Randomised Trials and Propensity Score Analyses in Transcatheter Aortic Valve Replacement: How should we interpret the results?

Nick Freemantle [1, 2]

Domenico Pagano [2]

1 Institute of Clinical Trials and Methodology, University College London, London, UK
2 Quality Outcomes Research Unit, University Hospitals Birmingham NHS Trust, Birmingham, UK

Keywords

TAVR; CABG; clinical epidemiology; meta analysis

Address for Correspondence

Nick Freemantle PhD

Professor of Clinical Epidemiology & Biostatistics

Director

Comprehensive Clinical Trials Unit

University College London

90 High Holborn 2nd Floor

London WC1V 6LJ

UK

Email: nicholas.freemantle@ucl.ac.uk

Tel: +44 (0)20 3549 5017

Key Words

Randomised Trials, Propensity Score Analyses, Transcatheter Aortic Valve Replacement,

Introduction

Transcatheter aortic valve replacement (TAVR) is increasingly considered an established treatment for patients with severe symptomatic aortic stenosis who are at high risk of surgical mortality or who are not suitable for surgery [1,2]. Success in high risk patients has inevitably raised the potential to extend the use of TAVR technology to lower risk subjects, fuelled by both clinician innovation and commercial interests. The evidence base to support this development has included randomised trials and quasi experimental propensity score based analyses. In this paper we discuss the landmark available studies (summarised in Table 1), their strengths and limitations, and make some general evidence based recommendation on their interpretation.

[Table 1 Here]

Device development & regulation

A challenge of device development and regulation is that each current version of a Class 3 medical device (be it a pacemaker or a TAVR system) is effectively a prototype for future more advanced versions. Unlike the field of pharmaceutical regulation, where the form of a treatment is described in the patent and specified exactly through the broader regulatory process, devices often experience an ongoing development based on a series of incremental steps. Often individual innovation steps may seem superficial or simply additive (eg improving the durability of batteries in pacemakers), although a challenge exists in separating substantive development from less important, and a series of small steps may collectively raise real questions about differences in effectiveness and safety.

When to do a randomised trial

Randomised outcomes trials cost millions of dollars to conduct. They are normally required for regulatory purposes each time a new pharmaceutical is granted a marketing authorisation, although there are exceptions such as when a pharmaceutical is used in high risk patients and rare conditions [6] or where new treatment bridges from an existing formulation [7]. Similarly, Class 3 medical devices also often require randomised trials in order to support an application for marketing approval, however such devices may not require a new randomised trials each time a new device version is marketed. Edwards Life Sciences sponsored the PARTNERS 2 trial to evaluate the SAPIEN XT valve system compared with conventional surgery in patients with severe aortic stenosis and intermediate-risk clinical profiles [3]. SAPIEN XT differed from the previous SAPIEN having a thinner strut cobalt–chromium frame, a partially closed resting geometry of the bovine pericardial leaflets, the addition of a valve size that is 29 mm in diameter, and a reduced profile delivery catheter [3].

PARTNERS 2 randomised 2032 patients on a 1:1 basis between the two experimental conditions, and found a non significant reduction in the composite event of death from all cause or disabling stroke at 2 years, hazard ratio 0.89; 95% confidence interval [CI], 0.73 to 1.09; P = 0.25). The trialists' criterion for non inferiority was a hazard ratio upper confidence interval smaller than 1.2. Some may feel that this non inferiority boundary (equating to a risk difference in death or disabling stroke less than 4.2% on an absolute scale) to include values of the difference which are actually clinically relevant, although of note it is more rigorous than those applied to the two other studies.

The SAPIEN 3 Observational study included patients with moderate symptomatic aortic stenosis treated with the SAPIEN 3 value system which is described as differing from the previous XT system in that previous versions with improved geometry of the trileafelet bovine pericardial valve; different cobalt alloy frame, which is longer than the early version of the balloon-expandable valve

system (SAPIEN XT valve; Edwards Lifesciences) with more open outlet cells and denser inlet cells; a polyethylene terephthalate fabric skirt sewn to the bottom portion of the interior and exterior of the frame (providing an external circumferential seal to reduce paravalvular leak); four valve sizes (20 mm, 23 mm, 26 mm, and 29 mm diameters); and lower-profile delivery catheters with more precise valve positioning inserted through 14 or 16 French expandable sheaths for transfemoral access.[4]

Propensity Scores

Rather than conducting a further randomised trial, Edwards Lifesciences undertook a propensity score analysis comparing the SAPIEN 3 observational study patients with the surgical patients in PARTNERS 2.[4] [8] Unlike the more rigorous propensity score matched approach, the authors merely stratified patients into 5 quintiles by propensity score to address confounders between the comparator groups.

The propensity score approach was developed by Rosenbaum and Rubin[9] in order to provide an efficient method for quasi experimental comparison between treatments using non randomised comparative data. The method requires the calculation of a propensity score for each subject (the likelihood that a patient will receive the treatment of interest) derived from a logistic regression model including patients' characteristics as explanatory variables and exposure to the experimental therapy as the response or dependent variable. An imperfect instrument to account for bias, the propensity score relies upon the inclusion of the appropriate observed characteristics, and that there are no important omissions of those characteristics, such that exposure to treatment carries no additional risk compared to control apart from that derived from the comparison of treatment strategies. However, the additional risks (if any) associated with exposure are completely confounded with that exposure, and are thus a latency in the data set which cannot be elicited directly.[10] In other words, the key assumption of the propensity score, that any additional risks, other than the effect of treatment, are conditioned out by the propensity score cannot be measured directly, and may lead to substantial bias in the results. Bias is particularly likely when a treatment exposure (eg TAVR) is the result of an expert clinical judgement, as inevitably the complexity of such judgements will not, indeed often cannot, be captured by an observational dataset. This problem is discussed extensively in the context of aldosterone antagonists in heart failure in Freemantle et al.[10] and is a form of *confounding by indication*. There is also a useful discussion of systematic bias in propensity scores in acute coronary syndromes by Dahabreh et al[11]

In Figure 1, the main results of PARTNERS 2[3] and Thourani[4] on all-cause mortality and disabling stroke are contrasted. The difference between the results is striking and may strain plausibility, implying through indirect comparison that the SAPIEN 3 value system is significantly better than the SAPIEN X system on this outcome with a hazard ratio of 0.66 (95% CI 0.48 to 0.90; p=0.008). In other words, it seems unlikely that the propensity score stratification has accounted adequately for confounding in the comparison and instead patients with a relatively good prognosis have been recruited to the SAPIEN 3 Observational study.[4]

Non inferiority margins

Because of the probabilistic nature of estimation, it is challenging to demonstrate that there is no difference between two treatments; all comparisons are undertaken in the context of measurement error. The concept of non inferiority [12] is best understood in the context of confidence intervals.

A study excludes a risk that is outside the confidence interval. In clinical areas where there is regular need to undertake non inferiority trials the non inferiority boundary is a given, declared by the regulatory authority on the basis of informed clinical experience. For example in diabetes, the non inferiority margin in regulatory studies is HbA1c <0.3%,[13] which is widely recognised to be a clinically trivial value.

In cardiac surgery there is no prespecified non inferiority boundary, and we observe substantial variability among the individual criteria specified in trials. Thus for SURTAVI the non inferiority boundary was an absolute risk difference of 7%,[5] for PARTNERS 2 [3] it was effectively an absolute risk difference of 4.2% (although somewhat unhelpfully specified on the ratio scale as a hazard ratio of 1.2 and thus depended upon the rate of events in the control condition). These boundaries may be considered surprisingly wide and varied; nearly 7 more subjects in one hundred treated experiencing major morbidity or death may not be considered trivial by patients and clinicians interpreting the trials. Regulators could usefully take a stronger position on this point given the importance of the clinical area and the relative commonness of the intervention.

Randomised trials of TAVR versus Surgery

Ronald Fisher, the father of biostatistics, commented in 1935 that 'the simple act of randomisation assures the internal validity of the test for significance'.[14] However in order to benefit from this protection the act of randomisation must be preserved in implementation of the trial and the analysis. In comparative trials, subjects must be prepared to receive either intervention on offer in the trial, and clinicians must also be content that either may be used. This can prove challenging. In the two main trials of TAVR[3,5] the baseline characteristic of the included subjects highlight that these are samples which may be considered optimal for TAVR, with for example substantial rates of prior CABG.

The intention to treat principle preserves randomisation regardless of the treatment actually received, having the consequence 'that subjects allocated to a treatment group should be followed up, assessed and analysed as members of that group irrespective of their compliance to the planned course of treatment'.[15]

In SURTAVI[5] there was a substantial imbalance in the extent to which subjects randomised to each experimental condition (TAVR or Surgery) actually received that intervention with 1.7% of the TAVR not receiving the allocated intervention versus 8.2% of the Surgery patients not receiving that treatment ($p<.0001$). The authors incorrectly undertake a biased, *Modified Intention to Treat*, analysis which counts patients as having been randomised when randomised treatment is attempted. This approach is reasonably used in randomised double blind pharmaceutical trials where the knowledge of which treatment the subject will take is obfuscated by blinding. However, in the unblinded case such as the SURTAVI trial, the subjects' decision to undertake treatment (and their clinicians propensity to give them that treatment) may be mediated by the knowledge of the treatment on offer, as was clearly the case in SURTAVI[5] and results in bias. Fortunately, the authors also provided the analysis based upon conventional intention to treat.

Both the PARTNERS 2[3] and SURTAVI[5] included patients intermediate-risk patients with severe aortic stenosis to undergo either TAVR or surgical replacement. However, the characteristics of the patient population across both trials may not be representative of that patient group in practice, with for example a relatively high rate of prior CABG (21%)[3,5]. Thus despite broad inclusion

criteria, the patient population actually recruited to the trials appears to be selective of likely candidates for TAVR.

Interpretation of the Results of PARTNERS 2 and SURTAVI

Figure 2a to 2c describe the pooled 2 year results for the SURTAVI Intention to Treat population[5] and the PARTNERS 2 trial population,[3] for all cause mortality plus disabling stroke, and each component of the composite primary outcomes separately on the absolute risk difference scale.

Comment

There are several observations of note on the comparative data for TAVR and Surgery. First, these comparisons represent the difference in effectiveness of TAVR and Surgery in a patient population that appears to be selected as candidates for TAVR and thus may over estimate benefits in a less selected population. Second, in the major comparative randomised trials, the 95% confidence intervals on all three pooled results are quite wide. The results reasonably exclude a benefit of Surgery compared to TAVR of 0.7% on disabling stroke, but only 1.5% on the composite outcome of disabling Stroke or all cause mortality, and 2.2% on all cause mortality. Third, given Surgery is the established therapy, the rational criteria for replacing Surgery with TAVI should be superiority or at least non inferiority of TAVR over Surgery, and we might expect tougher criteria for non inferiority than we observe in trials. Fourth, given the increasing activity in randomised trials comparing established surgical procedures with new and less invasive technologies, the regulatory agencies should to take a stronger view of what effect may be considered as a non inferiority boundary in these trials. Professional organisations in collaboration with patient representative groups could also usefully address this question together.

Table 1. Overview of Major TAVR Studies

PARTNERS 2[3]

Design: 2032 intermediate-risk patients with severe aortic stenosis were randomised in an open label trial to TAVR or surgical aortic valve repair.

Primary end point: composite of death from any cause or disabling stroke at 24 months.

Non inferiority boundary: The upper boundary of the two-sided 95% confidence interval for the hazard ratio of the primary end point at 2 years was below a hazard ratio of 1.20.

Funding: Edwards Life Sciences

Thourani[4]

Design: 963 patients from the SAPIEN 3 Observational study at intermediate risk with severe, symptomatic aortic stenosis and treated with TAVR were compared with 747 surgical aortic valve repair patients from PARTNERS 2[4] with 1 year follow up using a propensity score methodology.

Primary end point: composite of death from any cause, all strokes, and incidence of moderate or severe aortic regurgitation.

Non inferiority boundary: Considered non inferior if the lower confidence interval on the primary outcome excludes an absolute risk difference of 7.5%.

Funding: No funding declared for these analyses; PARTNERS 2 and SAPIEN 3 were funded by Edwards Life Sciences

SURTAVI[5]

Design: 1746 patients at intermediate risk with severe, symptomatic aortic stenosis were randomised in an open label trial to TAVR or surgical aortic valve repair.

Primary end point: composite of death from any cause or disabling stroke at 24 months.

Non inferiority boundary: Considered non inferior if the lower confidence interval on the primary outcome excludes an absolute risk difference of 7%.

Funding: Medtronic Inc.

1. Joint Task Force on the Management of Valvular Heart Disease of the European Society of Cardiology, European Association for Cardio-Thoracic S, Vahanian A, et al. Guidelines on the management of valvular heart disease (version 2012). Eur Heart J 2012; 33: 2451–96.

2. Nishimura RA, Otto CM, Bonow RO, et al. 2014 AHA/ACC guideline for the management of patients with valvular heart disease: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 2014; 63: e57–185.

3. Leon MB, Smith CR, Mack MJ, Makkar RR, Svensson LG, Kodali SK, et al. Transcatheter or Surgical Aortic-Valve Replacement in Intermediate-Risk Patients. N Engl J Med 2016;374:1609-20.

4. Thourani VH, Kodali S, Makkar RR, Herrmann HC, Williams M, Babaliaros V, et al. Transcatheter aortic valve replacement versus surgical valve replacement in intermediate-risk patients: a propensity score analysis. Lancet 2016; 387: 2218–25

5. Reardon MJ, Van Mieghem NM, Popma JJ, Kleiman NS, Søndergaard L, Mumtaz M et al. Surgical or Transcatheter Aortic-Valve Replacement in Intermediate-Risk Patients. New England Journal of Medicine DOI: 10.1056/NEJMoa1700456

6. Hatswell AJ, Baio G, Berlin JA, Irs A, Freemantle N. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999–2014. BMJ Open 2016;6:e011666. doi:10.1136/bmjopen-2016-011666

7. https://www.accessdata.fda.gov/drugsatfda_docs/label/2015/206538lbl.pdf accessed 15/02/2018

8. Barili F, Freemantle N, Folliguet T, Muneretto C, De Bonis M, Czerny M et al. The flaws in the detail of an observational study on transcatheter aortic valve implantation versus surgical aortic valve replacement in intermediate-risks patients. Eur J Cardiothorac Surg 2017; doi:10.1093/ejcts/ezx058.

9. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal EffectS. Biometrika 1983; 70: 41-55.

10. Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I. Making inferences on treatment effects from Real World data? Propensity Scores, Confounding by Indication and other Perils for the Unwary in Observational Research. BMJ, 2013 2013;347:f6409 doi: 10.1136/bmj.f6409.

11. Dahabreh IJ, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. European Heart Journal (2012) 33, 1893–1901 doi:10.1093/eurheartj/ehs114

12. Mauri L, D'Agostino Sr RB. Challenges in the Design and Interpretation of Noninferiority Trials. Engl J Med 2017;377:1357-67. DOI: 10.1056/NEJMra1510063

13. Wangge G, Putzeist M, Knol MJ, et al. Regulatory Scientific Advice on Non-Inferiority Drug Trials. Mintzes B, ed. PLoS ONE. 2013;8(9):e74818. doi:10.1371/journal.pone.0074818.

14. Fisher, Ronald A. (1971) [1935]. The Design of Experiments (9th ed.). Macmillan. ISBN 0-02-844690-9.

15.http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf accessed 4/1/2018