The paradox between randomized controlled trials and propensity score–matched real-world data: Moving from dissonance to dialog?

B. Zane Atkins, MD, and Gabriel S. Aldea, MD

Heart teams are increasingly challenged to inform individual patients with symptomatic aortic stenosis and high- or intermediate-risk features accurately to help select an optimal individualized treatment strategy. There is an emerging controversy regarding how to share and analyze available data most accurately. Is it best to report data from randomized, controlled trials (RCTs), or is it more accurate to present data from larger meta-analyses, propensity-scored matched controls, and real-world registry data and registries? Surprisingly, conclusions from these sets of data appear to be discordant.

Several foundational prospective RCTs reported comparable outcomes for transcatheter aortic valve replacement (TAVR) and surgical aortic valve replacement (SAVR). The PARTNER (Placement of Aortic Transcatheter Valve) 1 study1 screened 3105 patients; of these patients, 77% were screened out, 699 were randomly assigned to either TAVR or SAVR, and only 125 of those 699 patients (17.5%) were alive and had full follow-up to 5-years. The study reported equivalent 5-year mortalities for TAVR and SAVR (67.8% vs 62.4%). Stroke rates (10.4% vs 11.3%), myocardial infarction rates (2.9% vs 5.9%), mean valve area, mean gradient, and left ventricular mass index were also comparable for both therapies. The US CoreValve study2 screened 995 patients, randomly allocated 759, and reported equivalent 3-year outcomes for TAVR and SAVR with respect to mortality (32.9% vs 39.1%), stroke rate (8.1% vs 11.8%), and reintervention (2.4% vs 0.4%). A better effective orifice area was demonstrated for TAVR (1.79 vs 1.53 cm²), but TAVR also carried a higher incidence of moderate to severe paravalvular regurgitation (5.9% vs 0%). These studies, as well as comparable RCT studies of intermediate-risk patients, led to modification of American Heart Association and American College of Cardiology guidelines and European Society of Cardiology guidelines, which now present TAVR and SAVR as equally effective alternative therapies.

The article in this issue of the Journal by Armoiry and colleagues3 assessed real-world registry data by using the French Medical Information System to assess longer term (5-year) data in a larger patient cohort of propensity score–matched high-risk patients (n = 1598) with complete follow-up. Relative to SAVR, the study demonstrates a higher mortality for TAVR (51.5% vs 36.2%; hazard ratio [HR], 1.56; 95% confidence interval [CI], 1.1-2.5), and was comparable at 3 years to the US CoreValve study (33.7%), but mortality was lower than both RCTs for SAVR at 3 and 5 years (23.1% and 36.1%; P < .001). Relative to SAVR, TAVR also had higher HRs for stroke (1.64; 95% CI, 1.1-2.5), myocardial infarction (2.3; 95% CI, 1.1-4.7), and permanent pacemaker (HR, 2.4; 95% CI, 1.8-3.2). A trend toward reoperation was also noted (2.3% vs 1.1%; P = .09; HR, 2; 95% CI, 0.9-4.5). Finally, hospital stays (days) were comparable for the 2 interventions, but TAVR incurred a higher 5-year cost of approximately €13,500, presumably secondary to the differential costs of the valve. These dramatic results challenge many of the previously assumed ironclad assumptions of noninferiority.

The discordance among RCTs, propensity score–matched real-world registry data, and meta-analyses of observational trials of propensity-matched cohorts is increasingly
recognized. 4 We understand the limitations of these “older” TAVR RCTs conducted 5 and 7 years ago with older equipment and strategies. For TAVR, the first-generation Sapien valve (Edwards Lifesciences, Irvine, Calif) was replaced with the Sapien XT and more recently the Sapien 3, and the CoreValve (Medtronic Inc, Minneapolis, Minn) was replaced with the Evolut and Evolut R; smaller, more flexible devices were introduced, reducing the 30% incidence of nonfemoral access to approximately 5%; there was an impact of the initial learning curve; more accurate valve sizing was obtained with computed tomographic angiography; and better algorithms were introduced for patient selection and perioperative management. For SAVR, newer valve designs with lower gradients, more rapid deployment, and better durability were introduced; better myocardial protection techniques were instituted; and comparable advances were made in perioperative patient selection and therapies. By design, most RCTs limit study populations to very specific patient populations treated at higher volume centers. Similarly, propensity score–matched real-world registry data have the disadvantage of treating a broader, less-selective patient population and often rely (as in this study) on administrative data sets. Propensity score matching algorithms that aim to correct for variance and treatment bias are imperfect but have the advantage of assessing a numerically larger and clinically broader (less restricted) population. Although the differences in myocardial infarction and cerebrovascular accident rates are more likely to reflect a patient selection bias with different comorbidities (such as presence of CAD and atrial fibrillation), a mortality difference of 50% (HR, 1.5) deserves a pause and more in-depth consideration.

We conclude that differences in reported outcomes of RCTs and propensity score–matched real-world registry data, rather than being contradictory, reflect different aspects of patient care. Given these differences, it seems wiser to broaden the conversation and engage rather than negate contradictory data. We suggest the following:

- RCTs were designed to assess specific (limited) patient populations. Generalizations of findings to broader, less-selected populations should be avoided.
- Data from larger propensity score–matched real-world registry studies should inform design of future RCTs and challenge analyses of previous, more focused RCTs.
- Initial sample size may limit interpretation and certainty of longer term clinical outcomes. This is particularly relevant when combined primary end points are used to design the sample size analysis. For example, the PARTNER 3 trial was designed to assess a combined end point of death, neurologic deficits, and hospital readmission.
- These outcomes have varied clinical relevance and weights (death and cerebrovascular accident are more clinically significant than a hospital readmission).
- Longer term data of specific components of combined clinical end points may be underpowered.
- Strict definitions of valve degeneration2 need to be applied, because reintervention rates are poor surrogates for valve durability and performance.
- Both TAVR and SAVR technologies continue to evolve, and lessons learned from one treatment strategy can and should be thoughtfully translated and applied to the other.
- Most importantly, cardiovascular practitioners should engage rather than ignore discrepant data, keep an open mind, and be disciplined to reanalyze and challenge current understandings, data, and guidelines.

References