Broken Hearts

Steven L. Shafer, MD

Lindsay Borg is one smart resident.

I was taking the Stanford Anesthesia Residents through an article that had just appeared "in press" in *Anesthesia & Analgesia*.[1] The article, The Impact of Anesthesiologists on Coronary Artery Bypass Graft Surgery Outcomes, described how Glance et al. had demonstrated that the cardiac anesthesiologist had a significant effect on who lives and dies after cardiac surgery. The paper had impressed our reviewers and editors. Chuck Hogue, the handling editor, wrote to me "this may be one of the more important papers I have handled." Grateful to have received such a landmark submission, I asked experts in cardiac anesthesia, patient safety, and outcomes research to submit editorials.[2–4] I even put the article on the cover of the March 2015 issue of *Anesthesia & Analgesia* (Fig. 1).[a]

As shown in Figure 2, the findings were clear. "The rate of death or major complications among patients undergoing coronary artery bypass surgery varies markedly across anesthesiologists."[1] There was a highly significant ($P < 0.001$) difference in patient outcomes between the best and worst cardiac anesthesiologists. The article demonstrated something we've always thought: some anesthesiologists are better than others.

Lindsay wasn't impressed. "Isn't that just a tautology?" she asked. "When you order anesthesiologists by outcomes, don't you always get a graph that looks like that?

I assured her that while ranking anesthesiologists by random events would always generate a *similar* graph, these results were more extreme than expected from random events alone. I also explained that the significance isn't so much the shape of the curve, but rather the estimates of the standard errors around each point. If these estimates are quite small, then the model is very sure that the anesthesiologists at the edges are outliers. However, if the estimates are large, and particularly if they all intersect an odds ratio of 1, then none of the anesthesiologists is an outlier. Although clueless about the statistical methodology, I explained that the robust variance estimators properly adjusted for the clustering of the observations within anesthesiologists.

Later that day, I received an e-mail from Dennis Fisher, a close friend and frequent collaborator,[5] who asked a troubling question: "Isn't this just a tautology?" I offered Dennis the same hand waving. Dennis wasn't convinced either. He sent me a graph made completely from random distribution of events that looked a lot like the caterpillar graph in Figure 2, without the standard errors.

I checked with Frank Dexter, our statistical editor. "These outcomes aren't just random noise, right?" "Yes" Frank assured me. "The P value is < 0.001 while making multiple assumptions which would tend to increase the P value."

Then the Letters to the Editor started to arrive.[6–9] These respected colleagues collectively asked "isn't this just a tautology?" Doubts arose. I asked Larry Glance, the primary author, to respond authoritatively to these concerns.

Larry and his colleagues responded quickly. Their 22-page, 4700-word reply was comprehensive, elegant, authoritative, and convincing. Indeed, it was overwhelming.

Nathan Pace,[10] an anesthesiologist and statistician, also submitted a Letter to the Editor. Nathan's letter was a little different. To test for the possibility of a tautology, Nathan proposed a different analysis. I rejected Nathan's letter. He was proposing a new project, one too involved for a Letter to the Editor. However, I forwarded it to Larry for his consideration. Larry found Nathan's suggestions worthwhile and asked if he could reply. I agreed, expecting to wrap this up quickly.

Two months went by with no update. Then I received a sobering e-mail from Larry: "In replying to Nathan Pace's letter, I discovered that our use of clustered robust variance estimators led to downwardly biased estimates of the anesthesiologists standard errors—and to the wrong conclusion regarding the significance of the anesthesiologist effect."

In other words, the conclusion was wrong. It was a tautology after all. Lindsay was right, although she couldn't know that from looking at the figure without considering the standard errors.

Or was the original article right, and the new analysis wrong? We had to be certain! The problem was subtle, the difference between "robust cluster (anes)" and "robust" in the Stata program code. It seemed to me and the reviewers that the analysis should be clustered, as in the original article. If so, then Larry's new analysis was in error, not the original analysis. Over the course of 2 months and approximately 100 e-mails, we discussed the putative mistake with

From the Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Stanford, California.

Accepted for publication February 5, 2016.

Funding: None.

The author declares no conflicts of interest.

Reprints will not be available from the author.

Address correspondence to Steven L. Shafer, MD, Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, 300 Pasteur Dr., MC-5640, Stanford, CA 94305. Address e-mail to steven.shafer@stanford.edu.

DOI: 10.1213/ANE.0000000000001253

[a]Available at: http://journals.lww.com/anesthesia-analgesia/Publishing-Images/CoverArtGallery/1503_cover.jpg. Accessed February 2, 2016.

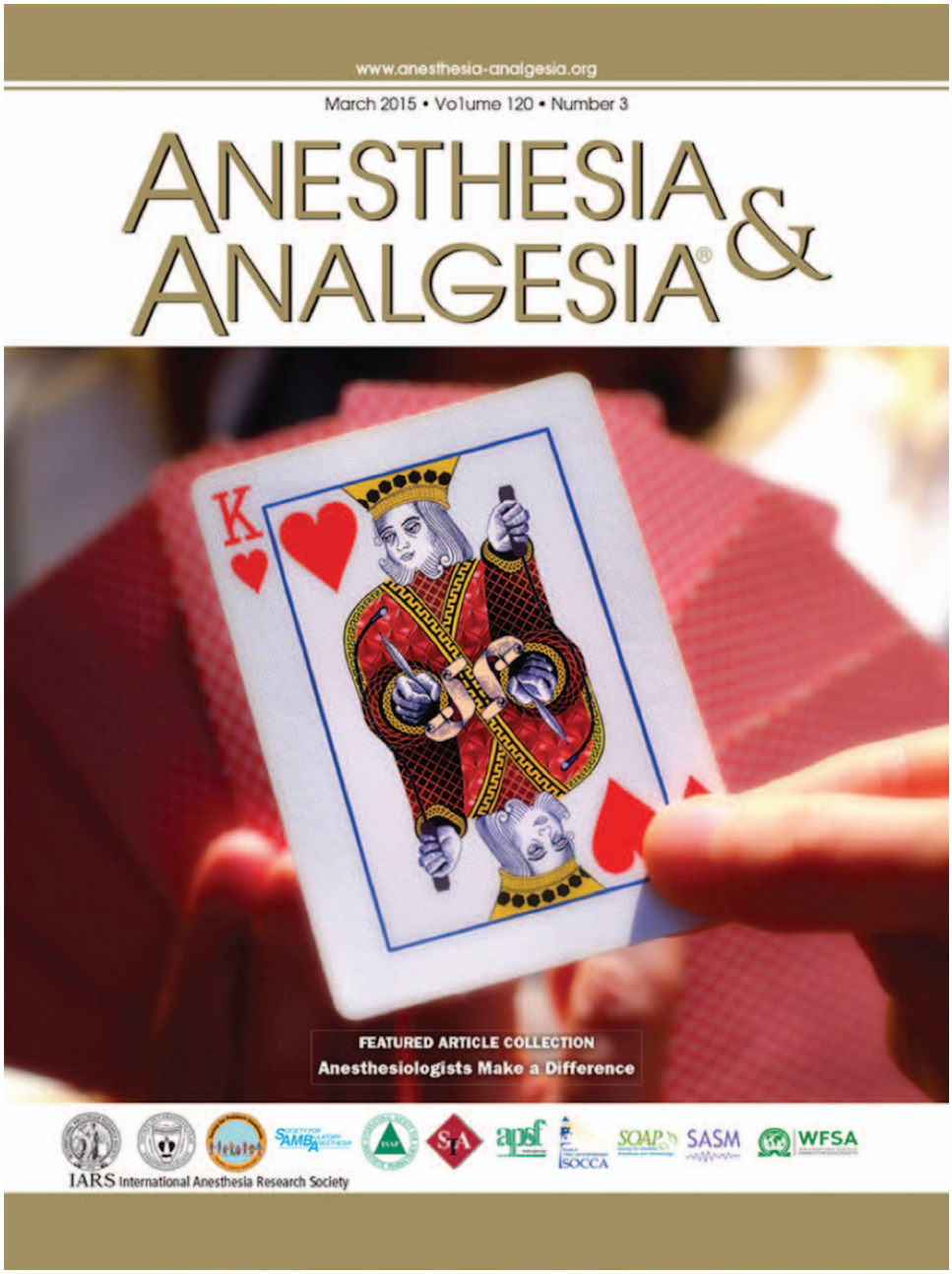

Figure 1. Cover of the March 2015 issue of *Anesthesia & Analgesia* (artwork by Naveen Nathan, MD).

Larry and his co-authors, other editors, and reviewers. We generated and exchanged multiple simulated data sets. We sought outside experts. According to Nichols and Schaffer, "when fixed effects and clustering are specified at the same level, tests that involve the fixed effects themselves are inadvisable (the standard errors on fixed effects are likely to be substantially underestimated, though this will not affect the other variance estimates in general)."[b] Larry e-mailed Austin Nichols, who confirmed that "The cluster-robust standard error is biased downward but consistent, so you need many, many more clusters than coefficients tested to reach asymptopia; if you are testing fixed effects, you may often find that the cluster-robust standard errors on fixed effects are essentially zero and you have effectively zero degrees of freedom but there is no error message."

Convinced that the March 2015 manuscript was wrong, Larry prepared a new response to the Letters to the Editor.[11] He explained the mistake in the analysis. He acknowledged that the differences in anesthesiologist performance seen in Figure 2 were due to chance alone.

The fundamental problem is that adverse events are so rare that even a seemingly large database (91 anesthesiologists and 7920 patients) was underpowered to find differences. The conclusion that the data showed performance differences was wrong. The article had to be retracted.

Larry asked if he could increase the size of the database and publish a new analysis of the results. I agreed but emphasized that the Journal had a responsibility to report the erroneous results within a reasonably short period of

[b]Nichols A, Schaffer M. Clustered Errors in Stata. Available at: http://repec.org/usug2007/crse.pdf. Accessed February 2, 2016.

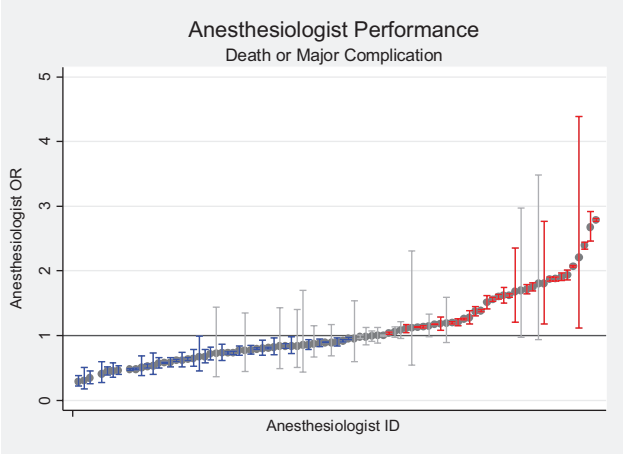

Figure 2. Each anesthesiologist is represented by the point estimate for his/her adjusted odds ratio, along with a 95% confidence interval (error bar). High-performance outliers are highlighted in blue and low-performance outliers in red (quoted verbatim from original article).[1]

time. Within 4 months, the authors obtained the expanded data set, performed a revised analysis, and submitted the new paper. The new paper went through 3 rounds of peer review to be certain it was right. This issue of *Anesthesia & Analgesia* has the retraction of the original article[12] and the new manuscript that replaces it.[13]

What can we learn from this? First, the most critical peer review occurs *after* a paper is published, when it is reviewed by dozens of experts. If something is wrong, it will likely be identified. Second, peer review is imperfect. Everyone who has received a rejection letter already knows that. Third, science is self-correcting when authors, reviewers, and editors have the will to do it.

The final lesson is that "retraction" is not a good word to describe the withdrawal of a manuscript when diligent authors discover and report a mistake in their work. There is too much baggage associated with "retraction." Nathan Pace proposed a thoughtful question. Larry Glance diligently pursued the question. In the process, he identified an error in his analysis. The error was subtle: clustered versus standard robust variance estimators. The impact was not subtle: erroneously reporting a *P* value of approximately 0[c] when the true *P* value was approximately 1. Once Larry had convinced himself, he had to convince his co-authors, then me, then our reviewers, and finally our readership.

[c]χ^2 = 1273 with 68 degrees of freedom. The *P* value in *R* is calculated as 1-pchisq(1273,68). The *P* value is so small that *R* returns 0. For comparison, 1-pchisq(213.68) returns 1.1e-16.

An author's finding a mistake and working diligently to correct the published record is a profound demonstration of commitment to academic integrity. "Retraction" seems a harsh reward for honesty. We need a better word to describe the result. We need a term that acknowledges honest error, a type of error familiar to every honest author. We also need to reward honesty with gratitude.

On behalf of the Journal, the Editors, our readers, and the patients whom we serve, I express appreciation to Dr. Glance and his co-authors for their commitment to scientific integrity.

DISCLOSURES

Name: Steven L. Shafer, MD.
Contribution: This author helped write the manuscript.
Attestation: Steven L. Shafer approved the final version of the manuscript.

RECUSE NOTE

Steven L. Shafer is the Editor-in-Chief for *Anesthesia & Analgesia*. This manuscript was handled by James G. Bovill, Guest Editor-in-Chief, and Dr. Shafer was not involved in any way with the editorial process or decision.

REFERENCES

1. Glance LG, Kellermann AL, Hannan EL, Fleisher LA, Eaton MP, Dutton RP, Lustik SJ, Li Y, Dick AW. The impact of anesthesiologists on coronary artery bypass graft surgery outcomes. Anesth Analg 2015;120:526–33
2. Shafer SL. Anesthesiologists make a difference. Anesth Analg 2015;120:497–8
3. Maxwell BG, Hogue CW Jr, Pronovost PJ. Does it matter who the anesthesiologist is for my heart surgery? Anesth Analg 2015;120:499–501
4. Wijeysundera DN, Beattie WS. Facing the uncomfortable truth: your choice of anesthesiologist does matter. Anesth Analg 2015;120:502–3
5. Fisher DM, Shafer SL. Allometry, shallometry! Anesth Analg 2016;122:1234–8
6. Gibbs NM, Weightman WM, Ho KM. Anesthesia outcome and chance. Anesth Analg;122:1717–8
7. Myles P, Kasza J. Physician performance rankings based on outcomes, confirmed by the same outcomes: a tautology. Anesth Analg 2016;122:1718–9
8. Mackay JH, Nashef SAM, Paprachristofi O, Sharples L. The impact of anesthesiologists on coronary artery bypass graft outcomes. Anesth Analg 2016;122:1719
9. Barbeito A, Raghunathan K, Albrecht R. Primus inter pares. Anesth Analg 2016;122:1719–20
10. Pace NL. Erroneous ranking of anesthesiologists? Anesth Analg 2016;122:1720–1
11. Glance LG, Dick AW. In response. Anesth Analg 2016;122:1722–7
12. Shafer SL. Notice of retraction. Anesth Analg 2016;122:1730
13. Glance LG, Hannan EL, Fleisher LA, Eaton MP, Dutton RP, Lustik SJ, Li Y, Dick AW. Feasibility of report cards for measuring anesthesiologist quality. Anesth Analg 2016;122:1603–13