Outcomes in cardiovascular surgery and cardiology have arguably been the most reported and transparent of any field of medicine. Data are rapidly becoming available for institutions and even individual physicians. With this increased visibility come the better recognition of outcomes and a push for increasing quality. But as W. Edwards Deming, who is considered a quality guru, might have said, “Uncontrolled variation is the enemy of quality.” The problem in medicine as opposed to manufacturing is that, in the words of William Osler,1 “Variability is the law of life, and as no two faces are the same, so no two bodies are alike, and no two individuals react alike and behave alike under the abnormal conditions which we know as disease.” To attempt to overcome this inherent human variability, we construct trials with inclusion and exclusion criteria designed to recruit a similar population and then randomly allocate between treatment arms to distribute equally any unrecognized variation. A series of randomized trials comparing transcatheter aortic valve replacement (TAVR) with surgical aortic valve replacement (SAVR) have been run, including the PARTNERII A trial (PARTNER II Trial: Placement of AoRTic TraNscatheter Valves—PII A).2-8 Even if we get similar populations to compare with this method, however, how do we understand, much less adjust for, differences between individual skill levels and institution abilities in these trials? Understanding the variability at this level helps the physician interpret the data from individual trials and further understand the complexity and inherent risk of attempting to compare between trials.

Greason and colleagues9 in this issue of the Journal attempt to understand the different levels of variability in the PARTNERII A randomized trial, which compared TAVR and SAVR. The PARTNERII A trial was the second in the series of trials randomly allocating patients to receive the transcatheter SAPIEN valve (Edwards Lifesciences Corporation, Irvine, Calif) or SAVR and included patients at an intermediate risk for surgical mortality. The trial has been previously described in detail.4 Perhaps not as well recognized by those not intimately involved in the TAVR field is the fact that the earlier PARTNER SAPIEN trial (PARTNER Trial: Placement of AoRTic TraNscatheter Valve Trial), with cohorts PARTNER I A2 and PARTNER I B,5 was the first and only TAVR trial available in the United States for a number of years. Competition to get into these early PARTNER cohorts was extensive, and consequently the trial leadership basically got their pick of sites. With this came very stringent requirements for surgical expertise to be allowed into the trials, with all surgeons having at least 5 years of experience and a large volume of high-risk aortic valve replacement cases to their credit. These sites for the initial trials were all virgin to TAVR; by PARTNERII A, however, with most sites being retained from the first trials, experience with TAVR was already present. With this background, Greason and colleagues9 looked at mortality and stroke to see how the PARTNERII A trial might be confounded by intrasite and intersite performance variability.

PARTNERII A randomly allocated more than 2000 patients at intermediate surgical risk to undergo TAVR with the second generation SAPIEN XT or SAVR with valve choice and implant technique left up to the individual surgeon.2 The current study by Greason and colleagues9 used a mixed effect statistical model to separate variability into understood components. Level 1 was variability at the level of the individual patients, who were independent from one another and subject to the inclusion and exclusion criteria and randomization. Level 2 was intrasite variability, and level 3 was intersite variability. Levels 2 and 3 are
dependent on multiple, difficult to control factors over and above just operator skill. The strength of the local heart team, quality of the decision making during procedures, anesthesia, and postoperative care are included in the variables affecting outcomes at these levels. Greason and colleagues concluded from their findings that “inter-site variability was similar for mortality in SAVR and TAVR, but variability for stroke was greater for SAVR than for TAVR. Intra-site events were similar or both SAVR and TAVR. These findings suggest that in performance-based trials, site variability and its sources should be taken into account in analyzing and interpreting trial results.”

We agree with their conclusions, but we find several areas of importance to surgeons interested in trial design and interpretation. The first is that the inter-site hazard ratios for SAVR mortality ranged from 0.581 to 1.93. This would appear to us to be an exceptionally large spread for surgical high-volume centers of selected surgical excellence, where the number of trial SAVRs done each year represents a fraction of the overall SAVR institutional and even individual surgeon volume. Surgeons with a consistent hazard ratio of 1.93 at these institutions would likely be looking for a new job. The inter-site mortality hazard ration for TAVR ranges from 0.49 to 2.03, which also appears to be a very large range for what are now experienced TAVR sites. This is likely due to fairly small number, because the confidence levels overlapped 1 for all but 1 site. The inter-site stroke risk for SAVR showed a variability of 0.17, with individual hazard ratios ranging from 0.44 to 2.26. Almost all the variability was in the first 30 days. For TAVR, this variability was essentially zero.

With these data, how much inter-site variability is related to each level? At the patient level, case mix accounts for 5% for SAVR and 18% for TAVR variability. That case mix can be different and account for some variability is expected, even in a randomized trial, but it is comforting to see that it is a low amount. Intrasite variability was found to be small. Trial volume is of interest, especially with volume levels being currently discussed for the national coverage decision. Trials sites with less than 5 cases had increased early but not late mortality. Although this finding is interesting, we find it difficult to interpret, because these are all high-volume SAVR sites, and the trial SAVRs represent only a small fraction of the total SAVRs being done by each surgeon and institution. This would seem to us more likely related to error introduced by small numbers.

We applaud this extensive exploration of intersite and intrasite variability in the PARTNERII A trial. With the large spread in mortality hazard ratio, we believe that one of the most important points in this article is stated at the end of the section on unique challenges of performance-based procedure trials. Here, Greason and colleagues state, “Quantification of inter-site variability—and exploration of its sources and implications for trial interpretation—is warranted. For example, with a 4-fold inter-site variability in PARTNER-2A, it is likely that either superiority or non-inferiority thresholds may have been exceeded at some sites, but not others. Similarly, identically conducted trials may have different outcomes due to differences in site selection or magnitude of inter-site variability.” Understanding the limitations of randomized trials and the inability to compare different trials directly is important in medicine. Variability is a constant in medicine, despite our statistical attempts to remove it.

References