Role of randomness in observed outcomes and implications for quality assessment

Jeffrey B. Michel, MD^a, Michael J. Mack, MD^b, and Gregory J. Dehmer, MD^c

^aCardiology Division, Baylor Scott and White Temple, Temple, Texas; ^bDepartment of Cardiovascular Surgery, The Heart Hospital Plano, Plano, Texas; ^cCardiology Division, Carilion Clinic, Roanoke, Virginia

Data collection and outcomes analysis are important tools in the practice of medicine. Scorecards and metrics give physicians and patients alike objective information concerning outcomes associated with medical care. However, determining the true quality of care using observed outcomes can be difficult and unreliable. For example, is a surgeon who begins practice by losing his or her first patient a bad surgeon? Should this surgeon be allowed to continue practice? Clearly, after only one operation, there are insufficient data to reliably answer that question. Nevertheless, important decisions are sometimes based on limited data that lack statistical validity.

Randomness can be thought of as a lack of predictability or pattern to observed events. Although the pattern of observed events may be unpredictable, the frequency of observed events is often well established. For example, suppose that after the collection of data on thousands of operations, it is known that a specific complication occurs in 1 out of 100 operations or 10 out of 1000 operations. Depending on the surgeon and the type of operation, 1000 operations could represent 10 years of practice. The principle of randomness would endorse that the specific complication need not occur at a nice even frequency of 1 per 100 cases. A surgeon may perform 300 cases without the specific complication, but then have a cluster of 4 complications in the next 100 cases. In other words, the outcome of interest may not be evenly spaced but rather may occur in clumps yet still have the accepted frequency of the complication.

New York State pioneered the collection, analysis, and public reporting of cardiac surgery mortality data in 1989.¹ Within 4 years, there was a significant reduction in observed mortality.² However, it was unclear whether surgery itself became safer. Some have suggested that surgeons responded by turning down high-risk cases, and outmigration of such

cases to other states was reported.³ Today, the outcomes of >95% of cardiac surgeons and hospitals in the United States are collected and analyzed by the Society of Thoracic Surgeons (STS).⁴ We assume that better observed outcomes reflect better quality, but the concept of randomness is often not appreciated in interval reports.

Consider the 2015 STS data for isolated coronary artery bypass graft outcomes at three New York hospitals: Bellevue, Basset, and Arnot Ogden, which reported 2, 2, and 1 mortalities on volumes of 113, 74, and 80 procedures, respectively.⁵ Is there a meaningful difference between these programs? The analysis provided 95% confidence intervals for risk-adjusted mortality to help readers understand the potential variation in the point estimates of mortality. The confidence intervals were 0.32 to 10.42 for Bellevue, 0.46 to 14.91 for Basset, and 0.03 to 10.42 for Arnot Ogden. All were very wide and overlap and thus indicate that the difference in mortality between hospitals is statistically insignificant. The concept of randomness contributes to the magnitude of these confidence intervals, so how should such data be interpreted?

This is a challenge for both patients and medical leaders. Psychologists report that people in general have poor statistical intuition, view data in an overly deterministic fashion, and tend to discount the role of chance in observed events.^{6,7}

If confidence intervals fail to convey the role randomness can play in observed events, is there another way? The so-called Monte Carlo method was developed in the 1940s to help solve key problems inherent to the design and manufacture of atomic bombs.⁸ This approach allows one or more variables to be held constant while using a random number generator to model possible outcomes. If the mortality rate is fixed at 1%, for example, any difference between 1% and the rate observed in the simulation is due entirely to chance.

Corresponding author: Jeffrey B. Michel, MD, Cardiology Division, Baylor Scott and White Memorial Hospital, 2401 South 31st Street, Temple, TX 76502 (e-mail: Jeffrey.Michel@BSWHealth.org)

Received December 20, 2019; Revised February 23, 2020; Accepted February 27, 2020.

Table 1. Observed vs. set mortality rates for simulated 6-month periods*

	Month in period						Observed MR (%)	Fixed MR (%)	Concordant (Yes/No)
	1	2	3	4	5	6			
Period 1 (Months 1 to 6)									
A	0	1	0	0	0	0	1.7%	1.0%	No
B	0	0	0	0	0	0	0.0%	2.0%	No
C	1	0	0	0	0	0	1.7%	3.0%	No
Period 2 (Months 7 to 12)									
A	0	0	0	0	1	0	1.7%	1.0%	No
B	0	0	0	0	0	0	0.0%	2.0%	No
C	0	0	1	0	0	0	1.7%	3.0%	No
Period 3 (Months 13 to 18)									
A	0	0	0	1	0	0	1.7%	1.0%	No
B	0	1	0	0	0	0	1.7%	2.0%	Yes
C	1	0	0	1	0	0	3.3%	3.0%	Yes
Period 4 (Months 19 to 24)									
A	0	1	0	0	0	0	1.7%	1.0%	No
B	1	1	0	0	1	0	5.0%	2.0%	No
C	0	1	0	0	0	1	3.3%	3.0%	Yes
Period 5 (Months 25 to 30)									
A	0	0	1	0	0	0	1.7%	1.0%	No
B	0	0	0	0	0	0	0.0%	2.0%	No
C	0	0	1	1	0	0	3.3%	3.0%	Yes

*A Monte Carlo simulation created three surgeons each performing 10 surgeries a month. Using Excel, a column of 10 cells using the [rand()] function assigned to each cell was used to generate random numbers between 0 and 1. A second column was created with the function [IF A > 0.99, "1", "0"] with A being the adjacent cell in the column that contains a randomly generated number. Two columns represent each hospital A, B, and C, with the first column generating the random number and the second column reporting the outcome. A > 0.99 defined our high-performing center while A > 0.98 and A > 0.97 defined the average and low-performing hospitals, respectively. Thirty runs created 30 months of data. MR indicates mortality rate. Concordance between observed and set mortality rates was determined. Results are a single simulation.

Using Monte Carlo methods, we can set a 1% mortality rate for a hypothetical surgeon and simulate 10 “surgeries,” with the series yielding between 0 and 10 mortalities. Zero will be the most common observed outcome, though the rate is in fact 1%. From time to time 1 mortality will be observed in any given series of 10, though the mortality rate remains 1%, not 10%. Run the simulation hundreds of times and it is likely that two or more mortalities may be observed in a given series. Again, when observed, such clusters do not indicate a change in the underlying mortality rate, which remains 1%. It is even possible for 10 consecutive mortalities to occur in a single series, though the odds of this happening are extremely low, on the order of 1 in 10.⁹

Are events such as 10 surgical mortalities in a row actually seen? A good person to ask might be Evelyn Adams, a woman who famously won the New Jersey State Lottery *twice in a row*, once in 1985 and again in 1986. At the time, it was estimated that the odds of a person winning the lottery once was greater than 1 in a million. The odds of her winning twice were estimated to be greater than 1 in a trillion or 10 to the 12th power.⁵ If our hypothetical surgeon had a spotless record with no mortalities in 9 years but then had 10 mortalities after performing 10 consecutive surgeries, could it be argued that the odds against this were extraordinarily low and resulted from randomness? However, with 10 consecutive mortalities, it is hard to imagine a hospital administrator who would not start asking questions or even temporarily suspend surgical privileges.

Consider the following example using a Monte Carlo simulation with more realistic numbers. A hospital has three cardiac surgeons. Surgeon A has a mortality rate of 1% and is a “high performer.” Surgeon B is “average” with a 2% mortality rate, while Surgeon C is a “low performer” with a mortality rate of 3%. Is the difference between surgeons significant? In absolute terms, Surgeon C has a mortality rate of 2%, which is higher than that of Surgeon A, but the difference can be described in a more dramatic manner by saying Surgeon C’s mortality is threefold or 300% higher than that of Surgeon A.

You are the chief quality officer for the hospital at which these surgeons work. Each surgeon does 10 surgeries a month, for a combined total of 30 surgeries a month for the hospital. You meet with the CEO every 6 months to review data. *Table 1* shows the results from a single simulation over the course of 30 months and demonstrates how the pattern of observed outcomes can unfold, knowing fixed event rates.

Period 1. The CEO is pleased. There is concern about mortalities that occurred in the first 2 months, but it is hard to blame you for those since you just started. The CEO is happy to see that Surgeon B has improved to high performing and is excited that your hard work with Surgeon C has paid off by moving out of the low-performing group. However, it is noteworthy that after 6 months, none of the observed outcomes match the true mortality rates set in the simulation.

Period 2. The CEO is concerned. There has been no improvement. However, administrators from other systems have called to learn how Surgeon B became high performing. Surgeon B has been invited to give a lecture titled “Good to Great: A Case Study in Excellence” at a national meeting. However, again, observed outcomes do not match the actual mortality rates, even after 12 months and 360 surgeries by the three surgeons.

Period 3. The CEO is disappointed that Surgeon C has regressed back to low performing. It was expected that Surgeon C’s performance would remain average. In addition, Surgeon B is no longer high performing. As the chief quality officer, you are given 12 months to improve cardiac surgery mortality rates.

Period 4. The CEO expresses frustration with the poor performance and wants an explanation of what happened in

month 20 when there was a cluster of mortalities. You are given 6 months to turn it around. However, it is important to emphasize that the simulation holds the mortality of individual surgeons constant; thus, all fluctuation in observed mortality is entirely due to chance.

Period 5. Your final meeting with the CEO is uncomfortable. The CEO notes that you have not fixed low-performing Surgeon C. This is true. Surgeon C has a 3% mortality, which is reflected by observed mortality. Moreover, Surgeon A appears to have gotten worse, but that is incorrect as the simulation fixed Surgeon A's mortality rate at 1%. Also incorrect is the apparent improvement in Surgeon B in the past 6 months, as Surgeon B's mortality rate was fixed at 2% by the simulation.

You correctly argue that aggregating data from the entire 30-month period improves the ability to determine the actual mortality rate for each surgeon. However, with an observed rate of 1.7% over 30 months, Surgeon A is above the 1% mortality rate set in the simulation. The observed rate for Surgeon B (1.3%) underestimates the actual 2% rate. In fact, Surgeon B appears to perform better than Surgeon A, which is false. Surgeon C (2.7%) is accurately classified as low performing. Month 20 had 3 mortalities, one for each surgeon, creating the appearance of a "spike" in mortality to 10%.

It is difficult for medical leaders to wait years to act on data, especially when those data suggest poor performance. Many leaders might interpret this as a "call to action." Meetings and performance improvement plans would likely follow. When no mortalities were observed for the 2 months following Month 20, leadership might think such efforts were effective. But the observed rise and fall in mortality was entirely random and did not reflect a true change in performance. The "spike" was meaningless.

What lessons can we learn from our simulation? First, fluctuations in observed data occur randomly, such that poor outcomes can cluster without any change in the true quality of care delivered. The mortality rates set in our simulation were fixed, yet observed month-to-month mortality fluctuated. It is obvious, yet not always intuitive, that a 1% mortality rate does not mean that one mortality will be observed in any given series of 100 surgeries. Any given series of 100 might yield no mortalities. Alternatively, two or more might be observed based on chance alone.

The number of observed events matters. So does the level of discrimination being sought. As poor outcomes become less frequent (wider confidence intervals), substantially more events must be observed to accurately gauge performance, i.e., narrow the confidence intervals. Looking frequently, say monthly, at small numbers of events can be a fool's errand.

Taking a longer view works better. In our simulation, observed mortality failed to reflect actual rates in 11 of 15 (73%) 6-month periods. No 6-month period (0%) captured the true mortality rates for all three surgeons. Given the frequency of current reporting and review (usually 12 months or less), health care leaders may not have enough data to

estimate quality with confidence. Yet it is hard not to feel compelled to act when a cluster of adverse events occurs.

Returning to Bellevue, Bassett, and Arnot hospitals in the subsequent year, 2016, there were 5, 2, and 0 mortalities on volumes of 144, 85, and 53, with confidence intervals of 1.47 to 10.67, 0.17 to 5.43, and 0.00 to 7.13, respectively. It was not possible to predict 2016 outcomes from a review of 2015 data, nor vice versa. The inability to predict results for future years from past years was recently shown in a large dataset of outcomes for percutaneous coronary intervention.⁹ It is critical to understand that the observed variation within and between hospitals over short time spans may result from chance alone. Differences between individual surgeons or individual hospitals may exist but are often more difficult to demonstrate than simply looking at the numbers. Risk adjustment is helpful but cannot correct for randomness in data.

But as we improve care and "drive down" observed event rates, it becomes more difficult to distinguish high and low performers. Given the limited number of observed events available to decision makers, observed data usually offer a "best guess," rather than absolute truth. The decisions that must be made are rarely black and white. No one will fault efforts to improve quality when data look bad. Perhaps the bigger challenge is to remain vigilant when the data look good. A week, a month, or even a year without observed adverse events should not lull us into a false belief that we have achieved perfection.

1. Hannan EL, Kilburn H Jr, O'Donnell JF, Lukacik G, Shields ES. Adult open-heart surgery in New York State: an analysis of risk factors and hospital mortality rates. *JAMA* 1990;264(21):2768–2774. doi:10.1001/jama.1990.03450210068035.
2. Hannan EL, Kumar D, Racz M, Siu AL, Chassin MR. New York State's cardiac surgery reporting system: four years later. *Ann Thorac Surg*. 1994;58(6):1852–1857. doi:10.1016/0003-4975(94)91726-4.
3. Omoigui NA, Miller DP, Brown JK, et al. Outmigration for coronary bypass surgery in an era of public dissemination of clinical outcomes. *Circulation*. 1996;93(1):27–33. doi:10.1161/01.CIR.93.1.27.
4. Shahian DM, Jacobs JP, Edwards FH, et al. The Society of Thoracic Surgeons national database. *Heart*. 2013;99(20):1494–1501. doi:10.1136/heartjnl-2012-303456.
5. New York State Department of Health. *Adult Cardiac Surgery in New York 2014–2016*. April 2019. https://www.health.ny.gov/docs/2014-2016_adult_cardiac_surgery.
6. Taleb NN. *Fooled by Randomness: The Hidden Role of Chance in the Markets and in Life*. New York: Texere; 2001:182–221.
7. Tversky A, Kahneman D. Judgements under uncertainty: heuristics and biases. *Science*. 1974;185(4157):1124–1131. doi:10.1126/science.185.4157.1124.
8. Rubinstein RY, Kroese DP. *Simulation and the Monte Carlo Method*. 2nd ed. New York: John Wiley & Sons; 2007.
9. Sandhu AT, Kohsaka S, Bhattacharya J, Fearon WF, Harrington RA, Heidenreich PA. Association between current and future annual hospital percutaneous coronary intervention mortality rates. *JAMA Cardiol*. 2019;4(11):1077–1083. doi:10.1001/jamacardio.2019.3221.