Published in final edited form as:

Eur J Heart Fail. 2020 February ; 22(2): 177–180. doi:10.1002/ejhf.1668.

Transforming the interpretation of significance in heart failure trials

Muhammad Shahzeb Khan¹, Simra Irfan², Safi U. Khan³, Mandeep R. Mehra⁴, Muthiah Vaduganathan^{4,*}

¹Department of Medicine, Cook County Health and Hospital System, Chicago, IL, USA;

²Department of Medicine, Dow University of Health Sciences, Karachi, Pakistan;

³Department of Medicine, West Virginia University, Morgantown, WV, USA;

⁴Heart and Vascular Center, Brigham and Women's Hospital, Boston, MA, USA

Introduction

Heart failure (HF) therapeutics are held to a high scientific and regulatory standard. Conclusions and interpretation of randomized clinical trials (RCTs) have been viewed narrowly on whether primary endpoints meet thresholds of significance based on achieving a pre-specified '*P*-value'. Reductionist interpretations of *P*-values at traditional cut-offs have been recently questioned. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group (www.gradeworkinggroup.org) suggested that clinical decision-making should not depend on binary *P*-value thresholds. The American Statistical Association has promoted the continuous interpretation of *P*-values and abandonment of language surrounding 'statistical significance'.^{1,2} The Heart Failure Collaboratory, a multidisciplinary working group hosted a meeting at the US Food and Drug Administration on 25 July 2019 reappraising statistical approaches and reporting in emerging HF clinical trials. Citing concerns surrounding the reproducibility of RCT findings, researchers have even gone to another extreme by proposing a reduction of the generally accepted threshold of statistical significance from a *P*-value of 0.05 to 0.005 and to reclassify values falling between 0.05 and 0.005 as merely 'suggestive'.³ In light of this shifting landscape, we examined the distribution of observed *P*-values across recently conducted HF RCTs and critically discuss issues related to interpreting and communicating uncertainty within trial results in an effort to guide better clinical usefulness and interpretation.

*Corresponding author. Brigham and Women's Hospital Heart & Vascular Center and Harvard Medical School, 75 Francis St., Boston, MA 02115, USA. Tel: +1 857 307-4000, Fax: +1 857 307-1222, mvaduganathan@bwh.harvard.edu.

Conflict of interest: M.R.M. is a consultant for Abbott, Medtronic, Janssen (a division of Johnson and Johnson), Mesoblast, NupulseCV, Inc., FineHeart, Leviticus, Triple Gene, Bayer and Portola. M.V. is supported by the KL2/Catalyst Medical Research Investigator Training award from Harvard Catalyst (NIH/NCATS Award UL1TR002541), serves on advisory boards for Amgen, AstraZeneca, Bayer AG, Baxter Healthcare, and Boehringer Ingelheim, and participates on clinical endpoint committees for studies sponsored by Novartis and the NIH. All other authors have no conflicts to declare.

Distribution of P -values in contemporary heart failure trials

Two independent investigators (S.I., S.U.K.) searched PubMed/MEDLINE from 1 January 2013 to 31 December 2017 for phase 3 or 4 HF RCTs using the following limits: publication year, 'heart failure', 'trial*', and 'randomized'. An arbitrary range of 5 years was considered appropriate to cover a large number of contemporary HF trials for the analysis. Only trials published in the *New England Journal of Medicine*; *Lancet*; *Journal of the American Medical Association*; *Journal of the American College of Cardiology*; *JACC: Heart Failure*; *Circulation*; *Circulation: Heart Failure*; *European Heart Journal*; and *European Journal of Heart Failure* were considered. These journals were selected based on their high impact factors, broad readership and reputation of publishing important HF trials. We excluded trials with <100 participants ($n = 77$), non-inferiority or Bayesian analyses ($n = 5$), pooled analyses ($n = 4$), or trials which did not list P -values for the primary endpoint ($n = 3$). We calculated the proportion of trial endpoints that would be considered statistically significant with a threshold of $P < 0.005$ and the proportion that would be re-categorized as suggestive ($P < 0.05$ but ≥ 0.005). We also calculated the fragility index (number of events that would need to occur in the comparator arm to change directionality of significance) for endpoints having $P < 0.05$.⁴ The fragility index is an easy to calculate measure that provides clinicians with an alternative, more intuitive way to understand the robustness of trial results in addition to traditionally used metrics.

Of 161 identified, 72 RCTs with 90 primary endpoint comparisons met inclusion criteria. Of these 90 endpoints, 43 (48%) were clinical endpoints (such as mortality or hospitalization). A $P < 0.05$ was reported for 37 endpoints. Of these 37 endpoints, 20 were significant at $P < 0.005$, while 17 would be reclassified as suggestive (Figure 1). Excluding continuous primary endpoints, there were 15 binary endpoints with $P < 0.05$ and sufficient reported data to calculate the fragility index. Median fragility index of durably significant trials was 23 (range 5–118), while that of suggestive trials was one (range 0–6). Key characteristics of trials that were reclassified vs. durably significant are presented (Table 1). Forty-seven trials reported P -values for all-cause mortality of which nine had $P < 0.05$ (three were durably significant at $P < 0.005$ while six would be reclassified as suggestive). Similarly, 18 trials reported P -values for HF hospitalizations of which only one had a $P < 0.005$.

Interpreting P -values in contemporary heart failure trials

In our empirical analysis, we observed a broad range of P -values for primary endpoints included in published reports of contemporary HF clinical trials. How should clinicians, regulatory bodies, guideline committees, and other key decision-makers interpret these measures of uncertainty? We explore a few proposals and address limitations regarding the current framework of interpretation of P -values.

Is lowering the P -value threshold the solution?

Based on our assessment, nearly half of HF clinical trial primary endpoints that were previously significant at $P < 0.05$ would be reclassified to suggestive if the statistical significance threshold were lowered from 0.05 to < 0.005 . Such reclassification also tracks

closely with the fragility index, as expected, indicating that suggestive outcomes are associated with greater fragility and less robustness. On average, among suggestive trials, a single additional event in the comparator group would have changed a result to non-significant. Even among articles published in major general medical journals with the highest impact factors, a significant proportion of primary findings would be reclassified at lower P -value boundaries.⁵

Lowering the significance threshold to $P < 0.005$ may guard against false-positive results and reassure regulators, payers, and clinicians on the confidence of a scientific finding. However, a smaller alpha threshold may result in potentially effective treatments being dismissed (higher false-negative rates), especially in populations with high phenotypic heterogeneity or where the disease is difficult to diagnose, such as HF with preserved ejection fraction. It will also result in an adverse consequence of lower feasibility for certain interventions as higher sample sizes with greater trial costs will be required. Investment in the cardiovascular RCT enterprise has remained stagnant,⁶ and stricter thresholds for significance may discourage trialists and sponsors. This may be particularly challenging for breakthrough therapeutics, including devices, and investigation of rare disease, where there are usually far smaller patient pools from which to recruit patients.⁷ Indeed, a lower proportion of durably significant studies were trials examining devices (compared with RCTs testing pharmacotherapies or other interventions).

It is important to note that changing the statistical significance threshold does not solve the fundamental problems of P -values viewed in a conventionally dichotomous way to refute or support a hypothesis. Even if the statistical significance threshold is lowered, P -values cannot be equated to the treatment effect being valid and data quality, integrity, and clinical meaningfulness will still need to be examined independently.

Are two trials better than one?

The US Food and Drug Administration has traditionally required two adequately powered and well-conducted RCTs that meet statistical significance at $P < 0.05$ for regulatory approval. Contemporary American College of Cardiology/American Heart Association guidelines provide an evidence hierarchy such that claims supported by more than one RCT, meta-analysis, or an RCT and a high-quality registry study are given a level of evidence A designation, while claims supported by a single RCT are given a level of evidence B-R designation.⁸ Replication of findings in a separate set of trial conditions may have theoretical benefits to improve generalizability and confidence in study findings.⁹ Trials conducted independently may be more likely to capture diverse and representative cohorts. However, in select circumstances, single, well-designed and well-executed, large trials with effect estimates highly inconsistent with the null hypothesis have been accepted as sufficient regulatory evidence. A single trial based on stringent thresholds that provides adequate regulatory confidence may reduce overall burden and cost of trial conduct and even accelerate approval of drugs or devices. A single trial may also be adequately sized to identify less common safety signals that would have been missed in individual smaller trials.

Conducting a second trial after a convincingly ‘positive’ initial trial with clinically meaningful endpoints may also be ethically challenging.

Sacubitril/valsartan was approved for use in HF with reduced ejection fraction based on a single large trial without replication as its composite primary endpoint was significant at $P=4.0 \times 10^{-7}$.¹⁰

Given its large sample size and as it was overpowered, the pivotal trial could have been split into multiple smaller trials and still individually shown a benefit on important endpoints.¹¹ As long as conducted with the same scientific rigor with valid comparators, we believe that data collected in the context of one large RCT conveys similar statistical confidence and thus should be interpreted under the same statistical framework as two or more smaller trials. Quality, rather than solely quantity, of evidence must be used to guide interpretation.¹¹

A path forward?

Pre-specified statistical rules

At the present juncture, statistical significance testing and formal declaration of an alpha threshold may still be relevant in select settings, such as for regulatory decision-making. A preferred approach may involve researchers justifying an alpha level when designing the RCT based on whether controlling for type 1 or type 2 error will be more consequential in the framework of the particular research objective. It is likely that different significance thresholds may be suitable for different circumstances. For instance, a lower statistical bar may be appropriate for rare (orphan) conditions or disease states with high mortality and limited therapeutic options, such as cardiac amyloidosis, in which the consequences of type 2 error are great (dismissing potentially important treatments). Conversely, more stringent alpha thresholds should be considered when testing a multiplicity of endpoints to adequately limit type 1 error. Tailored pre-specified statistical plans that justify the choice of alpha may facilitate more appropriate statistical interpretation. Similarly, in trials employing alternative emerging methods (such as Bayesian approaches), structured and pre-specified analytic plans will allow for appropriate interpretation.

Clinically meaningful effects

Clinicians and other decision-makers should consider the magnitude of the effect size and clinical meaningfulness of findings. A robust P -value in a trial constructed with a composite endpoint that includes less meaningful components (such as biomarkers) may not be equivalent to a trial with a larger P -value in a trial evaluating a more robust endpoint (such as survival). Decision-makers should also consider whether trial findings are generalizable to local populations of interest. Additionally, to make well-informed and balanced decisions in interpreting a trial’s findings, clinicians should also integrate findings from previous studies and examine the totality of evidence from secondary endpoints, key subgroups, and sensitivity analyses.

Communicating uncertainty

Based on the wide distribution of P -values reported across HF clinical trials of *published* reports, we anticipate an even broader spread of larger P -values if unpublished RCTs were to be considered. P -values should be reported in a continuous fashion, an approach that may be less prone to publication bias or spin. Reporting continuous P -values (e.g. ' $P = 0.051$ ') will avoid unnecessary dichotomization (e.g. 'not statistically significant at an alpha threshold of 0.05'). Additionally, as supported by new statistical reporting guidelines in the *New England Journal of Medicine*¹² and an Editors' statement in the *Journal of the American Medical Association*,¹³ all P -values should be accompanied by effect estimates and margins of error.

Trial A with a P -value of 0.049 and trial B with a P -value of 0.051 may be viewed by some from a binary lens despite the narrow difference in probability estimates. The global phase 3 PARAGON-HF (Prospective Comparison of ARNI with ARB Global Outcomes in HF with Preserved Ejection Fraction) trial examining sacubitril/valsartan in HF with preserved ejection fraction narrowly missed statistical significance for its primary endpoint at an alpha threshold of 0.05.¹⁴ As the HF clinical trial community interprets aggregate data results from this trial and other important RCTs due to report in the next year, we are hopeful that individual P -values are appropriately contextualized and not interpreted narrowly. Moving forward, flexible but pre-defined statistical analysis plans, robust study designs targeting clinically meaningful endpoints, and transparent reporting of results remain essential in communicating scientific findings from HF RCTs. As stated in a special issue of *The American Statistician*¹ on the P -value: 'As we venture down this path, we will begin to see fewer false alarms, fewer overlooked discoveries, and the development of more customized statistical strategies'.

References

1. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond " $p < 0.05$ ". *The American Statistician* 2019;73:1–19.
2. Wasserstein RL, Lazar NA. The ASA statement on p -values: context, process, and purpose. *The American Statistician* 2016;70:129–133.
3. Ioannidis JP. The proposal to lower P value thresholds to .005. *JAMA* 2018;319:1429–1430. [PubMed: 29566133]
4. Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, Molnar AO, Dattani ND, Burke A, Guyatt G, Thabane L, Walter SD, Pogue J, Devereaux PJ. The statistical significance of randomized controlled trial results is frequently fragile: a case for a fragility index. *J Clin Epidemiol* 2014;67: 622–628. [PubMed: 24508144]
5. Wayant C, Scott J, Vassar M. Evaluation of lowering the P value threshold for statistical significance from .05 to .005 in previously published randomized clinical trials in major medical journals. *JAMA* 2018;320:1813–1815. [PubMed: 30398593]
6. Fordyce CB, Roe MT, Ahmad T, Libby P, Borer JS, Hiatt WR, Bristow MR, Packer M, Wasserman SM, Braunstein N, Pitt B, DeMets DL, Cooper-Arnold K, Armstrong PW, Berkowitz SD, Scott R, Prats J, Galis ZS, Stockbridge N, Peterson ED, Califf RM. Cardiovascular drug development: is it dead or just hibernating? *J Am Coll Cardiol* 2015;65:1567–1582. [PubMed: 25881939]
7. Vaduganathan M, Tahhan AS, Greene SJ, Kelkar AA, Georgiopoulou VV, Kalogeropoulos AP, Fonarow GC, Gheorghiade M, Butler J. Contemporary cardiovascular device clinical trials (trends and patterns 2001 to 2012). *Am J Cardiol* 2015;116:307–312. [PubMed: 25963222]
8. Halperin JL, Levine GN, Al-Khatib SM, Birtcher KK, Bozkurt B, Brindis RG, Cigarroa JE, Curtis LH, Fleisher LA, Gentile F, Gidding S, Hlatky MA, Ikonidis J, Joglar J, Pressler SJ,

- Wijeysundera DN. Further evolution of the ACC/AHA clinical practice guideline recommendation classification system: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* 2016;133:1426–1428. [PubMed: 26399660]
9. Haslam A, Prasad V. Confirmatory trials for drugs approved on a single trial. *Circ Cardiovasc Qual Outcomes* 2019;12:e005494. [PubMed: 31137954]
10. McMurray JJ, Packer M, Desai AS, Gong J, Lefkowitz MP, Rizkala AR, Rouleau JL, Shi VC, Solomon SD, Swedberg K, Zile MR; PARADIGM-HF Investigators and Committees. Angiotensin–neprilysin inhibition versus enalapril in heart failure. *N Engl J Med* 2014;371:993–1004. [PubMed: 25176015]
11. Packer M Unbelievable folly of clinical trials in heart failure: the inconvenient truth about how investigators and guidelines weigh evidence. *Circ Heart Fail* 2016;9:e002837. [PubMed: 27009552]
12. Harrington D, D’Agostino RB, Gatsonis C, Hogan JW, Hunter DJ, Normand ST, Drazen JM, Hamel MB. New guidelines for statistical reporting in the *Journal*. *N Engl J Med* 2019;381:285–286. [PubMed: 31314974]
13. Bauchner H, Golub RM, Fontanarosa PB. Reporting and interpretation of randomized clinical trials. *JAMA* 2019;322:732–735. [PubMed: 31454026]
14. Solomon SD, JJ MM, Anand IS, Ge J, Lam CS, Maggioni AP, Martinez F, Packer M, Pfeffer MA, Pieske B, Redfield MM, Rouleau JL, van Veldhuisen DJ, Zannad F, Zile MR, Desai AS, Claggett B, Jhund PS, Boytsov SA, Comin-Colet J, Cle-land J, Düngen HD, Goncalvesova E, Katova T, Kerr Saraiva JF, Lelonek M, Merkely B, Senni M, Shah SJ, Zhou J, Rizkala AR, Gong J, Shi VC, Lefkowitz MP; PARAGON-HF Investigators and Committees. Angiotensin–neprilysin inhibition in heart failure with preserved ejection fraction. *N Engl J Med* 2019;381:1609–1620. [PubMed: 31475794]

Figure 1.

Distribution of P -values of the primary endpoints identified in selected recent heart failure randomized clinical trials. Trials that would be durably significant ($P < 0.005$) are highlighted in red, while trials that would be reclassified if the P -value threshold is lowered to <0.005 ($P < 0.05$ but $P \geq 0.005$) are highlighted in blue.

Table 1

Characteristics of 72 identified heart failure clinical trials

Trial characteristics	Durably significant ($P < 0.005$) [$n = 15$]	Suggestive ($P = 0.005$ to $P < 0.05$) [$n = 15$]	Non-significant ($P \geq 0.05$) [$n = 42$]
Age (years) ^a	65±6	67±5	65±5
Women, n (%)	303 (41)	133 (32)	244 (30)
Diabetes mellitus, n (%)	345 (26)	153 (39)	280 (37)
Hypertension, n (%)	834 (63)	268 (63)	493 (65)
NYHA class III or IV, n (%)	524 (52)	165 (55)	461 (50)
Heart failure with reduced ejection fraction, n (%)	8 (53)	13 (87)	30 (71)
Chronic heart failure, n (%)	13 (87)	11 (73)	32 (76)
Intervention, n (%)			
Drug	10 (67)	4 (27)	24 (58)
Device	1 (6)	5 (33)	9 (21)
Other	4 (27)	6 (40)	9 (21)
Sample size in control arm, median (IQR)	152 (67–506)	106 (77–301)	187 (78–497)
Sample size in intervention arm, median (IQR)	152 (71–506)	110 (79–299)	199 (110–505)
Funding source, n (%)			
Industry	6 (40)	5 (33)	20 (47)
Public	2 (13)	4 (27)	12 (29)
Mixed	4 (27)	4 (27)	8 (19)
Other/no support/not reported	3 (20)	2 (13)	2 (5)
Multicentre, n (%)	13 (87)	8 (53)	37 (88)
Multinational, n (%)	9 (60)	4 (27)	20 (48)
Mortality-inclusive primary endpoint, n (%)	4 (27)	6 (40)	25 (60)
Clinical primary endpoint ^b , n (%)	6 (40)	8 (53)	26 (62)

IQR, interquartile range; NYHA, New York Heart Association.

^aMean (standard deviation) for age was calculated from aggregate data.^bDefined as major adverse cardiovascular events or death or hospitalization or any composites involving these endpoints.