Clin Orthop Relat Res (2013) 471:3745–3749 / DOI 10.1007/s11999-013-3337-4

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

Published online: 19 October 2013

Not the Last Word: Harvard Beats Yale and Other Fallacies

Joseph Bernstein PhD

One of the great academic rivalries in the United States is between the nation's oldest institution of higher learning, Harvard University, and its New England neighbor to the south, Yale University. Both of these Ivy League schools typically stand near the top of national rankings, and jostling for the number one spot is a favored pastime among many of their alumni. When the 1968 Harvard versus Yale football game ended in a draw, for example, the Harvard student newspaper published the headline, "Harvard beats Yale 29-29."

Both Harvard and Yale boast top notch orthopaedic surgery departments, yet the orthopaedic residency at Yale is more likely to come up short on one Residency Review Committee (RRC) guideline for program evaluation, namely, that "a board certification rate [of its graduates] greater than 75% is required to maintain Residency Review Committee Accreditation" [4].

The assertion that Yale is more likely to violate this 75% RRC standard says nothing about the superior didactic programs or clinical experiences at Harvard. Rather, it relies on a long-known, but frequently forgotten mathematical phenomenon: variance increases with decreasing sample size. Therefore, a smaller program is more likely to be found at the extremes of performance, both good and bad. The larger a program is, the more it is protected against low-probability events (such has having too many recent graduates fail to pass their boards). The residency program at Harvard graduates 12 residents per year, and the program at Yale graduates five. Advantage: Harvard.

Here's why. Assuming that every graduate has a p chance of passing the boards examination, one can employ the binomial distribution to determine the chance that there will be at least m students out of n who pass in a given year [1]. For the Harvard program, if each of the 12 residents has an 80% chance of passing, there is 79.4% chance that in a given year at least nine will pass the exam. For the Yale program, with five residents with that same individual success rate of 80%, there is only a 73.7% chance that it will avoid sanction in a given year. For a 4-year average (48 versus 20 residents), the difference is even greater. With an 80% individual success rate, there is less than 1% chance that fewer than 32 of 48 Harvard graduates will pass. The probability that fewer than 15 out of 20 from Yale will pass is 20%. Under the current rules, we could easily imagine a report from the Residency Review Committee headlined "Harvard Beats Yale 80% to 80%."

This example illustrates the ease with which one can be misled when drawing inferences from extreme performance without adequately accounting for sample size differences. Wainer and Zwerling [13] provide a few additional examples. For instance, they found that the lowest rates of kidney cancer are found in rural areas, but caution against the inference that small town living is somehow healthier. That is because they also found that other

The author certifies that he, or a member of his immediate family, has no funding or commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research*® editors and board members are on file with the publication and can be viewed on request.
The opinions expressed are those of the writers, and do not reflect the opinion or policy of *Clinical Orthopaedics and Related Research*® or the Association of Bone and Joint Surgeons®.

J. Bernstein PhD (✉)
Department of Orthopaedic Surgery,
University of Pennsylvania,
424 Stemmler Hall, Philadelphia,
PA 19104, USA
e-mail: orthodoc@uphs.upenn.edu

Not the Last Word

small towns had the highest rates of cancer. Indeed, small areas are likely to be found at both extremes: a village with 100 inhabitants and no cancer will have the lowest rate, whereas a village of 100 inhabitants and but a single case of cancer will have a rate among the highest. One additional (or one fewer) case in a large city, by contrast, imposes a trivial effect on incidence.

To appreciate the powerful influence sample size has on variance, and in turn, the probability of witnessing a rare event, imagine that you are administering Part II of the American Board of Orthopaedic Surgery exam. A candidate has presented his experience of 30 total hip replacements. You come to learn that three of these arthroplasties, 10%, have dislocated. You also recall that the literature reports a dislocation rate of 3.2%. Should you condemn the work of a surgeon whose hips dislocate at more than triple the expected number? At first glance, the 10% rate does seem to be shoddy work, yet the binomial distribution preaches caution. If the examinee's true rate of a dislocation is 3.2% (ie, the rate that would be seen after a very large number of cases), then in the small sample of thirty cases there is a 37% chance that there will be no dislocations, a 37% chance that there will be exactly one, a 17% chance that there will be two, and 6% chance that there will be three. Further, assuming that an event with a 6% probability should be discounted as noise (a fair interpretation of the "$p > 0.05$" criterion for statistical significance), this candidate should not be censured. What is seen may just represent a statistical anomaly, owing to a small sample size.

Consideration of sample size is of even greater importance as clinicians are increasingly critiqued and graded on every aspect of their performance. If the orthopaedic surgery community does not want to be held accountable by the public for random events — a 10% dislocation rate after 30 hip replacements comes to mind — then it behooves us to not hold each other accountable for random events. We can start by dropping this Residency Review Committee standard for residency performance, or at least modifying it to account for sample size. (For example, one could calculate a rolling average of its most recent 100 graduates). If we account for sample size, the department at Harvard may still one day be able to claim superiority, but not by hiding behind a statistical quirk.

Saam Morshed MD, PhD

Assistant Professor, Department
of Orthopaedic Surgery, University
of California, San Francisco

In this thought-provoking piece, Dr. Bernstein has eloquently reminded us of the importance of understanding basic statistical principles such as the binomial distribution and sample size. The danger of small sample size and risk for Type II errors (false-negative result) that plague the orthopaedic literature [8] are now well documented. However, a lesser-known risk is that erroneous conclusions of all kinds can easily be drawn from small trials, including those that are randomized [7]. This is due to the disproportionate impact of outcome events in smaller studies due to the size of the denominator. Devereaux and colleagues [3] have labeled this phenomenon "fragility." Deveraux et al. [3] used fragility to argue for larger clinical trials where the increased sample size effectively buffers results from the sway of a small number of incident cases. Dr. Bernstein uses several analogous examples of spurious conclusion that may be drawn out of a lack of appreciation for the implications of inferences based on small samples. In medical school and residency training programs, we often teach these concepts to our students in order to make them more critical thinkers and consumers of the literature, more able to successfully practice in an evidence-based way, and to inspire some to go on to create knowledge. After reading this work, I hope we can all appreciate how vital skills in quantitative reasoning are to understanding the merits and perils of broader issues and policies that dictate

Not the Last Word

everything — from the way that we evaluate training programs to our own clinical performance.

Howard Wainer PhD

Distinguished Research Scientist,
National Board of Medical
Examiners

It is said that there are two kinds of lawyers; the first tells you that you cannot do it, and the second tells you how to do it. In the remarks that follow, I fear that I will come off as the wrong kind of lawyer, for I focus on why one should avoid drawing inferences from insufficient samples, and not on how to draw such inferences safely. I do this for two reasons. First, because of sensible space restrictions; there is no room for a useful description of how to do it. Second, the details require a moderate amount of algebra in their description, and a parallel amount of effort to understand them. I opted to augment the consciousness-raising in Dr. Bernstein's amusing and informative paper with facts. In so doing, I hope to instill some fear in the reader — fear that if they ignore Dr. Bernstein's caveat, they might end up looking foolish.

Dr. Bernstein's paper is about an equation that first appeared in a 1730 paper [2] by the French mathematician

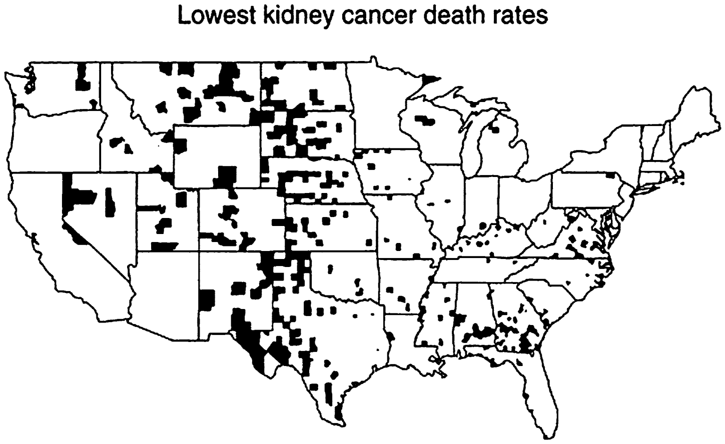

Fig. 1 The counties of the United States with the lowest 10% age-standardized death rates for cancer of kidney/urethra for U. S. males, 1980–1989 [5]. Reprinted with permission from Oxford University Press. Gelman A, Nolan DA. *Teaching Statistics: A Bag of Tricks*. Oxford University Press; 2002:14.

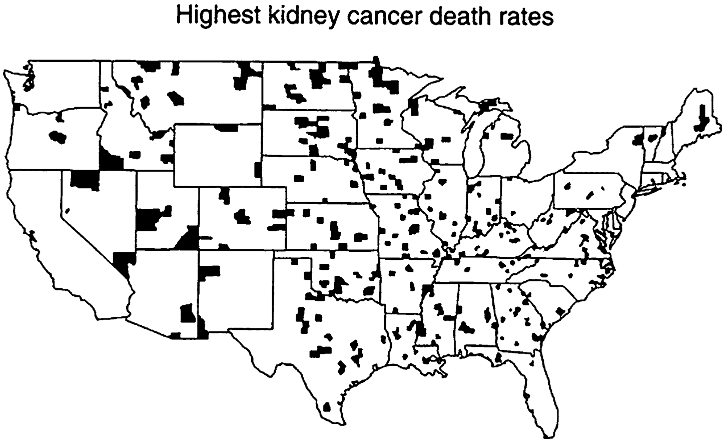

Fig. 2 The counties of the United States with the highest 10% age-standardized death rates for cancer of kidney/urethra for U.S. males, 1980–1989 [6]. Reprinted with permission from Oxford University Press. Gelman A, Nolan DA. *Teaching Statistics: A Bag of Tricks*. Oxford University Press; 2002:15.

Not the Last Word

Abraham de Moivre that showed the relationship between the variability of a mean and the sample size on which it was calculated. As Dr. Bernstein has pointed out, after almost 300 years, it is still a mystery to many; the damage such ignorance has caused over the centuries has led to it being a prime candidate for the title "The Most Dangerous Equation" [11, 12].

A map of age-adjusted kidney cancer rates (Fig. 1) will amplify Bernstein's description. As I have described in previous publications [11, 12] on the subject, the counties shaded are in the lowest decile of the cancer distribution. These areas have a tendency to be rural, Midwestern, southern, and western counties. This could be due to the clean living of the rural lifestyle — less air and water pollution.

The second map (Fig. 2) shows a similar map of age-adjusted kidney cancer rates. Although comparable to the first map, the second map differs in one essential detail — the shaded counties represent the highest decile of the cancer distribution. Again, these counties have a tendency to be rural, Midwestern, southern, and western [11, 12]. We can surmise that this outcome could be directly due to the poverty-stricken and unhealthy eating habits sometimes associated with the rural lifestyle [11, 12].

If we were to plot the first map on top of the second, many of the shaded counties on the first map would be adjacent to the shaded counties from the second map [11, 12]. This is a function of de Moivre's equation. The deviation of the mean is inversely relative to the square root of the sample size, and small counties have more variance than big counties [11, 12]. As Bernstein mentioned in his column, a county with, approximately 100 inhabitants that has no cancer deaths would be in the lowest category. But if it has even one cancer death, it would be among the highest [11, 12].

The age-adjusted cancer rates plotted against county population show a clearer picture (Fig. 3). We can see an extensive variation in cancer rates in the less populated regions (left side of the graph), but little variation when county populations are large (right side of graph) [11, 12].

It is easy for someone studying Figure 1 to conclude that people at risk of kidney cancer should move to the wide open spaces of rural America. But we know (thanks to better understanding de Moivre's equation) that this

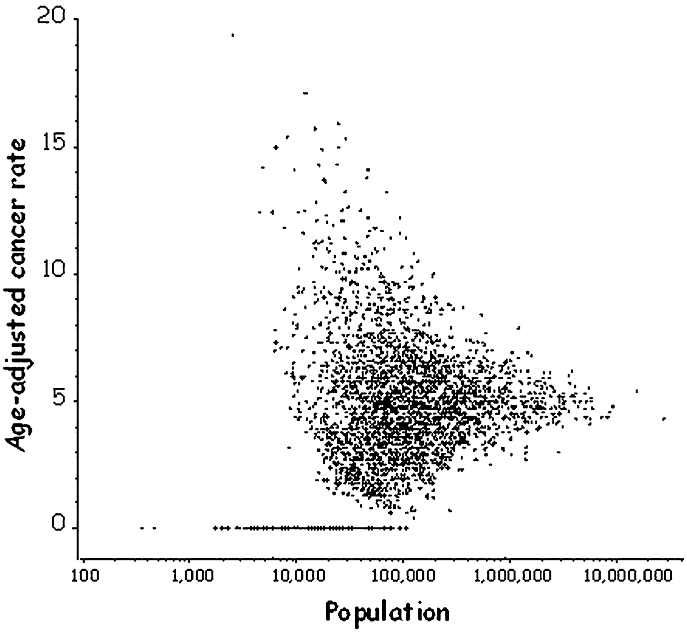

Fig. 3 Age adjusted kidney cancer rates for all U.S. counties in 1980–1989 shown as a function of the log of the county population [10]. Reprinted with permission from Oxford University Press. Wainer H. *Medical Illuminations: Using Evidence, Visualization, and Statistical Thinking to Improve Healthcare.* Oxford University Press; 2013:18.

Not the Last Word

would be incorrect. Being aware of the dangers of chasing noise when samples are small is important. An efficacious way to boost the reliability of estimates in small samples is by borrowing strength from other parts of the design. This borrowing is done well using empirical Bayes procedures [9], and is a more powerful cousin to the sort of averaging that Dr. Bernstein suggested.

But alas, that is a longer story must await another day.

References

1. Binomial calculator: online statistical table. Available at http://stattrek.com/online-calculator/binomial.aspx. Accessed: March 19, 2013.
2. de Moivre, A. *Miscellanea analytica de seriebus et quadraturi.* London, United Kingdom: Tonson and Watts; 1730.
3. Devereaux PJ, Chan MT, Eisenach J, Schricker T, Sessler DI. The need for large clinical studies in perioperative medicine. *Anesthesiology*. 2012;116: 1169–1175.
4. Dougherty PJ, Walter N, Schilling P, Najibi S, Herkowitz H. Do scores of the USMLE step 1 and OITE correlate with the ABOS part I certifying examination?: a multicenter study. *Clin Orthop Relat Res.* 2010;468: 2797–2802.
5. Gelman A, Nolan DA. *Teaching Statistics: A Bag of Tricks.* Oxford, United Kingdom: Oxford University Press; 2002:14.
6. Gelman A, Nolan DA. *Teaching Statistics: A Bag of Tricks.* Oxford, United Kingdom: Oxford University Press; 2002:15.
7. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294:218–228.
8. Lochner HV, Bhandari M, Tornetta P, 3rd. Type-II error rates (beta errors) of randomized trials in orthopaedic trauma. *J Bone Joint Surg Am.* 2001; 83-A:1650–1655.
9. Rubin DB. Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association.* 1980;75: 801–816.
10. Wainer H. *Medical Illuminations: Using Evidence, Visualization, and Statistical Thinking to Improve Healthcare.* London, United Kingdom: Oxford University Press; 2013:18.
11. Wainer H. *Picturing the Uncertain World: How to Understand, Communicate, and Control Uncertainty through Graphical Display*. Princeton, NJ: Princeton University Press; 2011:5–20.
12. Wainer H. The most dangerous equation. *American Scientist*. 2007;95: 249–256.
13. Wainer H, Zwerling HL. Questioning evidence for smaller schools: evidence that smaller schools do not improve student achievement. *The Phi Delta Kappan.* 2006;88: 300–303.