Received: 22 April 2020 | Revised: 18 August 2020 | Accepted: 25 August 2020

DOI: 10.1002/pst.2070

VIEWPOINT

WILEY

The promise and pitfalls of composite endpoints in sepsis and COVID-19 clinical trials

P. M. Brown[1,2] | Tormod Rogne[1,3] | Erik Solligård[1,3]

[1]Gemini Center for Sepsis Research, Clinic of Anesthesia and Intensive Care, St. Olav's hospital, Trondheim University Hospital, Trondheim, Norway

[2]Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

[3]Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence
P. M. Brown, Gemini Center for Sepsis Research, Clinic of Anesthesia and Intensive Care, St. Olavs Hospital, AHL-senteret, 6 etg. Prinsesse Kristinas gate 3, 7030 Trondheim, Norway.
Email: paul.brown@ntnu.no

Summary

Composite endpoints reveal the tendency for statistical convention to arise locally within subfields. Composites are familiar in cardiovascular trials, yet almost unknown in sepsis. However, the VITAMINS trial in patients with septic shock adopted a composite of mortality and vasopressor-free days, and an ordinal scale describing patient status rapidly became standard in COVID studies. Aware that recent use could incite interest in such endpoints, we are motivated to flag their potential value and pitfalls for sepsis research and COVID studies.

Composite endpoints that concoct a univariate score from multiple outcomes appear in regulatory guidelines, reflecting their rising use in research. The ITT principle and clinical considerations are often motivations, see, for example, the WHO's eight-point ordinal scale for COVID studies.[1] Composites intend to enhance statistical power and design phase II trials that better anticipate phase III results by combining clinical outcomes (low event rates) with, for example, biomarkers, intermediate events and/or patient-reported outcomes. An additional claim has been made, therefore, that the resulting endpoint is a concise measure of the "overall" impact of disease[2] and more relevant for patients.[3]

Composite endpoints reveal the tendency for convention to arise locally (eg, within industry/academia, geographical region, disease area) as preference for one approach over another is inculcated through imitation and repetition. Composites were described decades ago in HIV research[4] and are familiar today in cardiovascular trials, yet almost unknown in sepsis. The VITAMINS[5] trial in patients with septic shock adopted a composite of mortality and vasopressor-free days, but it was not referred to as such and it was not derived thoroughly as an unmatched win-ratio.[6] Aware that recent trials of vitamin C could incite interest in such endpoints (composites were raised at the first Critical Care Clinical Trials Workshop in 2019[7]), we are motivated to flag their potential value and pitfalls for sepsis research and COVID studies specifically.

We will use CITRIS-ALI[8] for illustration. This randomized trial is a good candidate for a global rank composite because three co-primary outcomes were nominated: Sequential Organ Failure Assessment (mSOFA), C-reactive protein (CRP) and thrombomodulin (TM). To control the overall type I error-rate, the decision rule required that all three P values be <.05 and the smallest <.02 and the next smallest <.03. Statistical significance was not achieved. However, under these circumstances, a composite obviates the need to apportion the significance level into unattainable pieces. In this way, more than any other, it may be considered to imply superior power.

The global rank[9,10] seeks to order patients from the most adverse response to the most favourable by arranging variables in a hierarchy that prioritizes definitive outcomes: the patient with the shortest survival time receives rank 1; patients with censored survival are ranked on the next outcome. We will order outcomes as follows: mortality-mSOFA-CRP-TM. A threshold for "failure" must be specified to determine when to move to the next outcome, for example, for mortality, a 28-day cut-off is used. If the death rate is low it implies that the composite makes greater use of subsequent data. Note that the global rank easily handles missing data and competing risks and a composite of four outcomes is not unusual; a survey of trials, reporting in 2008, found a median of 3 and a range of 2 to 9.[11] (Full detail is in our SAS code: https://gitlab.com/pmbrown/citris-ali.)

For power estimation, we generated 1000 random samples of correlated Normal variates before transforming them to the relevant scale, for example, exponential survival times for mortality. Iteration achieved the desired correlations. Moderate effect sizes were assumed based on VITAMINS (mortality) and CITRIS-ALI. For each random sample, we derived the global rank and ran the Wilcoxon rank-sum test to compare our nominal groups "Vitamin C" and "Placebo." Power is then estimated as the proportion of samples yielding a P value <.05. We consider two scenarios for the CITRIS-ALI decision rule: one assuming complete data, and the second assuming (conservatively) data are missing if the patient dies before day 28.

Figure 1 shows that the increase in power, observed for the composite, seems immediately compelling; however, it is uncertain and should be interpreted with caution. For example, the required coding is intricate and, because assumptions are made across multiple outcomes, the range of plausibility for power is enlarged. Moreover, power is sensitive to the weighting of component outcomes, that is, the representation of outcomes within the composite, which is affected by elicited opinion (required for construction), and also variable and uncertain.[12] Thus, the claim that composites retain power while prioritizing clinical outcomes is a sleight of hand: our simulations showed 60% of patients ranked on TM,

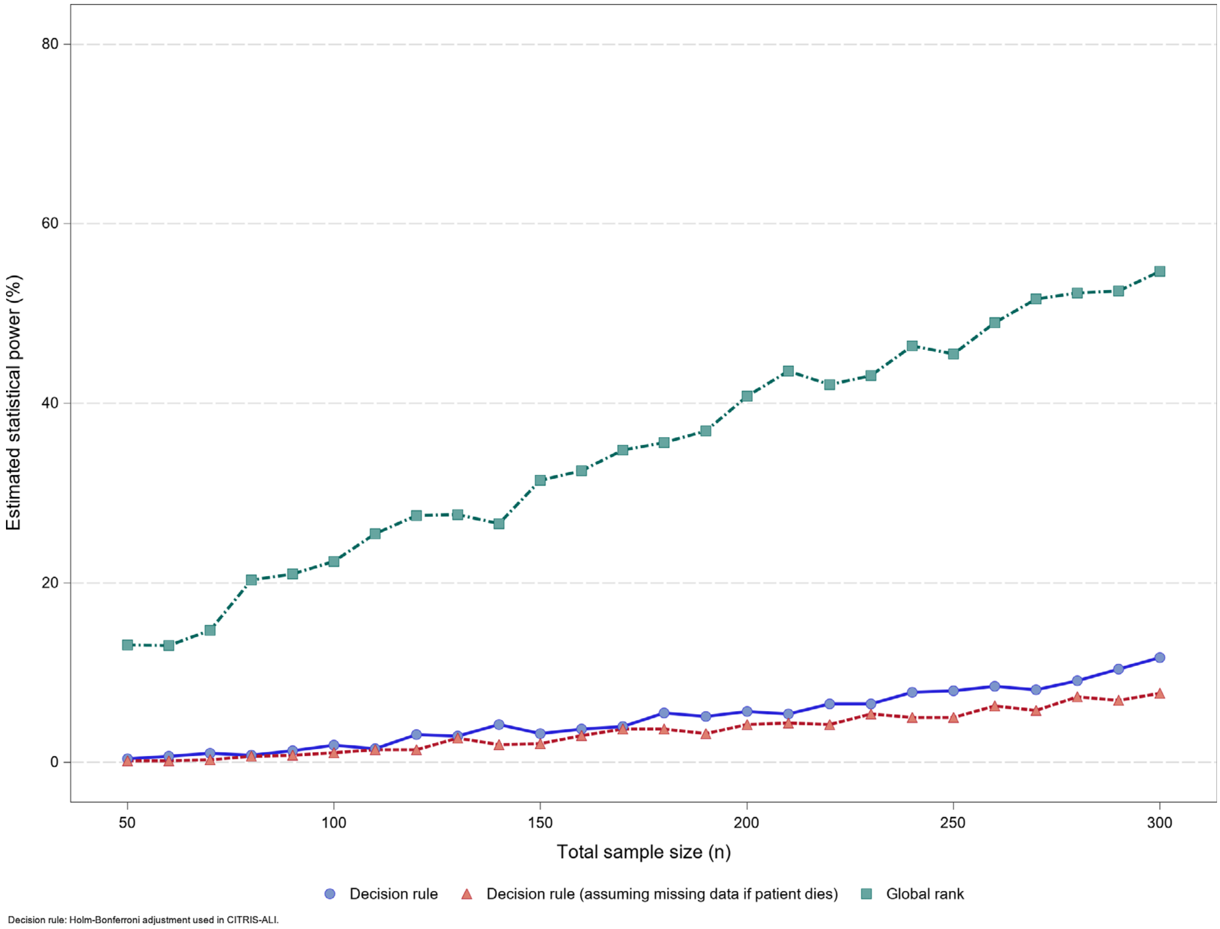

FIGURE 1 Estimated power by assumed total sample size

25% on mortality. If we remedy this by pre-specifying weights there is likely an inverse relationship between the relative importance weight and incidence, and power is low. In some cases, the trialist will plan an interim assessment of power at which time the composite may be expanded if found lacking.[13] Although with additional outcomes we increase the risk of counteracting effects, which further depletes power. Hence a composite makes sense only if heterogeneity (direction) of the effect can be ruled out a priori. In any case, convention dictates that the component outcomes are analysed and presented separately,[14] and these analyses will lack power, possibly discouraging further research.

We might then ponder the default popularity of composites and notice it feels ironic: (a) While "reproducibility" is the rallying cry, bespoke composites infused with subjective weights are offered by proponents. (b) As researchers lament the existence of *P* values, composites inadvertently shift emphasis from estimation onto significance testing (because the resulting scale is unfamiliar). (c) Amid a common move towards "translational medicine", popular composites can be difficult to translate (eg, if effects are not consistent across components). (d) When Pocock suggested the time-to-first composite in 1997,[15] he described it as "simple," "additional" and noted "interpretation can be difficult". This composite has since become a standard primary outcome in cardiovascular trials. (e) The concluding remarks of a literature review 25 years ago describe "serious deficiencies" that remain relevant: "many authors develop ad hoc arbitrarily constructed [composites] for immediate use (often as a primary outcome measure) … without evaluation of their measurement properties." This is "scientifically debatable."[16] (f) The draft EMA guideline on acute heart failure (AHF)[17] in 2012 specifically discouraged the use of trichotomous composites while AHF trials were appearing in NEJM with this endpoint as the primary outcome.[18] Current EMA recommendations against adopting net-benefit composites[19] remain unheeded outside industry.[20] (g) Composites reduce multivariate data to univariate while MRC guidance on complex interventions states: "a single primary outcome may not make best use of the data"[21] and the designation is incongruous in a Bayesian scheme.[22] With enhancements in computation and software, crude tools ought to be superseded by more sophisticated alternatives.

Research on composites is often prefaced by concern for statistical power and clinical relevance despite faring poorly against more informative joint modelling alternatives.[23-26] A few researchers can be found explicitly rejecting composites in favour of GEE modelling for their primary analysis.[27,28] It is unusual in a research paper to explain why a certain statistical method was used; it is even more unusual to find authors explaining why a certain method was *not* used. It betrays a conscious departure from "the norm" that statisticians negotiate in service to clients when statistical sophistication and an appeal to "clinical relevance" or convention do not coincide. As a collective, statisticians have appeared to waver in their willingness to appease[29,30] before settling on an unfortunate middle-ground where "assisting" now entails embellishing after-the-fact; for example, hierarchical composites incorporating "optimal weights that maximise power."[2]

Fortified by such literature, composites slowly become embedded in clinical research, driving go/no go decisions, despite unreliable power estimation and the ability to excite debate over trial results. For example, the first placebo-controlled remdesivir clinical trial for COVID that adopted an ad hoc ordinal scale (describing patient status) with obvious statistical shortcomings.[31] Disagreement over the results and choice of endpoint ensued.[32] The third remdesivir trial (using a similar ordinal scale) randomised patients to a 10 or 5-day course of remdesivir or standard care (1:1:1).[33] Statistical significance was declared for 5-day remdesivir vs standard care although the difference was of "uncertain clinical importance." An estimate was not even presented for 10-day remdesivir when the proportional odds assumption was found to be unrealistic. One commentary on the remdesivir trials noted the difficulty in translating the effect estimate from this scale "into a clinically meaningful statement for patients, clinicians, and policy makers."[34] (It should also be noted that power estimation is complicated by a lack of prior data for the outcome.) Desai and Gyawali surveyed the use of endpoints in COVID studies and found 43% used a similar ordinal scale, and only 6% used mortality. They expressed concern about equivocal results: "When lives are at stake, the trials should literally measure if lives can be saved."[35]

Given the rapid and uncritical propagation of composites in other disease areas, we are, therefore, motivated to promote caution. The EMA guideline on AHF has become quite agreeable, reflecting growing mainstream acceptance of composites in that field. The first version (2005) of the document stated matter-of-factly: "Composite endpoints at present are not acceptable unless well validated." Although the guideline still (2015) says "diverse and disparate measures would not be acceptable as components of a composite," based on the use of such endpoints in AHF there is now a push for regulatory agencies to accept them in sepsis trials. In particular, the trichotomous composite referred to above is promoted in sepsis[3] after criticism was removed from the 2012 draft guideline. Unsurprisingly, this composite has been shown to leak power.[36,37] To dissuade the use of ad hoc composites, recommendations could give more attention to the reporting and interpretation of results (eg, the ICH E9 R (1) addendum on estimands[38]). If the problems of composites are not given emphasis then only the curious analyst perceives just how wasteful and obscure composites are when whittling multivariate data down to a univariate catchall.

In phase II trials, adjusting alpha for multiple outcomes is undesirable and likely to diminish the chances of a successful trial. The issue of underpowered trials in sepsis has been flagged.[39] Hence there is an impulse to adopt a composite supported by precarious power estimates. In general, however, although data simulations are useful for evaluating the properties of composites, we should not be won over by contests that pit methods head-to-head in search of statistical power. First and foremost, the endpoint must remain cogent. There is a risk that future trials in sepsis and COVID will be designed around newly formed, ambivalent measures supported by fashion rather than science.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Gitlab at https://gitlab.com/pmbrown/citris-ali.

ORCID

P. M. Brown https://orcid.org/0000-0002-2411-3312

REFERENCES

1. WHO R&D blueprint novel coronavirus (COVID-19) therapeutic trial synopsis. World Health Organization 2020 Web site. https://www.who.int/blueprint/priority-diseases/key-action/COVID-19_Treatment_Trial_Design_Master_Protocol_synopsis_Final_18022020.pdf Published February 18, 2020. Accessed August 11, 2020.
2. Matsouaka RA, Singhal AB, Betensky RA. Optimal weighted Wilcoxon–Mann–Whitney Test for prioritized outcomes. *New Frontiers of Biostatistics and Bioinformatics*. Switzerland: Springer; 2018:3-40.
3. Mebazaa A, Laterre PF, Russell JA, et al. Designing phase 3 sepsis trials: application of learned experiences from critical care trials in acute heart failure. *J Intensive Care*. 2016;4:24-24.
4. Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. *Stat Med*. 1999;18:1341-1354.
5. Fujii T, Luethi N, Young PJ, et al. Effect of vitamin C, hydrocortisone, and thiamine vs hydrocortisone alone on time alive and free of vasopressor support among patients with septic shock: the VITAMINS randomized clinical trial. *JAMA*. 2020;323(5):423-431.
6. Novack V, Beitler JR, Yitshak-Sade M, et al. Alive and ventilator free: a hierarchical, composite outcome for clinical trials in the acute respiratory distress syndrome. *Crit Care Med*. 2020;48(2):158-166.
7. Harhay MO, Casey JD, Clement M, et al. Contemporary strategies to improve clinical trial design for critical care research: insights from the first critical care clinical Trialists workshop. *Intensive Care Med*. 2020;46(5):930-942.
8. Fowler AA 3rd, Truwit JD, Hite RD, et al. Effect of vitamin C infusion on organ failure and biomarkers of inflammation and vascular injury in patients with sepsis and severe acute respiratory failure: the CITRIS-ALI randomized clinical trial. *JAMA*. 2019;322(13):1261-1270.
9. Margulies KB, Anstrom KJ, Hernandez AF, et al. GLP-1 agonist therapy for advanced heart failure with reduced ejection fraction: design and rationale for the functional impact of GLP-1 for heart failure treatment study. *Circ Heart Fail*. 2014;7(4):673-679.
10. Felker GM, Maisel AS. A global rank end point for clinical trials in acute heart failure. *Circ Heart Fail*. 2010;3(5):643-646.
11. Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BMJ*. 2010;341:c3920.
12. Brown PM, Ezekowitz JA. Examining the influence of component outcomes on the composite at the design stage. *Circ Cardiovasc Qual Outcomes*. 2018;11(6):e004419.
13. Pocock SJ, Clayton TC, Stone GW. Challenging issues in clinical trial design: part 4 of a 4-part series on statistics for clinical trials. *J Am Coll Cardiol*. 2015;66(25):2886-2898.
14. O'Brien PC, Dyck PJ, Tilley BC. Composite endpoints in clinical trials. *Methods and Applications of Statistics in Clinical Trials*. Chichester: Wiley; 2014:246-251.
15. Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Control Clin Trials*. 1997;18(6):530-545.
16. Coste J, Fermanian J, Venot A. Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals. *Stat Med*. 1995;14(4):331-345.
17. Guideline on clinical investigation of medicinal products for the treatment of acute heart failure. Committee for Medicinal Products for Human Use (CHMP). CPMP/EWP/2986/03 Rev. 1 Web site. https://www.ema.europa.eu/en/clinical-investigation-medicinal-products-treatment-acute-heart-failure. Accessed February 25, 2020.
18. Massie BM, O'Connor CM, Metra M, et al. Rolofylline, an adenosine A1−receptor antagonist, in acute heart failure. *N Engl J Med*. 2010;363(15):1419-1428.
19. Guideline on multiplicity issues in clinical trials. European Medicines Agency. Committee for Medicinal Products for Human Use (CHM). EMA/CHMP/44762/2017 Web site. https://www.ema.europa.eu/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf. Published December 15, 2016. Accessed February 25, 2020.

20. Brown PM, Ezekowitz JA. Letter by Brown and Ezekowitz regarding article, "development and evolution of a hierarchical clinical composite end point for the evaluation of drugs and devices for acute and chronic heart failure: a 20-year perspective". *Circulation*. 2017;135 (15):e889-e891.
21. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth IMP. Developing and evaluating complex interventions: new guidance. Medical Research Council. https://www.mrc.ac.uk/documents/pdf/complex-interventions-guidance/. Published 2006. Accessed June 12, 2020.
22. Sjölander A, Vansteelandt S. Frequentist versus Bayesian approaches to multiple testing. *Eur J Epidemiol*. 2019;34(9):809-821.
23. Brown PM, Ezekowitz JA. Multitype events and the analysis of heart failure readmissions: illustration of a new modeling approach and comparison with familiar composite end points. *Circ Cardiovasc Qual Outcomes*. 2017;10(6):e003382.
24. Rogers JK, Jhund PS, Perez A-C, et al. Effect of Rosuvastatin on repeat heart failure hospitalizations: the CORONA trial (controlled Rosuvastatin multinational trial in heart failure). *JACC: Heart Fail*. 2014;2(3):289-297.
25. van Eijk RPA, Eijkemans MJC, Rizopoulos D, van den Berg LH. S N. comparing methods to combine functional loss and mortality in clinical trials for amyotrophic lateral sclerosis. *Clin Epidemiol*. 2018;10:333-341.
26. Mascha EJ, Sessler DI. Design and analysis of studies with binary- event composite endpoints: guidelines for anesthesia research. *Anesth Analg*. 2011;112(6):1461-1471.
27. Lee A, Chiu CH, Cho MWA, et al. Factors associated with failure of enhanced recovery protocol in patients undergoing major hepatobiliary and pancreatic surgery: a retrospective cohort study. *BMJ Open*. 2014;4(7):e005330.
28. Turan A, Grady M, You J, et al. Low vitamin D concentration is not associated with increased mortality and morbidity after cardiac surgery. *PLoS One*. 2013;8(5):e63831.
29. Senn S. Disappointing dichotomies. *Pharm Stat*. 2003;2(4):239-240.
30. Lewis JA. In defence of the dichotomy. *Pharm Stat*. 2004;3(2):77-79.
31. Wang Y, Zhang D, Du G, et al. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. *Lancet*. 2020;395(10236):1569-1578.
32. Herper M. Inside the NIH's controversial decision to stop its big remdesivir study. STAT Web site. https://www.statnews.com/2020/05/11/inside-the-nihs-controversial-decision-to-stop-its-big-remdesivir-study/. Published May 11, 2020. Accessed July 2, 2020.
33. Spinner CD, Gottlieb RL, Criner GJ, et al. Effect of Remdesivir vs standard care on clinical status at 11 days in patients with moderate COVID-19: a randomized clinical trial. *JAMA*. 2020. https://doi.org/10.1001/jama.2020.16349.
34. McCreary EK, Angus DC. Efficacy of Remdesivir in COVID-19. *JAMA*. 2020. https://doi.org/10.1001/jama.2020.16337.
35. Desai A, Gyawali B. Endpoints used in phase III randomized controlled trials of treatment options for COVID-19. *EClinicalMedicine*. 2020;23:100403.
36. Brown PM, Anstrom KJ, Felker GM, Ezekowitz JA. Composite end points in acute heart failure research: data simulations illustrate the limitations. *Can J Cardiol*. 2016;32(11):1356.e1321-1356.e1328.
37. Sun H, Davison Beth A, Cotter G, Pencina Michael J, Koch GG. Evaluating treatment efficacy by multiple end points in phase II acute heart failure clinical trials. *Circ Heart Fail*. 2012;5(6):742-749.
38. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. EMA/CHMP/ICH/436221/2017 Web site. https://www.ema.europa.eu/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf. Published February 17, 2020. Accessed February 12, 2020.
39. Wong JLC, Mason AJ, Gordon AC, Brett SJ. Are large randomised controlled trials in severe sepsis and septic shock statistically disadvantaged by repeated inadvertent underestimates of required sample size? *BMJ Open*. 2018;8(8):e020068.

How to cite this article: Brown PM, Rogne T, Solligård E. The promise and pitfalls of composite endpoints in sepsis and COVID-19 clinical trials. *Pharmaceutical Statistics*. 2021;20:413–417. https://doi.org/10.1002/pst.2070