounding: Did the study account for measured confounding variables using methods such as restriction, matching, stratification, standardization, multivariable adjustment, propensity score analysis, and/or generalized methods?
- Unmeasured confounding bias: Did the study account for unmeasured confounding variables using methods such as instrumental variables or regression discontinuity design? Alternatively, did the study simulate the effect of unmeasured confounders and the effect did not change interpretation (i.e., sensitivity analysis)? Did the study compare two interventions, rather than an intervention vs. nothing?
- Other selection biases: Were prevalent users excluded? Was immortal person-time avoided? Did the study enroll consecutive patients? Did the study enroll patients from the same population?
- Deviation bias: Were the interventions strictly adhered to? Were cointerventions likely due to patients or caregivers being aware of the intervention?
- Detection bias: Was the outcome assessor unaware of the intervention? Was the outcome a known effect of the intervention and more likely to be affected by confounding by indication, or an unknown effect (simulating blinding in a trial)?
- Attrition bias: Were all relevant data analyzed?
- Reporting bias: Were all of the measured outcomes reported?
- Other: Did the study have adequate power? Did the study justify multiple hypothesis tests?

If the answer is "No" to any of these questions, then low internal validity of the observational study may explain disagreements between results of observational studies and RCTs.

Was the internal validity of the RCTs optimized?

- Selection bias: Was there true random sequence generation that was blinded (i.e., concealed)?
- Performance bias: Were the patients and caregivers blinded?
- Detection bias: Were the outcome assessors blinded?
- Attrition bias: Was the data incomplete (i.e., loss to follow-up)?
- Reporting bias: Were all of the measured outcomes reported?
- Other: Did the trial have adequate power? Did the trial justify multiple hypothesis tests?

If the answer is "No" to any of these questions, then low internal validity of the randomized trial may explain disagreements between results of observational studies and RCTs.

Definition of abbreviation: RCT = randomized controlled trial.

Note that differences in achieving certain P value thresholds (e.g., $P < 0.05$) for similar study questions do not necessarily indicate disagreement between studies but, rather, may be a function of different study outcome rates and/or sample sizes. When evaluating "agreement" between studies, it is preferable to evaluate for similarity in effect estimates and inclusion of effect estimates within confidence intervals of the comparator studies.

during sepsis, whereas RCTs generally found it inefficacious. Later analyses determined that the observational studies were likely confounded by an inability to measure and analytically account for patient preferences to limit life-supporting interventions, which acted as confounders because they were common, associated with mortality, and produced a low likelihood of receiving activated protein C (35).

Selection bias. Selection bias in RCTs may occur if the study personnel influence the distribution of study interventions on the basis of perceived needs or if the participants in one study arm are more likely to be lost to follow-up than in another. A common cause of selection bias in observational studies occurs when studying the effectiveness of a treatment as compared with receipt of no treatment (i.e., confounding by indication). For example, if one was studying the effectiveness of a medication to treat COPD, one might use a database to identify groups of people with COPD who were and were not taking the medication and then determine how their outcomes differ. However, people are not usually started on medications unless they are sick, so the group receiving the medication is likely to be sicker than the one not receiving the medication and, thus, have worse health outcomes.

Prevalent user bias is related to prior use of the intervention of interest. Prevalent users have used the treatment being studied before the start of the study and are enrolled based on their adherence to the treatment. The fact that they have been using the treatment before study enrollment suggests that prevalent users have tolerated it and are more likely to be "healthy users" who have not developed adverse effects or outcomes. Inclusion of prevalent users generally biases results toward positive effects of the treatment under study. Observational studies including prevalent users are not comparable with most RCTs, in which follow-up starts at the beginning of treatment for all participants (36). For example, observational studies of statins and aspirin during critical illness may have been subject to prevalent user bias (34). As another example, reanalysis of observational studies of hormone replacement excluding prevalent users (and accounting for heterogeneity in treatment effects due to time from menopause) harmonized previously discrepant results between RCTs and observational studies (37). Including only new users of a treatment, which can be done in both observational studies and RCTs, allows detection of early adverse events and mitigates prevalent user bias.

Immortal person-time bias. In contrast to prevalent user bias, immortal person-time results from treatments started after the study begins. Take the example of a study that intends to compare death in patients who do and do not start a treatment after a hospitalization. If all patients in the treatment group are considered treated from the time they leave the hospital, but in fact they start treatment days or weeks after the hospitalization, they are essentially "immortal" for the time between the discharge and treatment start dates. Immortal person-time always favors the treatment of interest (38). There are many methods that can account for this common bias (39).

External validity. RCTs often use strict enrollment criteria to improve the likelihood of identifying efficacious interventions (40). Previous work has well documented how little RCT populations resemble real-world populations. For example, patients are often diagnosed with COPD in the real world without the use of spirometry, which differs from the situation in RCTs, where study participants have spirometry-confirmed COPD (41). This leads to a situation where some people who report being diagnosed with COPD by their physician do not have true airflow obstruction; thus, they differ from those evaluated in RCTs (42). As another example, observational studies may show lack of effectiveness and discrepant results to RCTs if there are differences in implementation of an intervention (e.g., poor fidelity to treatment algorithms), as shown by comparisons of studies investigating the effects of procalcitonin testing on antibiotic use in real-world and RCT settings (43–45). Enrolling study participants within narrow inclusion criteria in tightly controlled RCTs also may result in an increased risk of adverse events when the results are applied to frailer patients in real-world clinical practice (46). Long-term observational studies conducted outside of a trial setting may be better able to detect adverse effects.

Pairing RCTs and observational studies provides an opportunity to learn about treatment efficacy (47, 48) and effectiveness by comparing RCTs and observational studies, as well as rare side effects and problems with implementation. This allows for a more complete evaluation of interventions in the real world.

Maximizing the Quality of Observational Studies

What Are Some Methods Used to Strengthen the Internal Validity of an Observational Study?

Observational studies tend to have weaker internal validity than RCTs, generally related to confounding. Several strategies can be used to reduce confounding, and often more than one is used.

Confounding can be attenuated through proficient study design. For example, an observational study designed to look at unexpected outcomes minimizes confounding, because such outcomes are less likely to be associated with indications for treatment, severity of illness, and/or cointerventions. Likewise, studies that compare two interventions with similar indications (i.e., an active-comparator design) minimize confounding from selection bias, indication for treatment, prevalent users, and immortal time more than studies that compare an intervention to no intervention. For example, comparison of dopamine and norepinephrine in septic shock resulted in similar estimates of mortality between observational (RR, 1.09; 95% CI, 0.81–1.41) and RCTs (RR, 1.12; 95% CI, 1.01–1.20), with the observational study further enabling comparison of effects across clinically important subgroups not evaluated in RCTs (49, 50).

Restriction refers to restricting the comparator populations so that they are similar. For example, a study could focus on only hospitalized patients with COPD, increasing likelihood that the disease severity is similar in treatment and control groups. The limitation of this approach is that treatment effectiveness is only determined in people with severe disease, with generalizability problems similar to those encountered in some RCTs. A better design is to explore effect estimates across differing degrees of cohort restriction.

Matching involves pairing people between two comparison groups who have similar characteristics (51). Observational studies often match by age and sex but may also match by severity of disease and other factors. Matching can be one to one, one to two, or greater, depending on how many are in each group and how similar they are.

Stratification involves dividing the study population into strata on the basis of characteristics that have the potential to cause confounding and then conducting analyses in each stratum separately. For example, an investigator may include people with mild disease in one stratum, moderate disease in a second, and severe disease in a third. If results are consistent across strata, it indicates that disease severity did not likely alter treatment effectiveness.

Standardization or simple adjustment involves adding covariate-specific risk in the reference (i.e., unexposed) population to the covariate distribution in the study (i.e., exposed) population, yielding the expected risk in the study population if it had been unexposed. The exposed and unexposed groups then become independent of the confounder and comparable (52). Consider an example in which death is the outcome. If the age-specific mortality rates for a population of interest are known, they can be used to create a standard population that is used to compare age-adjusted mortality rates.

Multivariable adjustment or outcome regression is a statistical approach where one can theoretically look at the relationship of an exposure and a continuous, binary, or a time-to-event outcome. Unlike stratification, however, it allows simultaneously accounting for the effects of more than one confounding factor. The number of people with a study outcome often limits the number of variables that one can account for in a multivariable regression, with a rule of thumb that one confounding variable can be considered for every 10 people with an outcome.

Propensity score analysis is statistical technique that addresses some shortcomings of the multivariate adjustment method, particularly when there are many potential confounders and/or exposed and unexposed individuals are very different (53). A propensity score is the probability that an individual will be exposed (or will be allocated to a treatment group) based on his/her baseline characteristics (54). It can be used to adjust for confounders in several ways, including propensity score–adjustment, –stratification, or –matching (55). Each of these methods essentially creates groups of individuals that are balanced on the characteristics used to calculate the propensity scores that can then be compared.

Robins' generalized methods (i.e., g methods) are a family of statistical techniques developed to address time-dependent confounding in the presence of time-varying exposures (56). Time-dependent confounding occurs in longitudinal studies when factors that affect an outcome and an exposure are themselves affected by a previous exposure. For example, let us assume that we are interested in comparing the effects of two antihypertensive medications in patients with COPD on cardiovascular outcomes, where individuals' characteristics, blood pressure, and medication use are recorded longitudinally over several visits. A high blood pressure, recorded at a given visit, increases the risk of a cardiovascular adverse outcome and also creates an indication for antihypertensive use. This results in blood pressure confounding the association of subsequent antihypertensive use and the outcome, while at the same time it is itself affected by prior antihypertensive use. Because such factors (i.e., blood pressure level) are on the causal pathway between prior antihypertensive medication use and the outcome, traditional methods of adjusting for confounding would mask a portion of the effect of the preceding exposure on outcomes (57); therefore, none of the standard methods of adjustment for baseline confounding (e.g., restriction, matching, stratification, or regression adjustment in a linear, logistic, or Cox hazard model) can be used. G-methods include g-formula, marginal structural models that are estimated through inverse probability weighting of exposure or censoring, and structural nested models that are estimated by g-estimation technique (57). G-estimation was previously implemented to mitigate confounding by indication by including factors that predicted treatment initiation (58).

Are There Any Methods to Strengthen the Internal Validity of an Observational Study if the Confounders Are Unknown or Cannot Be Measured?

All the methods above address confounding when confounders can be measured. If confounders are unknown or cannot be measured, other approaches can be used to balanced exposure and control groups on measured and presumably also unmeasured variables.

An instrumental variable is a variable that, by design, correlates with the exposure but not with the outcome of interest (59). Dividing the study population by this variable creates groups that differ with respect to exposure, but not necessarily outcome, and, without the usual selection pressures, are presumably balanced on all other variables. In other words, if the outcome differs between the groups it is likely as a result of the exposure. For example, in a study of elderly patients hospitalized with an acute myocardial infarction, regional catheterization rate was the instrumental variable used to determine how cardiac catheterization impacted mortality (60). Instrumental variables are useful to counter confounding by indication (61).

Regression discontinuity design is used to estimate the effects of treatments on outcomes when treatments are initiated based on the threshold of a continuous variable (62, 63). For example, a post-bronchodilator FEV_1/FVC ratio of 0.70 is often used to confirm COPD and initiate therapy. In regression discontinuity designs, patients are quasi-randomized by the noise in this continuous variable—patients with a ratio of 0.69 likely do not differ from those with a ratio of 0.70, but they are more likely to be treated. Thus, it is most likely treatment and not patient characteristics that influence outcomes in patients with FEV_1/FVC ratio around the 0.70 threshold. This approach to reduce unmeasured confounding may also be used to evaluate health policy or health services interventions. For example, novel policies can be implemented on a threshold of a continuous risk score and studied using regression discontinuity designs (63). Time can be used as a continuous threshold on which to study an intervention, whereas interrupted time series designs (64) and difference-in-difference approaches (65) can be used to study new health services interventions by controlling for the confounding effects of secular trends.

Are There Ways to Explore Whether or Not Unmeasured Confounding Accounts for Results?

Sensitivity analyses may be used to determine how robust findings are to different methods of analysis or to the presence of unmeasured cofounding. One commonly used sensitivity analysis quantifies the extent of unmeasured confounding that would invalidate an observed association. It creates a hypothetical confounding variable that has sufficient strength and prevalence to explain away the results. A healthcare provider, using his or her clinical knowledge, can then judge the likelihood such a variable exists (and is not already controlled for). This approach has been formalized in a measure called the E-value (66–68).

Another sensitivity analysis to assess for unmeasured confounding examines the association between the intervention and a tracer condition. This is a condition associated with potential confounders but, by design, not the intervention of interest. If an association is found, it suggests that unmeasured confounding is present. For example, in a study comparing two COPD medications, if there were more upper gastrointestinal bleeds (a condition not likely to be associated with COPD medication use) in one group than the other group after adjustment for all covariates, it would suggest one group was more frail than the other and prone to poor health outcomes, including those related to COPD.

Are There Any Other Uses for Sensitivity Analyses?

Sensitivity analyses are also useful for detecting measurement error or misclassification. Observational studies often use secondary data sources (i.e., data not collected for research but reasons such as billing or healthcare delivery), where recorded services may not fully represent diseases or the treatments or exposures received by patients. For example, people identified as having COPD using health administrative data may not have received spirometry, because a large proportion of real-world patients do not (41). To address uncertainty, an imperfect COPD indicator, such as "physician-diagnosed COPD" (69), may be used.

When one uses an alternative to a gold standard measure, its accuracy should be established through a validation study (70). Poor correlation to the gold standard, however, does not necessarily negate its value if it is something that is of value to patients, physicians, and decision-makers. Rather, differences should be clearly delineated so that readers can evaluate its value. Indeed, the real-world nature of a secondary data source measure can be a strength, depending on the research question; in our example, a large proportion of people with physician-diagnosed COPD do not receive spirometry, yet they still contribute significantly to disease burden, take medications, get admitted to the hospital, and so on, which reflects real-world circumstances.

When differences between an alternate measure and a gold standard exist or a relevant validation study has not been done, uncertainty can be explored through a sensitivity analysis. For example, one could determine the hypothetical degree of misclassification that would have to occur (i.e., between diagnoses of COPD and asthma) to explain away the results found in a study (71). Then one could turn to other data or clinical experience to judge whether this seems likely. There are many approaches to sensitivity analysis for misclassification bias, including sensitivity-specificity imputation, positive and negative predictive values direct imputation, and probabilistic bias analysis by Monte Carlo or Bayesian analysis (72).

How Do You Assess the Quality of an Observational Study?

Various instruments have been developed to help consumers of medical literature identify high-quality observational studies, including the Risk of Bias in Non-randomized Studies—of Interventions (ROBINS-I) (73), Good Research for Comparative Effectiveness (GRACE) (74), and Methodological Evaluation of Observational Research (MORE) (75) assessment tools. The ROBINS-I tool is a representative instrument that evaluates seven types of risk of bias that may occur within observational studies, including: *1*) bias due to confounders, *2*) selection bias (selection of subjects who are different from those whom the investigators want to study), *3*) classification bias (due to misclassification), *4*) bias due to deviation from the intended interventions, *5*) bias due to missing data, *6*) bias due to outcome measurement (assessors are aware of the intervention status, different methods are used to measure outcomes in different groups), and *7*) reporting bias (selective reporting of outcomes).

How Do You Decide Which Studies to Believe?

When Results of Observational Studies and RCTs Disagree

First, the clinician should ask whether the RCTs and observational studies asked the same questions. The answer is “no” if they included different populations, used different interventions, made different comparisons, or measured different outcomes. Second, the clinician should look at the quality and internal validity of each and determine if any biases are present. In addition to recommendations in this manuscript (Table 1), tools to assess methodological quality and bias can help (73–75). With thoughtful consideration of study details, clinicians can make informed decisions about whether and how to incorporate study findings into patient care.

When Results of Observational Studies and RCTs Agree

Although much of our discussion has focused on what we can learn from discordant observational studies and RCTs, concordant results are also important. Concordance demonstrates that treatments are efficacious (or not) in controlled experiments and effective (or not) in the messiness of routine clinical practice. We propose that such concordance represents the best evidence (Figure 2). Some examples of pulmonary and critical care interventions that show similar treatment effects regardless of study design include benefits of dual inhaled corticosteroid/long-acting bronchodilators to reduce asthma exacerbations (76, 77), noninvasive ventilation to reduce morality in acute hypercarbic respiratory failure (8), and hypothermia for improved outcomes of out-of-hospital cardiac arrest (78).

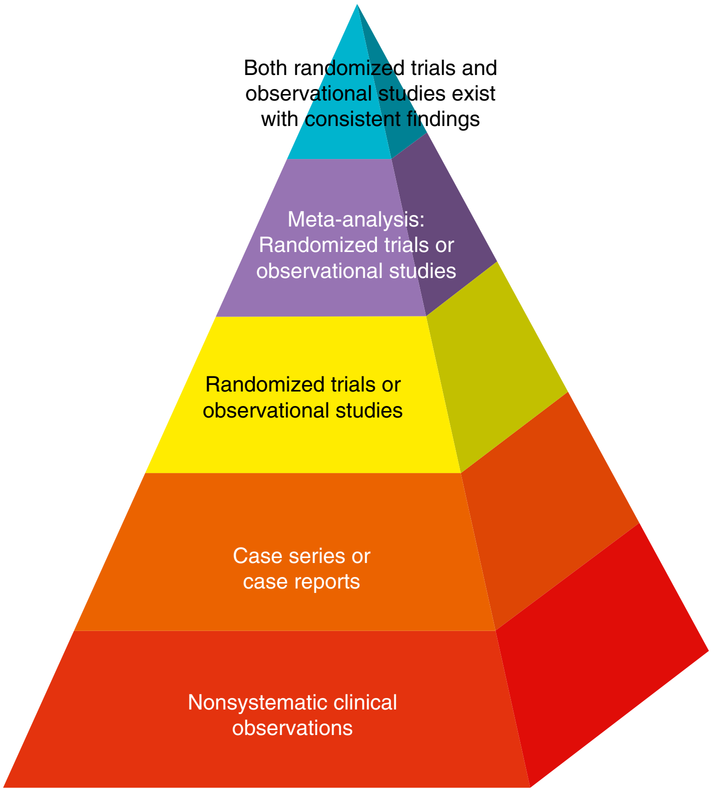
Both randomized trials and observational studies exist with consistent findings
Meta-analysis: Randomized trials or observational studies
Randomized trials or observational studies
Case series or case reports
Nonsystematic clinical observations

Figure 2. Proposed new hierarchy of evidence.

Evidence Synthesis and Guideline Development

Standards for the development of clinical practice guidelines, established by the Institute of Medicine (79), state that every question in a guideline requires a full systematic review of evidence to inform recommendations. However, adhering to these standards is time consuming and burdensome (79–81). In response, many organizations prioritize RCTs, which, if identified, cease the search for observational evidence. The effect is to bias guideline recommendations in favor of interventions that are efficacious, regardless of whether they have proven effective. This well-intentioned pragmatic approach may result in erroneous or no recommendations being made when RCTs and observational studies disagree and when high-quality observational evidence is overlooked. Guideline developers should appraise all relevant evidence and use the best-quality studies, whether randomized or observational, ideally using both to inform their recommendations.

Conclusions

Observational studies should be considered complementary, rather than inferior, to RCTs. High-quality observational studies provide valid estimates of exposure/treatment effects and may be more generalizable. Similarly, high-quality RCTs provide unbiased estimates of treatment effects but are often less generalizable. These trade-offs are not necessary when high-quality observational studies and RCTs coexist. We propose that the hierarchy of evidence be reconsidered and the existence of multiple concordant observational studies and RCTs increase certainty in clinical evidence (Figure 2). We further propose an approach when observational studies and RCTs disagree (Table 1). ■

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

References

1. Institute of Medicine. Integrating research and practice: health system leaders working toward high-value care: workshop summary. Washington, D.C.: The National Academies Press; 2015.
2. Califf RM, Sherman R. What we mean when we talk about data. 2015 Dec 10 [accessed 2018 Mar 15]. Available from: https://blogs.fda.gov/fdavoice/?s=What+we+mean+when+we+talk+about+data.
3. Agency for Healthcare Research and Quality. AHRQ research summit on learning health systems. 2017 Sept 15 [accessed 2018 Mar 15]. Available from: https://www.ahrq.gov/news/events/ahrq-research-summit-learning-health-system.html.
4. Mansournia MA, Higgins JP, Sterne JA, Hernán MA. Biases in randomized trials: a conversation between trialists and epidemiologists. *Epidemiology* 2017;28:54–59.
5. Manson JE, Shufelt CL, Robins JM. The potential for postrandomization confounding in randomized clinical trials. *JAMA* 2016:315:2273–2274.
6. Schmidt GA, Girard TD, Kress JP, Morris PE, Ouellette DR, Alhazzani W, *et al*.; ATS/CHEST Ad Hoc Committee on Liberation from Mechanical Ventilation in Adults. Official executive summary of an American Thoracic Society/American College of Chest Physicians clinical practice guideline: liberation from mechanical ventilation in critically ill adults. *Am J Respir Crit Care Med* 2017;195:115–119.
7. Fan E, Del Sorbo L, Goligher EC, Hodgson CL, Munshi L, Walkey AJ, *et al*.; American Thoracic Society, European Society of Intensive Care Medicine, and Society of Critical Care Medicine. An official American Thoracic Society/European Society of Intensive Care Medicine/Society of Critical Care Medicine clinical practice guideline: mechanical ventilation in adult patients with acute respiratory distress syndrome. *Am J Respir Crit Care Med* 2017;195:1253–1263.
8. Rochwerg B, Brochard L, Elliott MW, Hess D, Hill NS, Nava S, *et al*. Official ERS/ATS clinical practice guidelines: noninvasive ventilation for acute respiratory failure. *Eur Respir J* 2017;50:1602426.
9. Esteban A, Alía I, Gordo F, Fernández R, Solsona JF, Vallverdú I, *et al*.; The Spanish Lung Failure Collaborative Group. Extubation outcome after spontaneous breathing trials with T-tube or pressure support ventilation. *Am J Respir Crit Care Med* 1997;156:459–465.
10. Brochard L, Rauss A, Benito S, Conti G, Mancebo J, Rekik N, *et al*. Comparison of three methods of gradual withdrawal from ventilatory support during weaning from mechanical ventilation. *Am J Respir Crit Care Med* 1994;150:896–903.
11. Albert RK. “Lies, damned lies ...” and observational studies in comparative effectiveness research. *Am J Respir Crit Care Med* 2013;187:1173–1177.
12. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev* 2014;4:MR000034.
13. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;342:1887–1892.
14. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, *et al*. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;286:821–830.
15. Franklin JM, Dejene S, Huybrechts KF, Wang SV, Kulldorff M, Rothman KJ. A bias in the evaluation of bias comparing randomized trials with nonexperimental studies. *Epidemiol Methods* 2017;6:20160018.
16. Stampfer MJ, Colditz GA, Willett WC, Manson JE, Rosner B, Speizer FE, *et al*. Postmenopausal estrogen therapy and cardiovascular disease: ten-year follow-up from the nurses’ health study. *N Engl J Med* 1991;325:756–762.
17. Manson JE, Hsia J, Johnson KC, Rossouw JE, Assaf AR, Lasser NL, *et al*.; Women’s Health Initiative Investigators. Estrogen plus progestin and the risk of coronary heart disease. *N Engl J Med* 2003; 349:523–534.
18. Falagas ME, Makris GC, Matthaiou DK, Rafailidis PI. Statins for infection and sepsis: a systematic review of the clinical evidence. *J Antimicrob Chemother* 2008;61:774–785.
19. Truwit JD, Bernard GR, Steingrub J, Matthay MA, Liu K, Albertson TE, *et al*. SAILS: statins for acutely injured lungs (ARDS) from sepsis [abstract]. *Am J Respir Crit Care Med* 2014;189:A5088.
20. Christ-Crain M, Jaccard-Stolz D, Bingisser R, Gencay MM, Huber PR, Tamm M, *et al*. Effect of procalcitonin-guided treatment on antibiotic use and outcome in lower respiratory tract infections: cluster-randomised, single-blinded intervention trial. *Lancet* 2004;363: 600–607.
21. Christ-Crain M, Stolz D, Bingisser R, Müller C, Miedinger D, Huber PR, *et al*. Procalcitonin guidance of antibiotic therapy in community-acquired pneumonia: a randomized trial. *Am J Respir Crit Care Med* 2006;174:84–93.
22. Schuetz P, Christ-Crain M, Thomann R, Falconnier C, Wolbers M, Widmer I, *et al*.; ProHOSP Study Group. Effect of procalcitonin-based guidelines vs standard guidelines on antibiotic use in lower respiratory tract infections: the ProHOSP randomized controlled trial. *JAMA* 2009;302:1059–1066.
23. de Jong E, van Oers JA, Beishuizen A, Vos P, Vermeijden WJ, Haas LE, *et al*. Efficacy and safety of procalcitonin guidance in reducing the duration of antibiotic treatment in critically ill patients: a randomised, controlled, open-label trial. *Lancet Infect Dis* 2016;16:819–827.
24. Nobre V, Harbarth S, Graf JD, Rohner P, Pugin J. Use of procalcitonin to shorten antibiotic treatment duration in septic patients: a randomized trial. *Am J Respir Crit Care Med* 2008;177:498–505.
25. Bouadma L, Luyt C-E, Tubach F, Cracco C, Alvarez A, Schwebel C, *et al*.; PRORATA trial group. Use of procalcitonin to reduce patients’ exposure to antibiotics in intensive care units (PRORATA trial): a multicentre randomised controlled trial. *Lancet* 2010;375:463–474.
26. Hochreiter M, Köhler T, Schweiger AM, Keck FS, Bein B, von Spiegel T, *et al*. Procalcitonin to guide duration of antibiotic therapy in intensive care patients: a randomized prospective controlled trial. *Crit Care* 2009;13:R83.
27. Kopterides P, Siempos II, Tsangaris I, Tsantes A, Armaganidis A. Procalcitonin-guided algorithms of antibiotic therapy in the intensive care unit: a systematic review and meta-analysis of randomized controlled trials. *Crit Care Med* 2010;38:2229–2241.
28. Hohn A, Schroeder S, Gehrt A, Bernhardt K, Bein B, Wegscheider K, *et al*. Procalcitonin-guided algorithm to reduce length of antibiotic therapy in patients with severe sepsis and septic shock. *BMC Infect Dis* 2013;13:158.

29. Kalil AC, LaRosa SP. Effectiveness and safety of drotrecogin alfa (activated) for severe sepsis: a meta-analysis and metaregression. *Lancet Infect Dis* 2012;12:678–686.
30. Lai PS, Matteau A, Iddriss A, Hawes JC, Ranieri V, Thompson BT. An updated meta-analysis to understand the variable efficacy of drotrecogin alfa (activated) in severe sepsis and septic shock. *Minerva Anestesiol* 2013;79:33–43.
31. Young P, Bailey M, Beasley R, Henderson S, Mackle D, McArthur C, *et al*.; SPLIT Investigators; ANZICS CTG. Effect of a buffered crystalloid solution vs saline on acute kidney injury among patients in the intensive care unit: the SPLIT randomized clinical trial. *JAMA* 2015;314:1701–1710.
32. Semler MW, Self WH, Wanderer JP, Ehrenfeld JM, Wang L, Byrne DW, *et al*.; SMART Investigators and the Pragmatic Critical Care Research Group. Balanced crystalloids versus saline in critically ill adults. *N Engl J Med* 2018;378:829–839.
33. Higgins JPT, Altman DG, Sterne JAC, editors. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Version 5.1.0 [updated 2011 Mar; accessed 2018 Aug 27]. The Cochrane Collaboration; 2011. Available from: www.handbook.cochrane.org.
34. Kor DJ, Carter RE, Park PK, Festic E, Banner-Goodspeed VM, Hinds R, *et al*.; US Critical Illness and Injury Trials Group: Lung Injury Prevention with Aspirin Study Group (USCIITG: LIPS-A). Effect of aspirin on development of ARDS in at-risk patients presenting to the emergency department: the LIPS-A randomized clinical trial. *JAMA* 2016;315:2406–2414.
35. Bradford MA, Lindenauer PK, Wiener RS, Walkey AJ. Do-not-resuscitate status and observational comparative effectiveness research in patients with septic shock. *Crit Care Med* 2014;42:2042–2047.
36. Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, *et al*. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 2008;19:766–779.
37. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol* 1986;15:413–419.
38. Suissa S. Effectiveness of inhaled corticosteroids in chronic obstructive pulmonary disease: immortal time bias in observational studies. *Am J Respir Crit Care Med* 2003;168:49–53.
39. Suissa S. Immortal time bias in pharmaco-epidemiology. *Am J Epidemiol* 2008;167:492–499.
40. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol* 2013;42:1012–1014.
41. Gershon AS, Hwee J, Croxford R, Aaron SD, To T. Patient and physician factors associated with pulmonary function testing for COPD: a population study. *Chest* 2014;145:272–281.
42. Halpin DMG, Kerkhof M, Soriano JB, Mikkelsen H, Price DB. Eligibility of real-life patients with COPD for inclusion in trials of inhaled long-acting bronchodilator therapy. *Respir Res* 2016;17:120.
43. Lindenauer PK, Shieh MS, Stefan MS, Fisher KA, Haessler SD, Pekow PS, *et al*. Hospital procalcitonin testing and antibiotic treatment of patients admitted for COPD exacerbation. *Ann Am Thorac Soc* 2017; 14:1779–1785.
44. Chu DC, Mehta AB, Walkey AJ. Practice patterns and outcomes associated with procalcitonin use in critically ill patients with sepsis. *Clin Infect Dis* 2017;64:1509–1515.
45. Fisher KA, Landyn V, Lindenauer PK, Walkey AJ. procalcitonin test availability: a survey of acute care hospitals in Massachusetts. *Ann Am Thorac Soc* 2017;14:1489–1491.
46. Juurlink DN, Mamdani MM, Lee DS, Kopp A, Austin PC, Laupacis A, *et al*. Rates of hyperkalemia after publication of the Randomized Aldactone Evaluation Study. *N Engl J Med* 2004;351:543–551.
47. Kent DM, Kitsios G. Against pragmatism: on efficacy, effectiveness and the real world. *Trials* 2009;10:48.
48. Rothman KJ. Six persistent research misconceptions. *J Gen Intern Med* 2014;29:1060–1064.
49. Fawzy A, Evans SR, Walkey AJ. Practice patterns and outcomes associated with choice of initial vasopressor therapy for septic shock. *Crit Care Med* 2015;43:2141–2146.
50. De Backer D, Aldecoa C, Njimi H, Vincent JL. Dopamine versus norepinephrine in the treatment of septic shock: a meta-analysis. *Crit Care Med* 2012;40:725–730.
51. Mansournia MA, Hernán MA, Greenland S. Matched designs and causal diagrams. *Int J Epidemiol* 2013;42:860–869.
52. Keiding N, Clayton D. Standardization and control for confounding in observational studies: a historical perspective. *Stat Sci* 2014;29: 529–558.
53. Vansteelandt S, Daniel RM. On regression adjustment for the propensity score. *Stat Med* 2014;33:4053–4072.
54. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70: 41–55.
55. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med* 2010;29: 2137–2148.
56. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. *Int J Epidemiol* 2017;46:756–762.
57. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JA. Methods for dealing with time-dependent confounding. *Stat Med* 2013;32:1584–1618.
58. Joffe MM, Hoover DR, Jacobson LP, Kingsley L, Chmiel JS, Visscher BR. Effect of treatment with zidovudine on subsequent incidence of Kaposi's sarcoma. *Clin Infect Dis* 1997;25:1125–1133.
59. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006;17:360–372.
60. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA* 2007;297:278–285.
61. Joffe MM. Confounding by indication: the case of calcium channel blockers. *Pharmacoepidemiol Drug Saf* 2000;9:37–41.
62. Moscoe E, Bor J, Bärnighausen T. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. *J Clin Epidemiol* 2015;68:122–133.
63. Walkey AJ, Cordella N, Drainoni ML, Bor J. Advancing quality improvement with regression discontinuity designs. *Ann Am Thorac Soc* 2018;15:523–529.
64. Walkey AJ, Drainoni ML, Cordella N, Bor J. Advancing quality improvement with regression discontinuity designs. *Ann Am Thorac Soc* 2018;15:523–529.
65. Penfold RB, Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. *Acad Pediatr* 2013;13: S38–S44.
66. Wing C, Simon K, Bello-Gomez RA. Designing difference in difference studies: best practices for public health policy research. *Annu Rev Public Health* 2018;39:453–469.
67. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med* 2017;167:268–274.
68. Mathur MB, Ding P, Riddell CA, VanderWeele TJ. Website and R package for computing E-values. *Epidemiology* [online ahead of print] 14 Jun 2018 DOI: 10.1097/EDE.0000000000000864.
69. Gershon AS, Warner L, Cascagnette P, Victor JC, To T. Lifetime risk of developing chronic obstructive pulmonary disease: a longitudinal population study. *Lancet* 2011;378:991–996.
70. Gershon AS, Wang C, Guan J, Vasilevska-Ristovska J, Cicutto L, To T. Identifying individuals with physcian diagnosed COPD in health administrative databases. *COPD* 2009;6:388–394.
71. Gershon AS, Campitelli MA, Croxford R, Stanbrook MB, To T, Upshur R, *et al*. Combination long-acting β-agonists and inhaled corticosteroids compared with long-acting β-agonists alone in older adults with chronic obstructive pulmonary disease. *JAMA* 2014;312:1114–1121.
72. Corbin M, Haslett S, Pearce N, Maule M, Greenland S. A comparison of sensitivity-specificity imputation, direct imputation and fully Bayesian analysis to adjust for exposure misclassification when validation data are unavailable. *Int J Epidemiol* 2017;46:1063–1072.

73. Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, *et al*. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355: i4919.
74. Good Research for Comparative Effectiveness (GRACE) [accessed 2018 Apr 12]. Available from: https://www.graceprinciples.org.
75. Shamliyan TA, Kane RL, Ansari MT, Raman G, Berkman ND, Grant M, *et al*. Development of quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: pilot study of new checklists. Rockville, MD: Agency for Healthcare Research and Quality (US); 2011 Jan [accessed 2018 Aug 27]. Appendix A, Methodological evaluation of observational research. Available from: https://www.ncbi.nlm.nih.gov/books/NBK53279/.
76. Hirst C, Calingaert B, Stanford R, Castellsague J. Use of long-acting beta-agonists and inhaled steroids in asthma: meta-analysis of observational studies. *J Asthma* 2010;47:439–446.
77. Loymans RJ, Gemperli A, Cohen J, Rubinstein SM, Sterk PJ, Reddel HK, *et al*. Comparative effectiveness of long term drug treatment strategies to prevent asthma exacerbations: network meta-analysis. *BMJ* 2014;348:g3009.
78. Kitsios GD, Dahabreh IJ, Callahan S, Paulus JK, Campagna AC, Dargin JM. Can we trust observational studies using propensity scores in the critical care literature? A systematic comparison with randomized clinical trials. *Crit Care Med* 2015;43:1870–1879.
79. Institute of Medicine. Clinical practice guidelines we can trust. Washington, D.C.: The National Academies Press; 2011.
80. Institute of Medicine. Finding what works in health care: standards for systematic reviews. Washington, D.C.: The National Academies Press; 2011.
81. Schoenberg NC, Barker AF, Bernardo J, Deterding RR, Ellner JJ, Hess DR, *et al*. A comparative analysis of pulmonary and critical care medicine guideline development methodologies. *Am J Respir Crit Care Med* 2017;196:621–627.