Do all roads lead to Rome?

Critical reassessment of the individual patient meta-analysis on bypass grafts by Gaudino et al.

Freemantle N[1], Myers PO[2], Siepe M[3]

1. Institute for Clinical Trials and Methodology, University College London, London UK WC1V 6LJ
2. Division of Cardiac Surgery, CHUV, Lausanne University Hospital, Switzerland; La Tour Hospital, Geneva, Switzerland
3. Department of Cardiac Surgery, University Hospital Bern, University of Bern, Bern, Switzerland

Corresponding author:

Matthias Siepe, MD

Chairman and Professor

Department of Cardiac Surgery

Cardiovascular Center, Inselspital

Freiburgstrasse 18

3010 Bern

Phone: +41 31 664 26 00

E-Mail: matthias.siepe@insel.ch

Introduction

The recent individual patient data meta-analysis by Gaudino et al. [1], comparing outcomes with different grafts for coronary artery surgery from non-randomized trials, was met with questions on the methodology given that some of the results proved controversial.

The EJCTS reacted to various comments on the methodology and suggested an independent re-analysis of the data. The author group of the paper, led by Mario Gaudino, should be applauded for their scientifically sound cooperation in this exercise. In the spirit of open data and open science, the authors shared their entire dataset and the codes with the journal. The authors of this Editorial wish to express their gratitude and esteem for this behaviour which exemplifies excellence in science.

The Journal was particularly keen to have this process initiated because the topic is important, and the main players have considerable expertise in CABG research and statistics.

Data Reanalysis

Importantly, the authors used a legitimate way to analyse the data and the outline statistical analysis plan was predefined and registered together with another paper [2]. When performing the same analysis, we came to similar results, confirming that the data handling was numerically and statistically correct.

During the reanalysis process, we identified imbalances in the matching, which we preferred to analyse conventionally in separate propensity score matching for the two comparisons. Utilising two separate comparisons (BITA versus LITA+RA; BITA versus LITA+SVG), we achieved excellent balanced characteristics on observation of the matched baseline data in both comparisons (Tables 1&2).

Unlike the findings of Gaudino et al, we found no systematic difference between the groups on the important outcome of mortality (Figure 1). With the hazard ratio on mortality for BITA versus LITA+RA of HR 0.93 (95% CI 0.69 to 1.26; p= 0.66); and for BITA versus LITA+SVG of HR 1.02 (0.86 to 1.22; p= 0.81). The point of making two separate comparisons in the context of three separate treatment strategies of interest is to utilise conventionally the statistical principle of transitivity, that is if x>y and y>z then we can infer that x>z. This approach also matches one of the scenarios provided in the motivating example by Guadino and colleagues [1].

Comment

Different approaches to analysis can lead to qualitatively different results. In a fascinating and relevant study, Silberzahn et al. (2018) [3] found that analyses conducted independently by 29 analytical teams found important differences, with point estimates varying qualitatively (e.g. moving from increased risk to decreased risk in the vexed question of whether soccer referees were more likely to 'red card' dark-skin-toned players). Some 69% found a significant positive relationship, while the remainder did not.

Propensity score matched analyses use a statistical model (a logistic regression) to calculate a likelihood based score (described on the logit, or log(e)odds, scale) for each subject to identify their 'risk' of being treated with the strategy of interest, accounting for a range of potential patient level characteristics. Rosenbaum and Rubin who first developed the approach described how the method was unbiased when, having accounted for the propensity score, the actual exposure to the treatment of interest carried no extra information on the risk of the subject [4]. Therefore, in order to give the right answer, the propensity score has to capture all of the risks faced by a subject regardless of whether they received the treatment of interest or not.

Achieving the perfect adjustment described by Rosenbaum and Rubin is not practically possible. We have imperfect knowledge on the risks faced by individuals and we measure the available risk factors imperfectly. For example, a complex clinical process such as diabetes may be described as simply 'present' or 'absent'. A strength, but also a weakness, of the propensity score approach is that we can include many explanatory variables in the creation of the score without being concerned with technical issues such as overfitting because it is only the point estimate that is used from the model, not its uncertainty. While this allows us to include all relevant risk factors, it also facilitates the dilution of an important risk factor in the score because there are many less important risks included. An important risk factor may be hard to identify because statistical significance alone will not guide us; instead, we should assess the appropriateness of the achieved match between cases and controls, and only proceed to analysis of outcomes when we are content that the groups are closely alike.

While different analysis strategies are possible, propensity score matched analyses are attractive because they limit comparisons to subjects who are all potential candidates for the treatment of

interest, at least as described by the observed characteristics and thus the propensity score, rather than including subjects who could never be included. This is nicely described for example by work comparing TAVI and SAVR, where the matched subjects tend to be in the middle range of risk, [5] with lower risk subjects in the data set receiving SAVR, and high risk subjects receiving TAVI, the latter particularly are very unlikely to be candidates for SAVR and thus really have no place in a statistical model comparing the two strategies.

Because of concerns about the effectiveness of the matching process, we use approaches such as the standardised mean difference (SMD) to assess how similar the resulting groups are over each important subject characteristic. The SMD is helpful because, unlike a statistical test, it does not rely on having sufficient numbers and thus power. Threshold values of the SMD that are considered adequate are arbitrary, but we often look for values <.1 to suggest that a good match has been achieved.

The authors mentioned that the statistical plan was prespecified. However, the protection from bias of prespecification is rather less for observational based approaches compared with randomised comparisons (even when the data for the non-randomised comparisons come from randomised trials), as the risk of 'getting it wrong' is substantially greater for observational based approaches [6] than for randomised comparisons. The solution where possible is to contrast observational analysis with randomised trials using a target trial emulation approach [7] and use these as a cautious bridge into new analytic ground. That said, prespecification is very helpful as it may avoid the lure of an interesting answer. However, the prespecified analysis should be subject to considerable supportive or sensitivity analyses to establish its robustness to other reasonable strategies. The analyses we have conducted should be considered part of that validation strategy.

The analytic approach undertaken by the authors (contrasting three strategies in a single match, and using matching with replacement) does bring some extra challenges to pitch against the clear benefits of having a level playing field for comparison across all three strategies. While the standardised mean difference is quite helpful in identifying imbalance there is a potential challenge applying standard criteria for the SMD over 3 groups since an imbalance in one group may be somewhat obfuscated by similarity in the other two. Also, standard methods SMD for categorical variables (eg those with >2 levels) can give quite surprising results which may not reflect the importance of observed baseline differences. Replicating the matching achieved by the authors led to the identification of several systematic imbalances in the baseline characteristics of subjects and thus we question how well they are matched.

While the authors argue for the merits of a randomised trial to address this question, it is not clear from re-analysis of these data that this might come to firm conclusions if it is to consider mortality as the primary end point. However, if substantial clinical uncertainty remains, there is no substitute for a properly randomised and conducted trial to address this question. Such a trial will however need to be large.

Conclusion

The analyses presented by the authors are not under criticism as the approach is scientifically sound and the results of the implementation of the statistical methodology appear correct. We have conducted additional sensitivity analyses, which approach the question in slightly different although more orthodox ways, and we have not found the substantial differences in clinical outcome identified by the authors in their analyses. Firm conclusions await the results of properly conducted

randomised trials. We thank the authors for participating in this fascinating exercise in open science. We strongly believe that the overall rigor of scientific publishing profits from open data exchange and interpretation.

Conflict of interest:

NF declares that UCL receives a grant from EACTS to provide methodological and educational input, and that this work was conducted as part of that arrangement.

PM declares that he is Secretary General of EACTS.

MS declares that he is Editor-in-Chief of this Journal and Trustee of EACTS.

References

1. Gaudino M, Audisio K, Di Franco A, Alexander JH, Kurlansky P, Boening A et al. Radial artery versus saphenous vein versus right internal thoracic artery for coronary artery bypass grafting. European Journal of Cardio-Thoracic Surgery 2022, 62(1), ezac345
2. Gaudino M, Di Franco A, Alexander JH, Bakaeen F, Egorova N, Kurlansky P, et al. Sex differences in outcomes after coronary artery bypass grafting: a pooled analysis of individual patient data. European Heart Journal 2022; 43: 18-28
3. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. Advances in Methods and Practices in Psychological Science 2018, Vol. 1(3) 337 –356.
4. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 1983; 70: 41-55.
5. Beyersdorf F, Bauer T, Freemantle N, Walther T, Frerker C, Herrmann E et al. Five-year outcome in 18 010 patients from the German Aortic Valve Registry . Eur J Cardiothorac Surg 2021; doi:10.1093/ejcts/ezab216.
6. 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.
7. Gomes M, Latimer N, Soares M, Dias S, Baio G, Freemantle N, Dawoud D, Wailoo A, Grieve R. Target Trial Emulation for Transparent and Robust Estimation of Treatment Effects for Health Technology Assessment Using Real-World Data: Opportunities and Challenges. PharmacoEconomics (2022) 40:577–586

Table 1. BITA vs LITA-RA: Baseline characteristics for matched and unmatched groups

	Unmatched			Matched		
Characteristic	BITA (n=1510)	LITA-RA (n=1385)	SMD	BITA (n=595)	LITA-RA (n=595)	SMD
Female (%)	210 (13.9%)	185 (13.4%)	0.02	78 (13.1%)	93 (15.6%)	0.07
NYHA (%)	343 (25.9%)	209 (19.4%)	0.15	135 (22.7%)	135 (22.7%)	0
MI (%)	598 (39.6%)	501 (36.2%)	0.07	244 (41.0%)	39.5%)	0.03
PTCA (%)	221 (16.7%)	124 (11.5%)	0.15	87 (14.6%)	70 (11.8%)	0.08
Hypertension	1124 (74.4%)	919 (66.5%)	0.17	447 (75.1%)	444 (74.6%)	0.01
Diabetes	332 (22.0%)	499 (36.0%)	0.31	166 (27.9%)	172 (28.9%)	0.02
Renal Insufficiency	17 (1.1%)	23 (1.7%)	0.05	7 (1.2%)	3 (0.5%)	0.07
CVA	63 (4.8%)	57 (5.3%)	0.02	39 (6.6%)	39 (6.6%)	0
PVD	112 (8.4%)	79 (7.3%)	0.04	56 (9.4%)	58 (9.8%)	0.01
LVEF	429 (32.3%)	668 (62.0%)	0.62	286 (48.1%)	296 (49.8%)	0.03
Off Pump	591 (39.1%)	430 (31.1%)	0.17	221 (37.1%)	224 (37.7%)	0.01
Age (Median, IQR)	63.6 (57.1, 69.8)	65.0 (59.5, 70.3)	0.11	64.4 (58.2, 70.8)	64.3 (58.3, 71.0)	0.03
Creatinine (Median, IQR)	92.0 (80.0, 106.1)	84.0 (70.7, 101.0)	0.25	90.0 (79.6, 106.1)	89 (79.0, 104.0)	0.09
Grafts (Median, IQR)	3 (3, 4)	3 (3, 4)	0.29	3 (3, 4)	3 (3, 4)	0.05

Table 2. BITA vs LITA-SVG: Baseline characteristics for matched and unmatched groups

	Unmatched			Matched		
Characteristic	BITA (n=1510)	LITA-SVG (n=7361)	SMD	BITA (n=1273)	LITA-SVG (n=1273)	SMD
Female (%)	210 (13.9%)	1396 (19.0%)	0.14	185 (14.5%)	183 (14.4%)	0.00
NYHA (%)	343 (25.9%)	1382 (25.3%)	0.01	319 (25.1%)	309 (24.3%)	0.02
MI (%)	598 (39.6%)	2847 (38.7%)	0.02	508 (39.9%)	535 (42.0%)	0.04
PTCA (%)	221 (16.7%)	1176 (16.3%)	0.01	212 (16.7%)	212 (16.7%)	0
Hypertension	1124 (74.4%)	5587 (76.8%)	0.05	988 (77.6%)	993 (78.4%)	0.02
Diabetes	332 (22.0%)	2943 (40.0%)	0.40	312 (24.5%)	299 (23.4%)	0.03
Renal Insufficiency	17 (1.1%)	173 (2.4%)	0.09	9 (0.7%)	5 (0.4%)	0.04
CVA	63 (4.8%)	726 (10.0%)	0.20	60 (4.7%)	57 (4.5%)	0.01
PVD	112 (8.4%)	643 (8.9%)	0.02	101 (7.9%)	111 (8.7%)	0.03
LVEF	429 (32.3%)	4605 (63.7%)	0.66	423 (33.2%)	450 (35.4%)	0.04
Off Pump	591 (39.1%)	2662 (36.2%)	0.06	552 (43.4%)	525 (41.2%)	0.04
Age (Median, IQR)	63.6 (57.1, 69.8)	66.0 (60.0, 72.0)	0.28	64.3 (58.1, 70.5)	64.8 (58.1, 71.0)	0.03
Creatinine (Median, IQR)	92.0 (80.0, 106.1)	88.4 (79.6, 106.1)	0.01	92.0 (80.0, 106.0)	93.0 (81.0, 106.0)	0.01
Grafts (Median, IQR)	3 (3, 4)	3 (3, 4)	0.06	3 (3, 4)	3 (3, 4)	0.01