









# Critical Reading of cardiovascular trials with neutral or negative results

Claudio Rapezzi <sup>1,2†</sup>, Alberto Aimo <sup>3,4\*</sup>, Iacopo Fabiani <sup>4</sup>, Vincenzo Castiglione <sup>3,4</sup>, Roberto Ferrari<sup>1</sup>, Aldo Pietro Maggioni<sup>5</sup>, and Luigi Tavazzi <sup>2</sup>

<sup>1</sup>Cardiology Centre, Università degli studi di Ferrara, via Ludovico Ariosto, 35 - 44121 Ferrara, Italy; <sup>2</sup>GVM Care & Research, Maria Cecilia Hospital, Via Corriera, 1, 48033 Cotignola (Ravenna), Italy; <sup>3</sup>Scuola Superiore Sant'Anna, Piazza Martiri della Libertà 33, Pisa 56127, Italy; <sup>4</sup>Cardiology Division, Fondazione Toscana Gabriele Monasterio, Piazza Martiri della Libertà 33, Pisa 56127, Italy; and <sup>5</sup>Centro studi ANMCO, Via La Marmora 36, 50121 Firenze, Italy

Online publish-ahead-of-print 22 June 2023

## Graphical Abstract

Critical reading of cardiovascular trials with neutral or negative results					
Design	Execution	p-value	Interpretation	Possible causes	Possible examples
Good 	Good 	Far from 0.05	Treatment does not work	-	RED-HF
Good 	Good 	Close to 0.05	Assessment of all available evidence	-	PARAGON-HF
Good 	Difficult 	-	(inconclusive findings)	Protocol violations Premature interruption COVID-19 pandemic	TOPCAT GUIDE-HF
Some problems (good endpoint selection, insufficient power) 	-	-	(inconclusive findings)	Lower than expected event rates Smaller than expected effect size Follow-up period too short High rate of cross-over, treatment discontinuation or loss to FU	FLOWER-MI DOREMI STICH/STICHES CABANA, LIFE, CASTLE-AF
Some problems (suboptimal endpoint selection) 	-	-	(inconclusive findings)	Number of events "hypertrophied" by minor events or medical decisions Suboptimal endpoint definition	PROACTIVE OVERTURE

Proposed framework for the interpretation of superiority trials with neutral or negative results. See text for details. CABANA, catheter ablation vs. antiarrhythmic drug therapy for atrial fibrillation; CASTLE-AF, catheter ablation for atrial fibrillation with heart failure; COVID-19, coronavirus disease-19; DOREMI, dobutamine compared with milrinone; FLOWER-MI, flow evaluation to guide revascularization in multivessel ST-elevation myocardial infarction; FU, follow-up; GUIDE-HF, haemodynamic-GUIDEd management of heart failure; LIFE, LCZ696 in advanced heart failure; OVERTURE, Omapatrilat Versus Enalapril Randomized Trial of Utility in Reducing Events; PARAGON-HF, prospective comparison of ARNI with ARB global outcomes in HF with preserved ejection fraction; PROACTIVE, PROspective pioglitAzone Clinical Trial In macroVascular Events; RED-HF, reduction of events by darbepoetin alfa in heart failure; STICH, surgical treatment for ischaemic heart failure; STICHES, STICH Extension Study; TOPCAT, treatment of preserved cardiac function heart failure with an aldosterone antagonist.

\* Corresponding author. Tel: +39 050 3153521, +393477084391, Fax: +39 050 3152109, Email: [a.aimo@santannapisa.it](mailto:a.aimo@santannapisa.it); [aimoalb@ftgm.it](mailto:aimoalb@ftgm.it)

† Deceased.

Randomized controlled trials (RCTs) may change the standard of care of patients sharing the characteristics of trial participants. However, well-designed trials with neutral (i.e. where no significant difference between the treatment and its comparator) or negative findings (i.e. where the comparator seems more effective and/or the intervention caused harm) deserve scientific scrutiny to understand why the primary endpoint was not met. It is particularly important to differentiate well-designed and well-conducted RCTs from RCTs burdened by methodological issues. We propose a general framework to the interpretation of neutral or negative trials based on this differentiation and then on *P*-values (for well-designed and well-conducted RCTs) or on the identification of the possible methodological issue(s). We will focus on superiority trials, which assess whether an investigational approach is superior to the standard of care. Non-inferiority and equivalence trials are becoming increasingly important, but are relying on different statistical assumptions, and the possible reasons for type I errors (i.e. the erroneous acceptance of an inferior new treatment), type II errors (i.e. the erroneous rejection of a truly non-inferior treatment), or truly neutral findings would require an extensive discussion, and possibly a dedicated paper.

## Grid of interpretative hypotheses

### Good study design and execution, *P*-values far from significant

When a trial is adequately designed and conducted, but the *P*-value for the primary endpoint is far from significance, the reasonable conclusion is that the treatment is not more effective than its comparator in this specific setting. A possible example is the RED-HF trial (assessing the effects of darbepoetin alfa on clinical outcomes in patients with systolic heart failure—HF—and anaemia),<sup>1</sup> which was well designed and adequately powered.

### Good study design and execution, *P*-values close to significance

The *P*-value for the primary endpoint may approach the conventional threshold of  $P = .05$ , leading to uncertainties about result interpretation. A notable example is the PARAGON-HF trial ( $P = .06$ ),<sup>2</sup> testing sacubitril/valsartan in HF with preserved ejection fraction (HFpEF). The approach pursued by the Food and Drugs Administration when appraising the results of PARAGON-HF was to interpret the *P*-value for centrally adjudicated primary endpoint events on the light of investigator-reported events, and a subgroup analysis suggesting a greater efficacy for EF values  $<57\%$ . As 'the totality of information was positive', the Food and Drugs Administration granted an expanded indication to HFpEF patients with EF below normal.<sup>3</sup> This pragmatic approach seems reasonable, more than the simple categorical conclusion of non-superiority of sacubitril/valsartan reported in the first publication of results of PARAGON-HF.<sup>3</sup>

### Good study design, difficult execution

A well-known example of difficult trial execution is the TOPCAT trial, assessing spironolactone in patients with HFpEF. Patients enrolled in Russia or Georgia had a markedly lower event rate than patients from the Americas, and did not show any benefit from spironolactone.<sup>4</sup> The likely reason was severe and systematic adherence to protocol violations by study investigators in Russia and Georgia.<sup>5</sup>

Except when there is clear evidence of superiority at an interim analysis, all premature discontinuations lead to insufficient power. Possible

causes of early termination include no evidence of efficacy at interim analysis or a signal of harm, slow enrolment, or strategic decision of the sponsor. One of the many possible examples is the SOLOIST-WHF trial testing sotagliflozin in patients with diabetes and a recent HF decompensation. This trial 'ended early because of loss of funding from the sponsor'.<sup>6</sup>

The COVID-19 pandemic was an unexpected event leading to the premature discontinuation of many trials and affected the conduction of thousands of other trials. Besides discontinuation, the impact of the COVID-19 pandemics on the final study results was sometimes critical. An example is the GUIDE-HF trial comparing a haemodynamic-guided management (treatment) vs. usual care (control) in HF patients with a primary endpoint of all-cause mortality plus HF events. Before COVID-19, the primary endpoint rate was 0.553 vs. 0.682 events/patient-year in the treatment vs. control group (HR 0.81,  $P = .049$ ). No difference was evident during COVID-19 (HR 1.11,  $P = .526$ ) following a reduction in HF events possibly because of a 'fear-based avoidance of seeking healthcare services [...] hoping to avoid exposure to COVID-19'.<sup>7</sup>

### Problems in study design: good endpoint selection but insufficient power

Insufficient power may derive from (i) lower than expected event rates, (ii) smaller than expected effect size, (iii) a follow-up period too short to allow the benefit from treatment to become evident, or (iv) a high rate of cross-over, treatment discontinuation, or loss to follow-up.

In the FLOWER-MI trial, the expected rate of the primary outcome at 1 year was 10% with the fractional flow reserve-guided strategy vs. 15% with the angiography-guided strategy, while the observed rate was 6% and 4%, respectively. This discrepancy did not emerge before study termination since no interim analysis was planned.<sup>8</sup>

When the effect size is lower than expected, the trials are not properly underpowered, but rather appropriately powered for a specific effect size, which was not estimated correctly. For example, the DOREMI trial on milrinone vs. dobutamine to improve the outcome of cardiogenic shock was designed based on the expectation of a large treatment effect, i.e. a 20% lower rate of the primary endpoint in the milrinone group, while the actual difference was 5%.<sup>9</sup>

An interim analysis allows detecting lower event rates and/or a smaller effect size, and their effects may be mitigated by increasing the sample size or prolonging the follow-up.

In some cases, the follow-up period could be too short for the beneficial effects of an experimental treatment to emerge. This may be the case of studies testing primary prevention strategies. Another possible example is the STICH trial on patients with an ejection fraction of 35% or less and coronary artery disease amenable to surgical revascularization; the rates of all-cause death were not significantly different in the surgical revascularization and medical-therapy groups at a median follow-up of 4.7 years, while the rate of all-cause death was significantly lower over a median follow-up of almost 10 years.<sup>10</sup>

A loss of power may also occur because of a high cross-over rate. This may be the case of RCTs testing the efficacy of interventional procedures. An example may be the CABANA trial comparing atrial fibrillation ablation vs. pharmacological control therapy, which counted 37% of crossovers at the time of study termination (37%).<sup>11</sup> Finally, the experimental treatment may be discontinued in a high proportion of patients (e.g. 29% of those on sacubitril/valsartan and 21% on valsartan in LIFE),<sup>12</sup> or many patients can be lost to follow-up (e.g. 13% of patients in the catheter ablation arm of CASTLE-AF).<sup>13</sup>

## Problems in study design: suboptimal endpoint selection

The number of events drives the study power, and reaching an adequate power is easier when the primary endpoint includes multiple elements, some highly prevalent. Nonetheless, a composite endpoint should have clinical relevance and include elements that are presumably affected by the intervention based on prior knowledge. In some trials, the number of events may be 'hypertrophied' by minor events or medical decisions that are quite evenly distributed in the two arms diluting the inter-arm differences. An example is the PROACTIVE trial in patients with type 2 diabetes, including acute coronary syndrome, 'cardiac intervention', leg revascularization, or amputation in the primary endpoint, and resulting neutral, while the composite harder endpoint 'death, MI, or stroke', all spontaneous events, was positive ( $P = .03$ ).<sup>14</sup>

Endpoint definition is sometimes critical. In the OVERTURE trial, the results met pre-specified criteria for non-inferiority of omapatrilat vs. enalapril, but not for superiority (HR 0.94, 95% CI 0.86–1.03,  $P = .187$ ). Using the same definition of HF hospitalization as in the SOLVD trial (i.e. all HF hospitalizations rather than only the admissions requiring 'an intravenous treatment for HF within the first 3 days') led to an increase in the number of events (from 973 vs. 914 to 1041 vs. 941) and the finding of an 11% lower risk in patients on omapatrilat ( $P = .012$ ).<sup>15</sup>

In summary, despite great progress in trial design, several methodological issues can still be encountered when reading superiority trials with neutral results. Recognizing these issues is important to correctly interpret these trials and to decide if research on specific treatment strategies should be pursued.

## Declarations

### Disclosure of Interest

A.P.M. received personal fees, outside the present work, from Bayer, Astra Zeneca and Novartis for participation in clinical studies. The other authors do not report any conflict of interest.

### Data Availability

No new data were generated or analysed in support of this research.

### Funding

All authors declare no funding for this contribution.

## References

- Swedberg K, Young JB, Anand IS, Cheng S, Desai AS, Diaz R, et al. Treatment of anemia with darbepoetin alfa in systolic heart failure. *N Engl J Med* 2013;**368**:1210–1219. <https://doi.org/10.1056/NEJMoa1214865>
- Solomon SD, McMurray JJV, Anand IS, Ge J, Lam CSP, Maggioni AP, et al. Angiotensin-neprilysin inhibition in heart failure with preserved ejection fraction. *N Engl J Med* 2019;**381**:1609–1620. <https://doi.org/10.1056/NEJMoa1908655>
- [https://www.accessdata.fda.gov/drugsatfda\\_docs/nda/2021/207620Orig1s018.pdf](https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/207620Orig1s018.pdf).
- Pfeffer MA, Claggett B, Assmann SF, Boineau R, Anand IS, Clausell N, et al. Regional variation in patients and outcomes in the treatment of preserved cardiac function heart failure with an aldosterone antagonist (TOPCAT) trial. *Circulation* 2015;**131**:34–42. <https://doi.org/10.1161/CIRCULATIONAHA.114.013255>
- de Denu S, O'Meara E, Desai AS, Claggett B, Lewis EF, Leclair G, et al. Spironolactone metabolites in TOPCAT—new insights into regional variation. *N Engl J Med* 2017;**376**:1690–1692. <https://doi.org/10.1056/NEJMc1612601>
- Bhatt DL, Szarek M, Steg PG, Cannon CP, Leiter LA, McGuire DK, et al. Sotagliflozin in patients with diabetes and recent worsening heart failure. *N Engl J Med* 2021;**384**:117–128. <https://doi.org/10.1056/NEJMoa2030183>
- Zile MR, Desai AS, Costanzo MR, Ducharme A, Maisel A, Mehra MR, et al. The GUIDE-HF trial of pulmonary artery pressure monitoring in heart failure: impact of the COVID-19 pandemic. *Eur Heart J* 2022;**43**:2603–2618. <https://doi.org/10.1093/eurheartj/ehac114>
- Puymirat E, Cayla G, Simon T, Steg PG, Montalescot G, Durand-Zaleski I, et al. Multivessel PCI guided by FFR or angiography for myocardial infarction. *N Engl J Med* 2021;**385**:297–308. <https://doi.org/10.1056/NEJMoa2104650>
- Mathew R, Di Santo P, Jung RG, Marbach JA, Hutson J, Simard T, et al. Milrinone as compared with dobutamine in the treatment of cardiogenic shock. *N Engl J Med* 2021;**385**:516–525. <https://doi.org/10.1056/NEJMoa2026845>
- Velazquez EJ, Lee KL, Jones RH, Al-Khalidi HR, Hill JA, Panza JA, et al. Rouleau JL; STICHES investigators. Coronary-artery bypass surgery in patients with ischemic cardiomyopathy. *N Engl J Med* 2016;**374**:1511–1520. <https://doi.org/10.1056/NEJMoa1602001>
- Packer DL, Mark DB, Robb RA, Monahan KH, Bahnsen TD, Poole JE, et al. Effect of catheter ablation vs antiarrhythmic drug therapy on mortality, stroke, bleeding, and cardiac arrest among patients with atrial fibrillation: the CABANA randomized clinical trial. *JAMA* 2019;**321**:1261–1274. <https://doi.org/10.1001/jama.2019.0693>
- Mann DL, Givertz MM, Vader JM, Starling RC, Shah P, McNulty SE, et al. Effect of treatment with sacubitril/valsartan in patients with advanced heart failure and reduced ejection fraction: a randomized clinical trial. *JAMA Cardiol* 2022;**7**:17–25. <https://doi.org/10.1001/jamacardio.2021.4567>
- Marrouche NF, Brachmann J, Andresen D, Siebels J, Boersma L, Jordaens L, et al. Catheter ablation for atrial fibrillation with heart failure. *N Engl J Med* 2018;**378**:417–427. <https://doi.org/10.1056/NEJMoa1707855>
- Dormandy JA, Charbonnel B, Eckland DJ, Erdmann E, Massi-Benedetti M, Moules IK, et al. Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive study (PROspective pioglitAZone clinical trial in macroVascular events): a randomised controlled trial. *Lancet* 2005;**366**:1279–1289. [https://doi.org/10.1016/S0140-6736\(05\)67528-9](https://doi.org/10.1016/S0140-6736(05)67528-9)
- Packer M, Califf RM, Konstam MA, Krum H, McMurray JJ, Rouleau JL, et al. Comparison of omapatrilat and enalapril in patients with chronic heart failure: the omapatrilat versus enalapril randomized trial of utility in reducing events (OVERTURE). *Circulation* 2002;**106**:920–926. <https://doi.org/10.1161/01.cir.0000029801.86489.50>