pagepress

GLOBAL CARDIOLOGY
Global Cardiology 2024; **2**
Published online 28 June 2024 | **DOI**: 10.4081/cardio.2024.33

REVIEW

Using multiple primary endpoints in clinical trials with a focus on heart failure

Stefan D. Anker,[1] Javed Butler,[2] Khawaja M. Talha,[3] Tim Friede[4]

[1]Department of Cardiology (CVK) of German Heart Center Charité; Institute of Health Center for Regenerative Therapies (BCRT), German Centre for Cardiovascular Research (DZHK) partner site Berlin, Charité Universitätsmedizin, Berlin, Germany; [2]Baylor Scott and White Research Institute, Dallas, TX, USA [3]Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA; [4]Department of Medical Statistics, University Medical Center Göttingen; DZHK (German Center for Cardiovascular Research), partner site Göttingen, Germany

Abstract

The use of multiple primary endpoints in cardiovascular clinical trials could be useful addition to the arsenal of comprehensive evaluations of meaningful clinical outcomes. Particularly, it may be advantageous and more economic to use several primary endpoints, if several useful endpoint alternatives exist and when it is uncertain what degree of benefit a certain intervention to be tested can achieve, *i.e.* what power a trial has for a given endpoint. However, analysis of multiple endpoints gives rise to issues of multiplicity of outcomes and family-wise error rate. There are statistical adjustment models (single and multi-step) that modify the level of significance for each endpoint based on the number of endpoints considered overall to control the family-wise error rate. The Bonferroni method is a single step approach that divides the nominal significance level alpha equally across all endpoints but is considered a conservative approach in cases where the number of endpoints is large and where endpoints are correlated. The most used multi-step approaches include the Holm and Hochberg procedures. The Hochberg method is a more efficient, and less conservative approach towards alpha adjustment compared to the Holm procedure. The Bonferroni, Holm and Hochberg test procedures are all considered suitable analysis strategies for multiple primary endpoints with no need to determine a priori the order for the testing to be performed as is needed in all hierarchical test procedures that are most commonly used today. Furthermore, these strategies can also be used to protect the error rate when including secondary endpoints in an extended analytical procedure. The use of any of these methods needs to be specified a priori in the statistical analysis plan to ensure adequate statistical validity. Examples of clinical trials in the heart failure field that have used or are using such multiple primary endpoint approaches are: MIRACLE, ASCEND-HF, EVEREST, FAIR-HF, DELIVER, RESHAPE-HF2 and FAIR-HF2.

Key words: clinical trial, primary endpoints, type 1 error, multiplicity, cardiovascular disease.

Received: 3 June 2024; Accepted: 10 June 2024.
**Correspondence to: Stefan D. Anker, Department of Cardiology, German Heart Center Charité; Institute of Health Center for Regenerative Therapies, German Centre for Cardiovascular Research, partner site Berlin, Charité Universitätsmedizin, Berlin, Germany.*
E-mail: s.anker@cachexia.de

Late-phase clinical trials of drugs and devices are typically designed to assess safety and efficacy considering a range of endpoints, with the primary endpoint denoted as the key measure of therapeutic effectiveness. Clinical trials typically use one primary endpoint, which is either comprised of a single outcome measure (*e.g.,* all-cause death or 6-minute walk test), or is a composite endpoint (*e.g.,* the occurrence of cardiovascular death or hospitalization for heart failure) or a hierarchical composite of multiple different outcomes (*e.g.*, all-cause death, heart failure hospitalizations, and/or changes in symptoms or functional status and other measures). However, using a single primary endpoint has certain drawbacks. First, chronic diseases have a wide spectrum of clinical sequelae that may be important in the context of an individual's health and population-level outcomes, and one single primary endpoint cannot adequately reflect several or all of them. Second, although clinically meaningful, secondary endpoints are often underpowered as they are not used for estimation of sample size. Third, hierarchical composite outcomes account for multiple endpoints to provide an average or weighted estimate of a therapy's effectiveness across a range of outcomes and that may (or may not) increase statistical power, but a significant impact on individual end-

points is undermined if the overall composite outcome is not statistically significant.

When several clinical outcomes are important, typically one is selected as the primary endpoint, and others are termed secondary endpoints. If several clinically meaningful endpoints are available (be that for guideline guiding or regulatory purposes), and if a trial has sufficient power for them, selecting one of these endpoints as primary – and making others for this reason secondary – as akin to gambling. Typically, rationalizing of this approach involves then using arguments that one endpoint is more certain to get regulatory acceptance than others or that the power of the study for one of these endpoints is higher than others, but an element of *guessing* for what will work best often remains. Even if that is *informed* guessing, this exposes whole development programs of innovative drugs or devices to the fate of luck, one may argue, at least in part. This should not be how medical research is progressing.

If the type I error rate is controlled at a certain significance level such as 5% across secondary endpoints (α-*protected* or *error-protected*), then hierarchical ordering of secondary endpoints is typically planned. Of course, this approach requires that α is still available to test the endpoint again if the error protection is supposed to be intact. When the result for the primary endpoint is not significant, all is spent, and further testing cannot be considered error protected. Any hierarchy of endpoints comes with the risk of stopping too early, if one of them fails to be significant. Methods are available to de-risk the analytical approach by moving away from hierarchies without reducing the overall power.[1] Nevertheless, results for secondary endpoints may be nominally significant and may be considered clinically important. The debate is the regularly occurring whether such results can be deemed valid. Recent examples for such situations include the AFFIRM-AHF (Study to Compare Ferric Carboxymaltose with Placebo in Patients with Acute Heart Failure and Iron Deficiency), IRONMAN (Effectiveness of Intravenous Iron Treatment versus Standard Care in Patients with Heart Failure and Iron Deficiency) and EMPACT-MI (Effect of Empagliflozin on Hospitalization for Heart Failure and Mortality in Patients with Acute Myocardial Infarction) trials.[2-4] Meta-analyses, in part, can mitigate the issue of underpowered endpoints, but the issue of α-protection persists.[5]

In such cases, creating multiple primary endpoints should be more often considered as a useful alternative, and may be superior – *i.e.,* more effective in yielding meaningful and valid results – to single primary endpoint approaches.[6] Moreover, there are analytic techniques that control the error rate. Multiple primary endpoints can be designed in different ways.[7-15]

The statistically most simple approach is that multiple primary endpoints are designated as co-primary endpoints whereby all the endpoints need to have evidence for a statistically significant treatment effect to prove efficacy of the therapy.[7] This approach is useful if an effect needs to be demonstrated regarding different dimensions of the disease. For instance, it may be appropriate if one of the primary endpoints confirms a certain mode of action but alone would not lead to therapy approval (*e.g.,* prevent muscle wasting in a chronically ill patient at a severe risk for it), whereas the other primary endpoint provides the evidence for clinical efficacy.

However, using multiple primary endpoints should provide a spectrum of outcomes across which a therapy is effective, and it should further guide investigation into why certain outcomes are more affected than others.[16] If that is considered as the key reason, co-primary outcomes are not the appropriate methodological approach to take.

Analyses of multiple primary endpoints represents a complex issue in trial statistics mainly due to multiplicity of outcomes, increases in the rate of type 1 error, and concerns about the power for each component individually (usually overcome by an increase in sample size).[17] This is not an issue with co-primary endpoints where all endpoints are required to reach statistical significance to establish effectiveness of therapy, precluding the need for α adjustment. However, in cases where statistically significant changes in at least one of several primary endpoints is sufficient to demonstrate effectiveness, controlling for type 1 error becomes crucial. When type 1 error is observed across a group of endpoints, it is referred to as the family-wise error rate, which is the probability of incorrectly rejecting at least one null hypothesis.[16] In general, the higher the number of endpoints/comparisons/observations, the higher the family-wise error rate. For example, in case of 5 independent endpoints being tested at a significance level of α=0.05, the family-wise error rate amounts to 23%.[6] This can be calculated from the following formula: $1-(1-\alpha)^n$. That means for 20 independent endpoints, the family-wise error rate would go up to 64%.

Several further methods of α adjustment have been explored to minimize the effect of multiplicity and control the family-wise error rate – they are categorized as single-step or multi-step approaches. A key advantage is that these methods allow testing for several primary endpoints, but they do not assume a hierarchy (or pre-specified order) of testing for these several endpoints. The most common single-step method is the Bonferroni method which simply involves dividing the significance level (*i.e*. α) by the number of endpoints (n) which is considered for testing.[18] Each outcome is then tested against the adjusted α/n for statistical significance. The Bonferroni method is best used in cases with a small number of independent (*i.e.* uncorrelated) endpoints.[19] Of course, this is unlikely to occur in practice, since endpoints are likely to be correlated. The Bonferroni method is widely used due to its simplicity, although it is a conservative measure to control for family-wise error rate, especially in cases where correlation (positive or negative) between different endpoints exists and the number of endpoints is rather large.

The second approach to addressing multiplicity is using a multi-step approach, the most common being the Holm and the Hochberg procedures.[20,21] Both methods adjust the α in a data dependent manner, testing each subsequent endpoint at a more liberal level of significance, *i.e*. a higher p-value (Holm procedure), or at each subsequent step at a more stringent level of significance, *i.e.* lower p-value (Hochberg procedure). In the following both procedures are considered in more detail.

The Holm procedure is a multi-step procedure where the endpoints are ordered from lowest to highest p-values. The endpoint with the lowest p-value (P1) is chosen and tested against

a pre-specified α level (*e.g.* 0.05) divided by the total number of endpoints (n). If P1 is lower than the adjusted α, the endpoint is considered significant, and the next smallest p-value (P2) is tested. This p-value is evaluated against a less conservative level of significance [α divided by the remaining number of endpoints, α/(n-1)]. If P2 is higher than the new adjusted a level, the null hypothesis is not rejected, and all endpoints thereafter with higher p-values are also rendered not successful. This approach is more efficient than the Bonferroni method; it is less conservative and does not require pre-specified ordering of endpoints based on clinical significance.

An alternative multi-step method has been described by Hochberg, which is considered more efficient than the Holm approach. In the Hochberg procedure, the endpoints are sequenced in the order of decreasing p-values (Figure 1). The highest value is then tested against the original α (usually 0.05). If the p-value is greater than α, the next largest p-value is tested against a more stringent level of significance (α/2). If this p-value also does not surpass the adjusted α, the next p-value is then tested against an even more stringent α (α/3). Once the statistical significance is established at any point in the sequence of decreasing p-value, all endpoints with p-values lower than the one at which statistical significance was established are automatically rendered statistically significant.

Several trials have successfully employed the Hochberg and other procedure for testing multiple endpoints (Table 1). The Hochberg approach was utilized in the MIRACLE trial (Multicenter InSync Randomized Clinical Evaluation) study that evaluated the efficacy of cardiac resynchronisation therapy in patients with heart failure with reduced ejection fraction and intraventricular conduction delay.[8] The study had 3 primary endpoints and used the Hochberg method for α adjustment (although not naming the approach; Figure 2); for all 3 primary endpoints to be statistically significant, their p-values had to be significantly lower than the pre-specified level of significance (α=0.05). For two of the endpoints to be deemed statistically significant, both would require independent p-values to be less than a more stringent level of significance of 0.025 (α/2). For a single endpoint to be successful, it was to surpass a much more rigorous level of significance of 0.0167 (α/3). This approach is less conservative than the Holm approach and prioritizes preservation of α, since once an endpoint is successful, all subsequent endpoints are also rendered significantly significant without being formally tested against its endpoint specific sequential α level. In contrast, in the Holm procedure, the opposite happens when statistical significance is not established, all subsequent endpoints with higher p-values are also considered statistically not significant. The Hochberg procedure, unlike the Bonferroni and Holm procedure, is reliant on the assumption that endpoints are either independent or are positively correlated. However, the latter is a theoretical concern, as any outcome could also be considered from an inverse perspective (*e.g.* event rates become rates of freedom of an event *etc.*), rendering a negative correlation positive by definition.

There are several ongoing cardiovascular clinical trials that are incorporating multiple primary endpoints. The RESHAPE-HF2

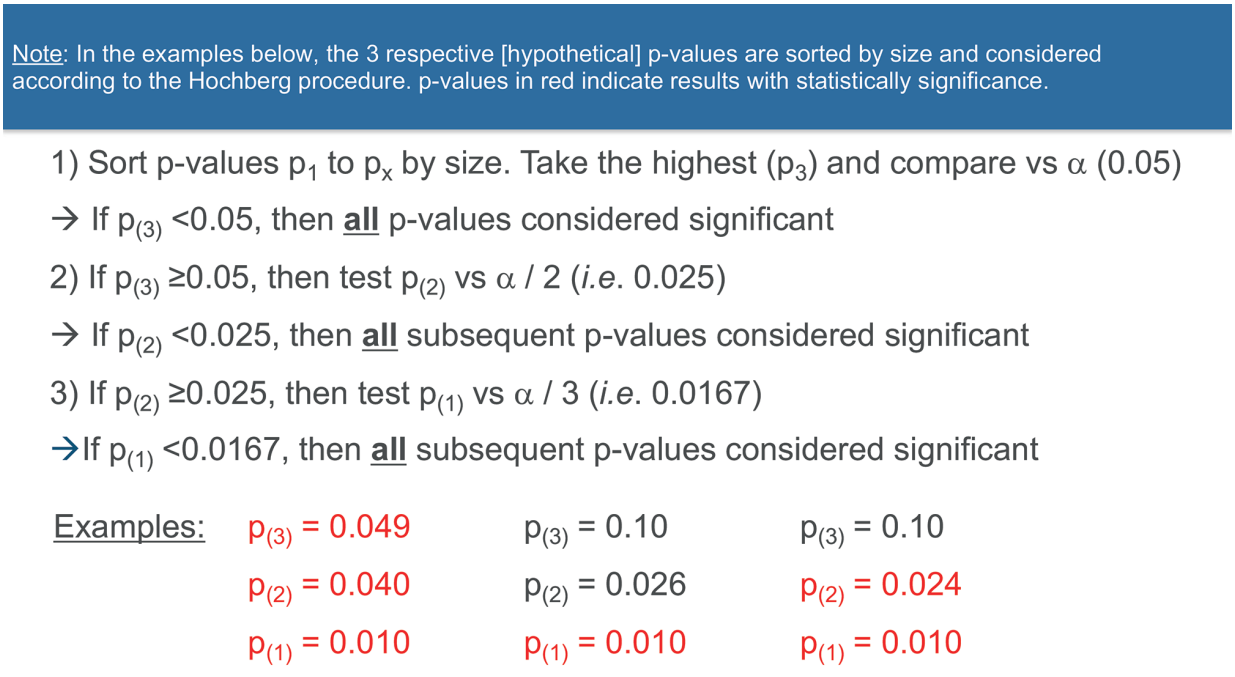

Figure 1. Hochberg procedure for 2 or more (primary) endpoints.

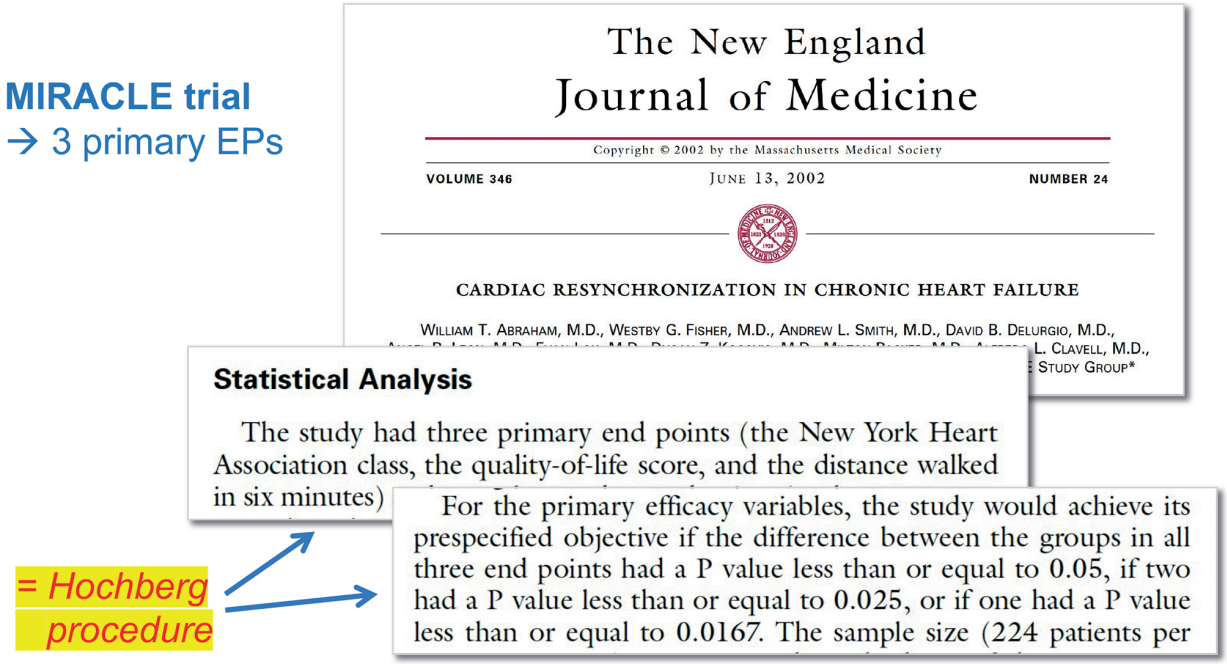

Figure 2. Historical example of using the Hochberg procedure for 3 primary endpoints in a clinical trial in chronic heart failure. Examples are screen shots of Abraham *et al.*, NEJM 2002[8]

(A Randomized Study of the MitraClip Device in Heart Failure Patients with Clinically Significant Functional Mitral Regurgitation trial) is an ongoing randomized controlled trial evaluating the efficacy of mitral trans-catheter edge-to-edge repair in patients with heart failure with reduced ejection fraction and functional mitral regurgitation.[22] The trial uses 3 primary endpoints including 1) composite of total heart failure hospitalizations and cardiovascular death during 24 months of follow-up, 2) total (first and recurrent) HF hospitalizations for 24 months, and 3) change from baseline to 12 months in the Kansas City Cardiomyopathy Questionnaire overall score. These endpoints will be analyzed using the Hochberg procedure for a adjustment to control the family-wise error rate. As an interesting additional feature, the Hochberg procedure will also be applied to the secondary endpoints of RESHAPE-HF2 (Figure 3). All the arguments that can be used for primary endpoints and selecting one over another, can also be used to reconsider the typical approach of pre-selecting a list of hierarchically selected secondary endpoints. Using the Hochberg procedure also there may provide additional benefits that enhance the way we plan and execute our clinical trials in innovative ways.[23]

Lastly, the FAIR-HF2 (Intravenous Iron in Patients with Systolic Heart Failure and Iron Deficiency to Improve Morbidity & Mortality) is the last ongoing trial of intravenous iron in patients with heart failure and a reduced ejection fraction, which will include multiple primary endpoints with Hochberg procedure used for α adjustment (personal communication).[24] The trial is designed to comprehensively assess the role of intravenous iron therapy in heart failure in an attempt to provide a definitive conclusion, given a recent run of equivocal trial results in this space owing to inherent issues with the definitions of iron deficiency and dosing criteria.[25,26] One of the earlier trials of intravenous iron in heart failure, FAIR-HF (Ferinject Assessment in Patients with Iron Deficiency and Chronic Heart Failure), also included dual primary endpoints, 1) self-reported Patient Global Assessment at 24 weeks, and 2) NYHA functional class at 24 weeks and were analyzed using the Hochberg procedure. The trial found that intravenous ferric carboxymaltose had a significantly favorable effect on both primary endpoints with p-value for both endpoints lower than the pre-specified α level of $p=0.05$ ($p<0.001$).[9]

There is a growing emphasis on the inclusion of multiple primary endpoints in cardiovascular clinical trials to allow for a more comprehensive evaluation of clinically meaningful outcomes. It does come with the issue of multiplicity resulting in loss of statistical power and high family-wise error rate. There are statistical adjustment models that modify the level of significance for each endpoint based on the number of endpoints to minimize the family-wise error rate. These adjustments apply primarily to alternative hypothesis and related analyses, and hence need to be planned during the design stage of clinical trials to ensure adequate power.[6,17] The Hochberg method appears to be a more powerful and less conservative approach toward α adjustment compared to the Holm and the Bonferroni procedures and may have greater utility in trials of cardiovascular disease where endpoints are usually independent of each other or are positively correlated.

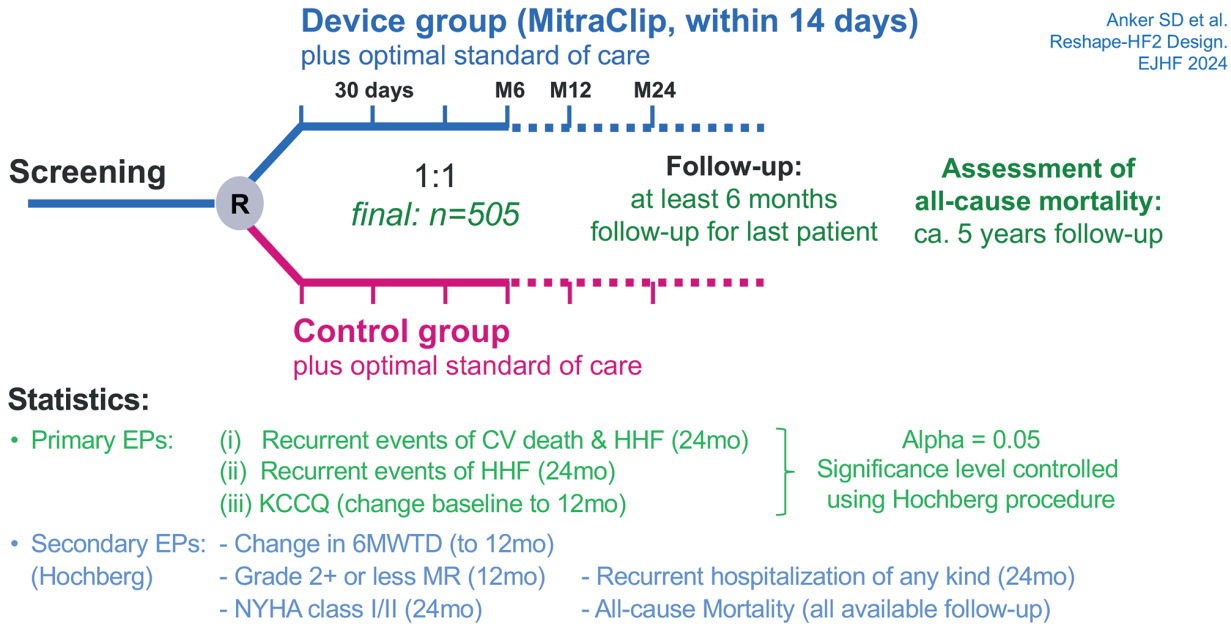

Figure 3. Trial design of RESHAPE-HF2 – testing transcatheter edge-to-edge repair (MitraClip) in the 3rd population, using the Hochberg procedure for primary and secondary endpoints.

Conflict of interest

SDA reports grants and personal fees from Vifor and Abbott Vascular, and personal fees for consultancies, trial committee work and/or lectures from Actimed, Astra Zeneca, Bayer, Bioventrix, Boehringer Ingelheim, Brahms, Cardiac Dimensions, Cardior, Cordio, CVRx, Cytokinetics, Edwards, Farraday Pharmaceuticals, GSK, HeartKinetics, Impulse Dynamics, Medtronic, Novartis, Novo Nordisk, Occlutech, Pfizer, Regeneron, Relaxera, Repairon, Scirent, Sensible Medical, Servier, Vectorious, and V-Wave. Named co-inventor of two patent applications regarding MR-proANP (DE 102007010834 & DE 102007022367), but he does not benefit personally from the related issued patents.

JB reports consulting fees from Abbott, American Regent, Amgen, Applied Therapeutic, AskBio, Astellas, AstraZeneca, Bayer, Boehringer Ingelheim, Boston Scientific, Bristol Myers Squibb, Cardiac Dimension, Cardiocell, Cardior, Cardiorem, CSL Bearing, CVRx, Cytokinetics, Daxor, Edwards, Element Science, Faraday, Foundry, G3P, Innolife, Impulse Dynamics, Imbria, Inventiva, Ionis, Lexicon, Lilly, LivaNova, Janssen, Medtronics, Merck, Occlutech, Owkin, Novartis, Novo Nordisk, Pfizer, Pharmacosmos, Pharmain, Pfize, Prolaio, Regeneron, Renibus, Roche, Salamandra, Sanofi, SC Pharma, Secretome, Sequana, SQ Innovation, Tenex, Tricog, Ultromics, Vifor, and Zoll, and honoraria from Novartis, Boehringer Ingelheim-Lilly, Astra Zeneca, Impulse Dynamics, Vifor.

KMT has nothing to declare.

TF reports payments to institution from Abbott, grants from Deutsche Forschungsgemeinschaft (DFG), Federal Joint Committee (G-BA) and European Commission; consulting fees from Actimed, Bayer, BMS, CSLBehring, Daiichi Sankyo, Galapagos, Immunic, KyowaKirin, LivaNova, Minoryx, Novartis, RECARDIO, Relaxera, Roche, Servier, Viatris, and Vifor, payments from Fresenius Kabi and PINK gegen Brustkrebs, is Trial Data Monitoring Committee member of Aslan, Bayer, BiosenseWebster, Enanta, Galapagos, IQVIA, Novartis, PPD, Recordati, Roche, VICO Therapeutics and is Trial Steering Committee member of SCLBehring.

References

1. Alosh M. Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. Stat Med 2014;33:693-713.
2. Ponikowski P, Kirwan BA, Anker SD, et al. Ferric carboxymaltose for iron deficiency at discharge after acute heart failure: a multicentre, double-blind, randomised, controlled trial. Lancet 2020;396:1895-904.
3. Kalra PR, Cleland JGF, Petrie MC, et al. Intravenous ferric derisomaltose in patients with heart failure and iron deficiency in the UK (IRONMAN): an investigator-initiated, prospective, randomised, open-label, blinded-endpoint trial. Lancet 2022;400:2199-209.
4. Butler J, Jones WS, Udell JA, et al. Empagliflozin after acute myocardial infarction. N Engl J Med 2024;390:1455-66.
5. Siddiqi TJ, Butler J, Coats AJS, et al. SGLT2 inhibitors and risk reduction for mortality in high-risk patients: a meta-analysis of randomized controlled trials. Global Cardiology 2023;1:2.
6. Li G, Taljaard M, Van Den Heuvel ER, et al. An introduction to multiplicity issues in clinical trials: the what, why, when and how. Int J Epidemiol 2017;46:746-56.
7. Crawford J, Prado CMM, Johnston MA, et al. Study design and rationale for the phase 3 clinical development program of enobosarm, a selective androgen receptor modulator, for the prevention and treatment of muscle wasting in cancer patients (POWER Trials). Curr Oncol Rep 2016;18:37.
8. Abraham WT, Fisher WG, Smith AL, et al. Cardiac resynchronization in chronic heart failure. N Engl J Med 2002;346:1845-53.
9. Anker SD, Comin Colet J, Filippatos G, et al. Ferric carboxymaltose in patients with heart failure and iron deficiency. N Engl J Med 2009;361:2436-48.
10. Anker SD, Colet JC, Filippatos G, et al. Rationale and design of Ferinject assessment in patients with IRon deficiency and chronic

Heart Failure (FAIR-HF) study: a randomized, placebo-controlled study of intravenous iron supplementation in patients with and without anaemia. Eur J Heart Fail 2009;11:1084-91.
11. Konstam MA, Gheorghiade M, Burnett JC, et al. Effects of oral tolvaptan in patients hospitalized for worsening heart failure: the EVEREST outcome trial. JAMA 2007;297:1319-31.
12. O'Connor CM, Starling RC, Hernandez AF, et al. Effect of nesiritide in patients with acute decompensated heart failure. N Engl J Med 2011;365:32-43.
13. Solomon SD, McMurray JJV, Claggett B, et al. Dapagliflozin in heart failure with mildly reduced or preserved ejection fraction. N Engl J Med 2022;387:1089-98.
14. Kosiborod MN, Abildstrøm SZ, Borlaug BA, et al. Semaglutide in patients with heart failure with preserved ejection fraction and obesity. N Engl J Med 2023;389:1069-84.
15. Kosiborod MN, Petrie MC, Borlaug BA, et al. Semaglutide in patients with obesity-related heart failure and type 2 diabetes. N Engl J Med 2024;390:1394-407.
16. Chen SY, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. J Thorac Dis 2017;9:1725.
17. FDA. Multiple endpoints in clinical trials guidance for industry. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry (accessed March 23, 2024).
18. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. Stat Med 2009;28:586-604.
19. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310:170.
20. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75:800-2.
21. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65-70.
22. Anker SD, Friede T, von Bardeleben RS, et al. Randomized investigation of the MitraClip device in heart failure: design and rationale of the RESHAPE-HF2 trial design. Eur J Heart Fail 2024;26:984-93.
23. Rosano GMC. Clinical trial design, endpoints and regulatory considerations in heart failure. Global Cardiology 2024;2:18
24. ClinicalTrials.gov. Intravenous iron in patients with systolic heart failure and iron deficiency to improve morbidity & mortality. Available from: https://clinicaltrials.gov/study/NCT03036462 (accessed May 29, 2024).
25. Rosano GMC, Farkas J. Evolving targets for heart failure: the journey so far. Global Cardiology 2023;1:6.
26. Talha KM, Butler J, Haehling S von, et al. Defining iron replete status in patients with heart failure treated with intravenous iron. Global Cardiology 2023;1:17.