incorrect or exaggerated results,[11] and the suggestion that the durability of medical knowledge is unrelated to methodologic quality.[12]

Even the best observational study is limited by an inability to draw causal inferences and by the presence of confounders. RCT design takes causality as a given and puts its trust in an ability to minimize—of course it does not eliminate—confounders by randomization. But the problem of "unknown unknowns" remains, and the greater the number of unknown confounders that exist, the greater the likelihood of an imbalance. This problem is common to RCTs and observational studies alike and is probably most likely in small studies where our understanding of disease pathogenesis is limited. In a study with total n ≈ 1,600, where five independent confounders exist, each with an incidence of 20%, the probability of an imbalance for at least one confounder is almost 25%.[13] So studies A and B might disagree because A has greater balance of unknown confounders than B, and thus a better balance of confounders in a large observational study might "trump" randomization in a small RCT. This does not upgrade the status of observational studies, but it does explain why well-designed observational studies often arrive at similar conclusions relative to RCTs, and why some of the time they will correctly contradict previous RCT data. The controversial articles by Karkouti *et al.*[4] and Mangano *et al.*[5] may exemplify this—as suggested by the results of the recent Blood Conservation Using Antifibrinolytics in a Randomized Trial.[14]

The article of Vincent *et al.* discusses whether leukoreduction might account for the findings but provides no data[2]; the editorial does not mention it.[1] Neither the original article nor the editorial provides any convincing explanation (*i.e.*, biologic basis) for the reported effect. We wonder whether additional analysis of the data in the article of Vincent *et al.*[2] might shed light on whether leukoreduction may be responsible for the apparently altered impact of transfusion, as has been suggested previously.[15,16]

The data of Vincent *et al.*[2] and the recent TRICC reanalysis by Deans *et al.*[9] suggest that outcome is changing over time and that the interpretation of the TRICC trial is more complex than we thought. It will be some time before we get a clearer picture, but in the meantime, we should not treat propensity scoring as a straw man. Reading the article of Vincent *et al.*,[2] we experience the judgment under uncertainty that pervades clinical life. Decisions to transfuse—and not to transfuse—are not made lightly, so it is a truism that these data should be viewed with caution. The function of the article, however, is to make us view with caution things that we think we know.

John F. Boylan, M.B., F.R.C.P.C.,* Brian P. Kavanagh, M.B., F.R.C.P.C. *St. Vincent's University Hospital, Dublin, Ireland. jboylan@iol.ie

References

1. Nuttall GA, Houle TT: Liars, damn liars, and propensity scores. ANESTHESIOLOGY 2008; 108:3–4

2. Vincent JL, Sakr Y, Sprung C, Harboe S, Damas P, on behalf of the Sepsis Occurrence in Acutely Ill Patients (SOAP) Investigators: Are blood transfusions associated with greater mortality rates? Results of the Sepsis Occurrence in Acutely Ill Patients Study. ANESTHESIOLOGY 2008; 108:31–9

3. Connors AF Jr, Speroff T, Dawson NV, Thomas C, Harrell FE Jr, Wagner D, Desbiens N, Goldman L, Wu AW, Califf RM, Fulkerson WJ Jr, Vidaillet H, Broste S, Bellamy P, Lynn J, Knaus WA: The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. JAMA 1996; 276:889–97

4. Karkouti K, Beattie WS, Dattilo KM, McCluskey SA, Ghannam M, Hamdy A, Wijeysundera DN, Fedorko L, Yau TM: A propensity score case-control comparison of aprotinin and tranexamic acid in high-transfusion-risk cardiac surgery. Transfusion 2006; 46:327–38

5. Mangano DT, Tudor IC, Dietzel C: The risk associated with aprotinin in cardiac surgery. N Engl J Med 2006; 354:353–65

6. Hébert PC, Wells G, Blajchman MA, Marshall J, Martin C, Pagliarello G, Tweeddale M, Schweitzer I, Yetisir E: A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. Transfusion Requirements in Critical Care Investigators, Canadian Critical Care Trials Group. N Engl J Med 1999; 340:409–17

7. Pitt B, Zannad F, Remme WJ, Cody R, Castaigne A, Perez A, Palensky J, Wittes J: The effect of spironolactone on morbidity and mortality in patients with severe heart failure. Randomized Aldactone Evaluation Study Investigators. N Engl J Med 1999; 341:709–17

8. Juurlink DN, Mamdani MM, Lee DS, Kopp A, Austin PC, Laupacis A, Redelmeier DA: Rates of hyperkalemia after publication of the Randomized Aldactone Evaluation Study. N Engl J Med 2004; 351:543–51

9. Deans KJ, Minneci PC, Suffredini AF, Danner RL, Hoffman WD, Ciu X, Klein HG, Schechter AN, Banks SM, Eichacker PQ, Natanson C: Randomization in clinical trials of titrated therapies: Unintended consequences of using fixed treatment protocols. Crit Care Med 2007; 35:1509–16

10. Benson K, Hartz AJ: A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000; 342:1878–86

11. Ioannidis JP: Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005; 294:218–28

12. Poynard T, Munteanu M, Ratziu V, Benhamou Y, Di Martino V, Taieb J, Opolon P: Truth survival in clinical research: An evidence-based requiem? Ann Intern Med 2002; 136:888–95

13. Shrier I, Platt RW, Steele RJ: Mega-trials *versus* meta-analysis: Precision *versus* heterogeneity? Contemporary Clin Trials 2007; 28:324–8

14. Fergusson DA, Hébert PC, Mazer CD, Fremes S, MacAdams C, Murkin JM, Teoh K, Duke PC, Arellano R, Blajchman MA, Bussières JS, Côté D, Karski J, Martineau R, Robblee JA, Rodger M, Wells G, Clinch J, Pretorius R, for the BART Investigators: A comparison of aprotinin and lysine analogues in high-risk cardiac surgery. N Engl J Med 2008; 358:2319–31

15. Hébert PC, Fergusson D, Blajchman MA, Wells GA, Kmetic A, Coyle D, Heddle N, Germain M, Goldman M, Toye B, Schweitzer I, vanWalraven C, Devine D, Sher GD: Clinical outcomes following institution of the Canadian universal leukoreduction program for red blood cell transfusions. JAMA 2003; 289:1941–9

16. Fergusson D, Hébert PC, Lee SK, Walker CR, Barrington KJ, Joseph L, Blajchman MA, Shapiro S: Clinical outcomes following institution of universal leukoreduction of blood transfusions for premature infants. JAMA 2003; 289: 1950–6

(Accepted for publication April 30, 2008.)

Anesthesiology 2008; 109:746–7

Propensity Scores Do Not Necessarily Lie!

To the Editor:—Recently, an Editorial View was published on propensity score methods.[1] The editorial describes strengths and weaknesses of propensity score methods in observational therapeutic studies. The authors apparently refer in their title to a quote said by the English Prime Minister Benjamin Disraeli (1804–1880) in the 19th century: "There are lies, damn lies and statistics." In general, we appreciate links to statements from outside the clinical research world. However, the title of the Editorial View may be misinterpreted as a statement against propensity scores. Readers of ANESTHESIOLOGY in general are not professional statisticians and may be reluctant to use propensity scores, even in appropriate situations, because of such a title.

The editorial is of value because it reviews an important problem of observational therapeutic studies. In such studies, investigators do not have control over who is or is not receiving the index treatment, which potentially results in imbalance of prognostic factors across the treatment arms. In the absence of randomization, treatment indication and assignment are typically related to the prognosis of the patient. For example, more patients with advanced disease may be given the index treatment than patients with early disease stages. As a consequence,

Supported by grant No. ZON-MW 917.46.360 from The Netherlands Organization for Scientific Research, Den Haag, The Netherlands.

the estimated treatment effect can be biased. This is known as confounding by indication and can be adjusted for in the statistical analysis. However, adjustments can be made only for prognostic factors that were measured in the study. Prognostic factors that were not measured may introduce hidden bias, for which adjustment is not possible. Any statistical method that aims to adjust for confounding by indication suffers from this problem, which is by no means restricted to propensity score methods! Propensity score methods may even have particular advantages over other correction methods. Therefore, the chosen title of the Editorial View was in our view very unfortunate.[1]

Prognostic factors can influence the treatment effect only if the factors are related both to the patient outcome and to the assignment of treatment. This implies that two different analytical strategies are possible. Conventionally, the measured prognostic factors are directly included in a regression model together with the assigned treatment and with the patient outcome as a dependent variable (treatment model). The propensity scores method contains two steps. First, the focus is on the association between the assigned treatment (dependent variable) and the prognostic factors, to develop a so-called propensity score. The propensity score predicts the probability of having received the index treatment based on the prognostic factors. Second, the focus is on the association between the patient outcome and the prognostic factors included as one combined variable (*i.e.*, the propensity score) together with the assigned treatment. The propensity score is here used to adjust the treatment effect for all prognostic factors.[2]

Nuttall *et al.* seem to suggest that both analytical methods are equally insufficient. We like to stress that propensity score methods have particular advantages when the outcome event is rare, the treatment is common, and many prognostic factors are collected.[3] The low number of outcome events in fact limits the number of prognostic factors that can be included in the conventional treatment model. A low ratio of "number of events over number of included factors" jeopardizes proper estimation of the treatment effect in the regression analysis. In contrast, the numbers of patients in the two treatment groups are generally high. This allows for adequate modeling of the association between the treatment assignment and many prognostic factors—a high ratio of "number of patients with the treatment over number of included factors." Subsequently, the treatment model includes only the assigned treatment and the propensity score, allowing for a proper and adjusted estimation of the treatment effect, despite the low number of outcome events. The efficiency of propensity scores in relation to the number of outcome events has been shown in a previous study, where propensity scores were found to produce less biased, more robust, and more precise estimates when fewer than seven events were available for each prognostic factor.[4]

Like any other correction method in observational therapeutic studies, propensity scores cannot control for hidden bias. However, sensitivity analysis has been proposed to indicate the magnitude of hidden bias that should be present to alter the conclusion of the study.[5] Furthermore, propensity scores cannot fix other potential methodologic bias, as discussed by Nuttall *et al.*, which again applies also to the conventional approach. Propensity scores do not pretend to solve these problems. Hence, propensity scores can not be considered as "liars."

In conclusion, Nuttall *et al.* discussed confounding by indication as an important weakness of observational therapeutic studies. However, when for ethical, economical, or practical reasons randomized trials can not be conducted, observational studies are the only appropriate alternative.[6] Imbalance in prognostic factors can be adjusted for in the analysis. Particularly when the number of outcome events is small, propensity score methods can more efficiently adjust for the imbalance than can conventional methods. Sensitivity analysis may complete the statistical analysis to study possible effects of hidden bias.

Yvonne Vergouwe, Ph.D.,* Wilton A. van Klei, M.D., Cor J. Kalkman, M.D., Karel G. M. Moons, Ph.D. *Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, The Netherlands. y.vergouwe@umcutrecht.nl

References

1. Nuttall GA, Houle TT: Liars, damn liars, and propensity scores. ANESTHESIOLOGY 2008; 108:3–4
2. Rosenbaum PR, Rubin DB: Reducing bias in observational studies. J Am Stat Assoc 1984; 79:516–24
3. Braitman LE, Rosenbaum PR: Rare outcomes, common treatments: Analytic strategies using propensity scores. Ann Intern Med 2002; 137:693–5
4. Cepeda MS, Boston R, Farrar JT, Strom BL: Comparison of logistic regression *versus* propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 2003; 158:280–7
5. Rosenbaum PR: Discussing hidden bias in observational studies. Ann Intern Med 1991; 115:901–5
6. Vandenbroucke JP: When are observational studies as credible as randomised trials? Lancet 2004; 363:1728–31

(Accepted for publication April 30, 2008.)

Anesthesiology 2008; 109:747

In Reply:—I would like to thank the journal ANESTHESIOLOGY for the opportunity to respond to these letters to the editor. In general, the purpose of our editorial[1] was to educate the anesthesiologist community on the strengths and weaknesses of propensity analysis. It was not meant advocate or demean this type of analysis. I will respond to each letter sequentially.

Dr. Engoren is correct that observational studies should be encouraged as a complement to prospective randomized studies. He is also correct that there are limitations and biases to prospective randomized control trials, which he enumerates. Despite these limitations, they are still considered the gold standard.

Drs. Vincent and Sakr are correct that one of the strengths of their study is the very large size of the Sepsis Occurrence in Acutely Ill Patients database.[2] It also is a weakness in that they are using data from another study that was designed for another purpose. They are correct that their statistical analysis is well performed. The comment in our editorial about the statistical process being opaque, simulating a "black box," was intended as a general comment about propensity analysis, not specifically their propensity analysis.

Drs. Boylan and Kavanagh are correct that our editorial was long on methods and short on biology. This was intentional because we had a limited word count and the goal of our editorial was to educate the anesthesiologist community on the strengths and weaknesses of propensity analysis. In searching through the literature, I found very few articles describing propensity analysis in the anesthesia literature. The authors do a very nice job describing the biology.

Dr. Vergouwe *et al.* are correct that propensity scores do not necessarily lie, but to nonstatisticians they are mysterious. The authors are correct that the title was a play on the quote by the English Prime Minister Benjamin Disraeli (1804–1880). Though the title was provocative, we tried to write a balanced editorial on the strengths and weaknesses of propensity analysis. I trust that the anesthesiologist community is smart enough not to be biased by a title in a single editorial.

Gregory A. Nuttall, M.D., Mayo Clinic College of Medicine, Rochester, Minnesota. nuttall.gregory@mayo.edu

References

1. Nuttall GA, Houle TT: Liars, damn liars, and propensity scores. ANESTHESIOLOGY 2008; 108:3–4
2. Vincent JL, Sakr Y, Sprung CL, Harboe S, Damas P: Are blood transfusions associated with greater mortality rates? ANESTHESIOLOGY 2008; 108:31–9

(Accepted for publication April 30, 2008.)