The flaws in the detail of an observational study on TAVI vs SAVR in intermediate-risks patients

Authors: Fabio Barili 1, M.D. Ph.D., Nick Freemantle 2 Ph.D., Lorenzo Menicanti 3, M.D., Thierry Folliguet 4, M.D., Claudio Muleretto 5, M.D., Michele De Bonis 6, M.D., Martin Czerny 7, M.D., Jean Francois Obadia 8, M.D., Nawwar Al-Attar 9, M.D., Nikolas Bonaros 10, M.D., Jolanda Kluin 11, M.D., Roberto Lorusso 12, M.D. Ph.D., Prakash Punjabi 13, M.D., Rafael Sabada 14, M.D., Malakh Shrestha 15, M.D., Piotr Suvalski 16,17, M.D., Volkmann Falk 18, M.D., Miguel Sousa Uva 19, Pieter A. Kappetein 20, M.D. Ph.D., M.D., Alessandro Parolari 21, M.D. Ph.D.

Institutions:
1 Department of Cardiac Surgery, S. Croce Hospital, Cuneo, Italy
2 Comprehensive Clinical Trials Unit, Department of Primary Care and Population Health University College London, London, UK.
3 Department of Cardiac Surgery, IRCCS Policlinico S. Donato, University of Milan, Milan, Italy
4 Department of Cardiac Surgery, Centre Hospitalo-Universitaire Brabois ILCV, Nancy, France
5 Department of Cardiac Surgery, University of Brescia Medical School, Brescia, Italy
6 Department of Cardiac Surgery, S. Raffaele University Hospital, Milan, Italy
7 Department of Cardio-Vascular Surgery, University Hospital Freiburg, Germany
8 Department of Cardio-Thoracic Surgery, Hopital Cardiothoracique Louis Pradel, Lyon, France
9 Department of Cardiac Surgery, Golden Jubilee National Hospital, Glasgow, United Kingdom
10 Department of Cardiac Surgery, Innsbruck Medical University, Innsbruck, Austria.
11 Department of Cardiac Surgery, AMC, Amsterdam, Netherlands.
12 Department of Cardio-Thoracic Surgery, Heart & Vascular Centre - Maastricht University Medical Hospital, Maastricht, Netherlands
13 Department of Cardio-Thoracic Surgery, Imperial College Heathcare NHS Trust and Imperial College School of Medicine, London, United Kingdom
14 Department of Cardiac Surgery, Hospital de Navarra, Pamplona, Spain

15 Department of Cardio-thoracic, Transplantation and Vascular Surgery, Hannover Medical School, Hannover, Germany

16 Department of Cardiac Surgery, Central Teaching Hospital of the Ministry of the Interior, Warsaw, Poland

17 Pulaski University of Technology and Humanities, Radom, Poland

18 Department of Cardio-Thoracic Surgery, Deutsches Herzzentrum Berlin, Charite Berlin, Germany

19 Department of Cardiac Surgery, Hospital Cruz Vermelha, Lisbon, Portugal

20 Thoraxcenter, Erasmus MC, Rotterdam, Netherlands.

21 Unit of Cardiac Surgery and Translational Research, IRCCS Policlinico S. Donato, University of Milan, Italy

Corresponding Author: **Fabio Barili, M.D., PhD,**

Department of Cardiac Surgery, S. Croce Hospital

Via M. Coppino 26, 12100 Cuneo, Italy

Tel: +39 0171642571       Fax: +39 0171642064

Email: fabarili@libero.it       barili.f@ospedale.cuneo.it
General consideration

The development and availability of a trans-catheter approach for treating severe aortic valve stenosis (TAVI) has warranted clinical trials and observational studies to evaluate the safety and short/long term outcomes of newly designed prostheses in order to compare them with surgical aortic valve replacement (SAVR), the gold-standard treatment [1, 2]. The new treatment has been initially reserved for patients with absolute contraindications to surgery. Subsequently the evidence of safety of the new devices, as well as the maturation of experience with this technology, has led to the expansion of indications to higher-risk patients [3, 4]. Nonetheless, technology runs fast, and new prostheses are regularly launched on the market claiming better performances and wider indications and hence requiring new evidence [5]. The PARTNER group recently published a comparison between the latest-generation SAPIEN3 TAVI system (Edwards Lifesciences, Irvine, CA, USA) and SAVR in intermediate-risk patients, apparently demonstrating superiority of the TAVI and suggesting that TAVI might be the preferred treatment method in this risk-class of patients [6]. These favourable results of transcatheter approach in intermediate risk-patients could lead the decision-makers and the scientific community to consider TAVI as the new standard of care in a wider population of patients with severe aortic stenosis. The recent Food and Drug Administration (FDA) approval for expanded indications for SAPIEN 3 device based on their data somewhat supports this position [7].

Nonetheless [6], assessment of the non-randomised methodology used in this comparison reveals challenges which should be addressed in order to elucidate the validity of the results. The study is observational, employing propensity scores (PS), risk scores that can be used to match patients with a similar likelihood of receiving treatment [8-10], since non-random differences in baseline will lead to bias in comparisons between treatment conditions [10]. PS analysis can be used to create a “quasi-randomized” comparison, but the approach has well-known intrinsic limitations and pitfalls including the misspecification of the PS, effects of unknown biases and confounding by indication [10-14]. Hence, unlike properly randomised trials, the use of the PS does not assure the internal validity of the analyses, and decision-makers and the scientific communities need to be wary of making inference from their results [12]. The PS study by Thourani and Colleagues has a number of major design flaws, and its results have clear signs of bias [6].
The assumption of “ignorability” and the effects of propensity score misspecification.

The first important step in PS analysis is the careful specification of the risk algorithm, as omission of important confounding factors (eg getting it wrong) will lead to biased estimation of treatment effect. The objective is that as a result of the PS conditioning of the relevant explanatory variables, the treatment will be independent of potential outcomes. This conditional independence assumption is called “ignorability”, “unconfoundedness”, “selection on observables” and crucially it is always held as an assumption, because it is not directly testable [15]. In order to assume that treatment assignment is “otherwise ignorable” [10-16], the very first step is the inclusion in the PS algorithm of all known and available confounding factors, as explanatory variables that meet the condition of affecting both treatment assignment and outcome confound the observed relationship between treatment and outcome [10, 16]. The propensity score is compromised when important variables influencing selection have not been collected or considered and misspecification of the propensity score by excluding known confounders has been demonstrated to lead to largely biased results [11].

The study by Thourani and Colleagues was designed to compare outcomes of an observational study on the latest-generation SAPIEN 3 TAVI System (Edwards Lifesciences, Irvine, CA, USA) with results of the surgical group of the PARTNER 2A trial [5, 6, 17]. The two groups were not homogeneous, as shown in baseline characteristics [6] Thourani and Colleagues planned propensity score stratification before analysing outcomes. The use of PS stratification rather than precise matching is surprising as it is by design limited in the extent to which systematic differences between the comparator groups may be accounted for. Indeed there were important differences between the comparator samples. The comparative analysis of patients’ baseline characteristics and baseline variables included in the PS algorithm showed that the most significantly different characteristics between the two groups (left ventricular ejection fraction LVEF, p-value <0.0001; STS score, p-value 0.0002; moderate or severe mitral regurgitation, p-value <0.0001) were omitted in the PS generation, together with other significant factors (frail condition and mean gradient). These different baseline characteristics are well-known predictors of early and late mortality [18-25] and hence, affecting both treatment assignment and outcomes, are major confounders that should be included in the PS. Their omission may violate the “ignorability” assumption and, consequently, may lead to bias.
Moreover, further potential confounders not collected in the study are the associated procedures, such as myocardial revascularization. They increase the risk of perioperative mortality and morbidity as widely demonstrated by STS score and EuroSCORE [18-28], and they could represent a major confounder to be included in the PS algorithm. Nonetheless, although patients with non-complex coronary disease requiring revascularization were included if a treatment plan for the coronary disease was agreed before enrolment [5, 6, 17], no information on associated myocardial revascularization in the TAVI group has been reported [6, 17]. Some information on the SAVR group can be derived from the published PARTNER 2A trial [5] where a total of 86 of 944 patients (9.1%) had concomitant procedures during surgery and 137 of 944 patients (14.5%) underwent associated coronary artery bypass grafting (CABG) [5]. Thus a proportion ranging between 14.5% and 23.6% had concomitant surgical procedures in the SAVR group of the PARTNER 2A trial, indicating an increased risk of mortality and morbidity and potentially a major confounder. The need for a deep analysis on associated procedures in the Thourani’s study is also strengthened by the significantly different proportion of myocardial revascularization in the PARTNER 2A trial (137/994, 14.5% in the SAVR; 39/994, 3.9% in the TAVI group; Chi-square p-value <0.0001) [5].

**Confounding by indication and assessing the performance of the propensity score.**

Confounding by indication is the situation where, although all known confounders have been balanced, allocation to treatment is not otherwise ignorable but instead subject to some latent (unrecognized or unmeasured) process associated with those who are treated. This confounding cannot be measured directly but only tangentially through its effects and hence the effort should be focused on performance analysis of PS [12].

The first useful precaution against unsafe inference from an observational study is to compare it with a known treatment effect and bridge from that point to consider further questions. A deeper step in diagnostic should be the evaluation of PS performance through testing the potential heterogeneity of the treatment effect across the range of the PS. A comparison between two well-balanced groups should lead to a homogeneous treatment effect across the range of the PS, while heterogeneous effects will raise concern.

The treatment effect of the observational study by Thourani and Colleagues [6] can be compared to the PARTNER 2A randomized trial [5]. As shown in Figure 1, the relative risk of the main outcome (all-
cause death or disabling stroke) significantly differs from the two studies (interaction p-value = 0.0001), which militates against drawing strong conclusions in the observational study. Moreover, a deeper analysis of the treatment effect across the PS quintiles shows that the treatment effect is not homogeneous across classes, showing a decreasing pattern through strata (Figure 2). Only the treatment effect in the fifth quintile is similar to the PARTNER 2A trial effect. It can be hypothesized that in patients with low likelihood of TAVI (lower quintiles of PS) there is important information that the PS did not capture and so the match was made with inappropriately low risk individuals, leading to a not otherwise ignorable treatment assignment.

[12]

To adjust or not to adjust, this is another question.

The concerns also increase in the second part of the study, the time-to-event analyses. The study is based on evidence that groups are different and biased estimated of treatment effects needs to be accounted for by balancing the covariates with PS methods [6]. Nonetheless, after employing PS stratification for comparing dichotomic outcomes, the Authors surprisingly did not undertake any type of adjustment in time-to-event analysis and presented simple unadjusted Kaplan-Meier estimates and curves, making inference on their results [6]. This is counterintuitive and the curves are not interpretable, as they are simply a first-step evaluation before adjustment. Stating in results “important differences between TAVR and surgery for each endpoint are observed in the first several months” is inappropriate until results are confirmed by adjusted results. Making inference on unadjusted outcomes derived from biased groups should be avoided [10, 14].

Is there an outcome missing?

In the PARTNER 2 SAPIEN 3 observation study, clinical outcomes were reported as defined by Valve Academic Research Consortium (VARC)-2 definitions [6, 29]. The VARC-2 definitions recommend capturing the cause of death with a careful review and, among mortality causes to be reported, all valve-related deaths are included. Valve-related mortality and morbidity represent the main outcomes to evaluate the safety and short/long-term follow-up after valvular treatment, as it is the most specific index of early-late performance. In a comparison between two treatment options for valvular disease considering two homogeneous groups, we might reasonably expect to observe a similar non-cardiovascular and cardiac non-
valve-related mortality, while the treatment effect would be expressed in differences in valve-related mortality [30]. Nonetheless, in the PARTNER 2 SAPIEN 3 observation study only all-cause mortality, non-cardiac and cardiac death were reported, with no information on valve-related mortality shown. Therefore as it is not possible to differentiate prostheses-related events from prostheses-unrelated deaths, such as these caused by non-embolic myocardial infarction, defined as cardiac but non-valve-related death [29,30]. In the Thourani’s study, 30-day non-cardiac mortality is higher in the surgical group as shown in the Appendix (0.1% and 1.1% in the TAVR and surgical group respectively, Chi-square p-value 0.0152) which suggests the presence of confounding by indication.

Conclusions

As shown, the study on the comparison between SAPIEN3 TAVR and surgical AVR [6] has demonstrated several major methodological concerns:

- suboptimal methods in propensity score analysis with evident misspecification of the PS (no adjustment for the most significantly different covariates: LVEF, moderate-severe MR, associated procedures)
- Use of PS quintiles rather than matching
- inference on not-adjusted Kaplan-Meier curves, although the Authors correctly claimed for the need of balancing score for adjusting for confounding factors in order to have unbiased estimates of the treatment effect
- evidence of poor fit
- lack of data on valve-related death

These methodological flaws invalidate direct comparison between treatments and cannot support Authors’ conclusions that TAVI with SAPIEN 3 in intermediate-risk is superior to surgery and might be the preferred treatment alternative to surgery.
Acknowledgment

Sources of funding: None

Authors’ contributions:

All Authors participated to conception of the manuscript, drafted and revised the article and gave their final approval to the text.

Conflict of interest:
References


30. Akins CW, Miller DC, Turina MI, Kouchoukos NT, Blackstone EH, Grunkemeier GL, Takkenberg JJ, David TE, Butchart EG, Adams DH, Shahian DM, Hagl S, Mayer JE, Lytle BW; Councils of the American Association for Thoracic Surgery; Society of Thoracic Surgeons; European Association for Cardio-Thoracic
Figures legend

Figure 1. Treatment effect of TAVR vs Surgery on all-cause mortality and stroke in PARTNER 2A randomized trial and PARTNER 2A SAPIEN 3 observational study.

Figure 2. Treatment effect of TAVR vs Surgery on composite outcome (death, stroke and moderate or severe aortic regurgitation at 1 year) across the quintiles of propensity score in the PARTNER 2A SAPIEN 3 observational study.
RELATIVE RISK TAVR/SURGERY (ALL-CAUSE DEATH OR STROKE)

PARTNER 2A SAPIEN 3 Study
Thourani 2016

0.57 (0.47 -0.70, p-value < 0.0001)
324/2121

PARTNER 2A Trial
Leon 2016

0.94 (0.80 -1.12, p-value = 0.5112)
416/2032

INTERACTION P-VALUE <0.0001

RR (95% CI; p-value)
events / total

1 Lancet; 2016; 387: 2218–25
### ODDS RATIO TAVR/SURGERY
(Death, Stroke, Moderate-Severe Aortic Regurgitation)

<table>
<thead>
<tr>
<th>Quintiles of propensity score</th>
<th>ATT weight</th>
<th>ODDS RATIO</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quintile 1</td>
<td>14.00%</td>
<td>0.41 [0.23, 0.72]</td>
<td>0.002</td>
</tr>
<tr>
<td>Quintile 2</td>
<td>18.00%</td>
<td>0.37 [0.20, 0.69]</td>
<td>0.002</td>
</tr>
<tr>
<td>Quintile 3</td>
<td>20.00%</td>
<td>0.49 [0.26, 0.89]</td>
<td>0.020</td>
</tr>
<tr>
<td>Quintile 4</td>
<td>23.00%</td>
<td>0.57 [0.33, 1.00]</td>
<td>0.051</td>
</tr>
<tr>
<td>Quintile 5</td>
<td>25.00%</td>
<td>0.74 [0.41, 1.34]</td>
<td>0.317</td>
</tr>
</tbody>
</table>

Overall (weighted) 100.00% 0.52 [0.40, 0.68] < 0.001

**ADVANTAGE TAVR**

---

**ADVANTAGE SURGERY**

---

0.14 0.37 1 1.65