Composite endpoints in health technology assessment: Part 1 – an illustration of best modeling practice

Journal of Comparative Effectiveness Research

Andrew Briggs*,[1] (iD), Aris Angelis[1,2,3] (iD), Jieling Chen[4], David Booth[5], Jason A Davis[5], Muthiah Vaduganathan[6] & Pardeep S Jhund[7] (iD)

[1]Department of Health Services Research & Policy, London School of Hygiene & Tropical Medicine, London, UK
[2]Department of Health Policy & LSE Health, London School of Economics & Political Science, London, UK
[3]General Secretariat for Strategic Planning, Ministry of Health, Athens, Greece
[4]Global Market Access & Pricing, AstraZeneca R&D, Gaithersburg, MD 20878, USA
[5]Health Economics & Outcomes Research Ltd, Cardiff, UK
[6]Cardiovascular Division & Center for Cardiometabolic Implementation Science, Brigham & Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[7]British Heart Foundation, University of Glasgow, Glasgow, UK
*Author for correspondence: andrew.briggs@lshtm.ac.uk

Composite endpoints amalgamate multiple clinical outcomes into a single measure, offering efficiency gains in clinical trials through increased event rates and reduced sample sizes, thus accelerating clinical development and regulatory approval. However, employing composite endpoints introduces complexities into health technology assessments (HTAs), particularly in economic modeling, due to the varying clinical significance and cost implications of the components. In this paper, we explore best modeling practice for HTAs that are based on clinical trials that employ composite endpoints. We examine regulatory guidance and discuss statistical solutions for differential component impacts, before presenting a case study based on a recent dapagliflozin submission for reimbursement in heart failure. Our investigation reveals that while composite endpoints can streamline trial analyses and hasten regulatory approval, they also pose a risk of bias in HTA if treatment effects for the components are inappropriately pooled. The paper discusses HTA principles in the context of composite endpoint trials and proposes strategies to develop modeling scenarios and interpret results, especially concerning whether to combine or split out estimates of component treatment effects. A particular focus is the accurate capture of uncertainty, both in terms of the parameter inputs to the model and over the ultimate decision to reimburse. This paper serves as a potential resource for researchers, practitioners and decision-makers, offering insights into best modeling practices that can unlock the full potential of composite endpoints in the pursuit of evidence-based healthcare decision-making.

Plain language summary
What is this article about? This article explores how composite endpoints – which combine multiple clinical outcomes into a single measure – are used in clinical trials and how they impact health technology assessment (HTA). Composite endpoints are widely used in clinical research to increase statistical power and speed up drug development, but they can create challenges in economic modeling when evaluating whether new treatments provide value for money. This paper reviews existing guidance from regulatory agencies and HTA bodies, presents statistical approaches to handling composite endpoints, and uses a case study of dapagliflozin for heart failure to illustrate key challenges and potential approaches to cost-effectiveness modeling.
What were the results? This study finds that while composite endpoints offer advantages for clinical trial efficiency, they present difficulties if treatment effects on individual components vary significantly. The case study of dapagliflozin highlights how different approaches to modeling composite endpoints can lead to different economic results, ultimately affecting reimbursement decisions:
- Using the composite endpoint as a single measure suggested dapagliflozin was highly cost-effective (i.e., resulted in greatest value for money).

Becaris PUBLISHING

- Separating the composite endpoint into its individual components showed similar results, but with greater uncertainty.
- Ignoring treatment effects on certain components (e.g., cardiovascular death) made dapagliflozin appear less cost-effective, despite trial data showing a possible benefit.

What do the results of the study mean? These findings highlight the importance of carefully handling composite endpoints to avoid biased conclusions. The study suggests that HTA evaluations should account for uncertainty in treatment effects on individual components, avoid exclusions of components based on statistical significance alone and align with best practices from regulatory agencies, while also ensuring economic models accurately reflect real-world clinical impact.

Overall, this paper provides practical guidance for researchers, policymakers and decision-makers on how to handle composite endpoints when presenting economic models to HTA decision-makers.

Composite endpoints, formed by aggregating multiple individual clinical endpoints into a common measure, have emerged as a valuable tool for the evaluation of healthcare interventions. Many clinical trials have embraced the use of composite endpoints [1]. The primary benefit of composite endpoints is the potential for greater efficiency (increased statistical power) in clinical trial design, either through combining several faster accruing nonfatal events or by combining nonfatal events with slower accruing fatal events [2,3]. This may allow for a more rapid completion of the trial, which is of importance given the resource constraints often seen in clinical research and development. Additionally, the use of a composite endpoint eliminates the need for multiplicity adjustments or other complex statistical approaches that are otherwise required when analyzing multiple individual endpoints separately. This streamlined approach may simplify statistical analyses and interpretation of trial results, potentially accelerating the translation of clinical evidence into practice. Meanwhile, the integration of both nonfatal and fatal events may better summarize the patient experience and provide clinicians with a summary estimate based on the totality of evidence.

In this context, the focus of this methodological paper (the first in a series of two) aims to provide guidance for conducting health technology assessments (HTAs) using data from regulatory clinical trials utilizing composite endpoints, with the express aim of exploring the value for money for products seeking reimbursement. Health economic models represent a key part of HTAs that are submitted by companies seeking reimbursement for their products. Identifying best practice approaches for modeling trials with composite endpoints is the main aim of this first manuscript.

The utilization of composite endpoints in clinical research is not without its challenges [4,5], and these challenges are magnified when considering their implications for HTA and associated health economic modeling of cost-effectiveness. There is an inherent assumption that all components of the composite endpoint hold equal value, yet this assumption rarely applies; for example, different clinical events may carry varying degrees of clinical significance, as well as having different impacts on patients and health service costs. Furthermore, treatments under evaluation may exert differential effects on each component of the composite endpoint, or may influence recurrent events as well as the first event. The relative timing of component events for the groups compared may differ in that nonfatal events will likely occur on average earlier in the time course of disease progression compared with fatal events. Understanding the relative importance of the individual components of a composite endpoint and the potential for differing treatment effects, is essential not only to inform clinical decision-making, yet also to inform cost-effectiveness analyses and HTA, as they directly influence the estimation of an intervention's value.

This methodological study summarizes the available guidance on the use of composite endpoints that has been offered by regulators and HTA bodies, as well as examining some of the proposed statistical solutions for addressing the hierarchy of components in a composite endpoint, the potential for components to be differentially impacted by treatment, and where treatment affects both first and recurrent events. A case study on dapagliflozin, a sodium-glucose co-transporter 2 inhibitor (SGLT2i), is also presented to illustrate the specific challenges for HTA and cost-effectiveness modeling. Finally, we discuss the principles of HTA and cost-effectiveness modeling in the context of composite endpoint trials and offer some guidance for how to develop modeling scenarios and interpret

the results, bearing in mind the potential conflict between the efficiency of employing a clinical endpoint and the danger of bias that can result if inappropriately employed. Throughout this study, the focus is to accurately capture the uncertainty in terms of the parameter inputs to the model, yet most importantly, assess the decision uncertainty arising from a model's outputs.

Good practice guidance from drug regulatory & HTA bodies

Adhering to good practice guidance from drug regulatory and HTA bodies is essential to ensure robust and reliable assessments of composite endpoints. This section highlights the key recommendations and guidance provided by prominent regulatory authorities and HTA organizations on the use of composite endpoints.

The EMA offers comprehensive guidance on the clinical evaluation of medicinal products, emphasizing the importance of composite endpoints in demonstrating the efficacy and safety of new treatments. EMA guidance on multiplicity in clinical trials includes a section on composite endpoints. This includes a number of important guidelines, such as the expectation that treatment should impact all components, that more clinically severe events (e.g., mortality) should be included alongside less severe components (e.g., hospitalization), and that one of the advantages of a composite endpoint is that it obviates the need for multiplicity testing. The EMA advocates that the components should be analyzed separately alongside the composite, and that evidence of differential treatment effects on the components will likely complicate the interpretation of results [6].

Similarly, the US FDA guidance on composite endpoints notes that [7]:

> "*The effect on the composite endpoint, however, will not be a reasonable indicator of the effect on all of the components or an accurate description of the drug's benefit if the clinical importance of different components is substantially different and the treatment effect is chiefly on the least important event.*"

The FDA also provides specific guidance for examining the components of a composite endpoint, including disaggregating a composite endpoint into individual components based on first event, where more than one of the component events occurs.

There is limited guidance on the use of composite endpoints from HTA bodies. The Pharmaceutical Benefits Advisory Committee (PBAC), Australia's advisory body on the cost-effectiveness of medicines, provides guidance on the evaluation of pharmaceuticals, which includes some limited guidance on the use of composite endpoints in terms of reporting, justifying and analyzing results [8].

> "*If one or more of the reported outcomes is a composite, discuss and compare the clinical importance of each of the components of the composite. Report whether the definition of the composite outcome was explicitly prespecified. Justify the inclusion of the components in the composite outcome, and the exclusion of any components that were considered but subsequently rejected. Disaggregate the composite outcome and present the results. . . Composite outcomes need to be appropriately handled when disaggregating the component outcomes so that the true estimate for each component outcome is appropriately captured.*" – Section 2.4.3 Outcomes, PBAC 2016 [8]

However, this information is reported in the 'Clinical Evaluation' section of the PBAC guidelines, and no direct guidance for how to interpret a composite endpoint from the perspective of the economic modeler is given.

Other prominent HTA bodies, such as the National Institute for Health and Care Excellence (NICE) in the UK, and Canada's Drug Agency, provide comprehensive methodological guidance on HTAs, though specific guidance on the use of composite endpoints is not provided in their relevant documents [9,10]. The Scottish Medicines Consortium (SMC) does not provide guidance on composite endpoints directly, though it does make the following specific request [11]:

> "*If the submitting company presents a base case analysis which includes some differences which were not statistically significant, within the sensitivity analysis they should include such an analysis with the non-significant differences removed.*" – Section 6.13.3 Uncertainty, SMC 2022 [11]

This guidance from SMC is presented without context but may be applicable to the components of a composite endpoint. It must be noted, however, that if this guidance is applied in this way, any hypothesis test would be underpowered by design; issues of power for detecting significant differences are not addressed in the approach described in Section 6.13.3.

Current guidelines identified in this study thus appear to devote limited attention specifically to the topic of managing results from clinical trials reporting composite endpoints, although the use of such endpoints is extensive.

Eventually, guidance will be needed on the development and application of methods relating to establishing a link between treatment response based on composite endpoints and health outcomes for patients in the context of valuation for HTA assessment [12].

Existing statistical approaches for analysis of composite endpoints

The aggregation of multiple individual clinical endpoints into a composite measure offers statistical advantages in terms of increasing events and, therefore, power for statistical testing, while avoiding the complexities associated with multiplicity testing. Nevertheless, as identified in regulatory guidance, the use of composite endpoints creates new challenges around the value of the component endpoints and the possibility that treatment effects could vary by component [13]. A number of statistical methods have been developed relating to composite endpoints, many of which are described in a recent review by Baracaldo-Santamaria *et al.,* which investigated the interpretation of composite endpoints in clinical research [4].

These statistical methods can be categorized according to the challenge they seek to address, including the hierarchical importance of the components of the composite (e.g., the Win-ratio [Pocock *et al.* [14]]) or weighted composite endpoints (e.g., Ozga and Rauch [15]), testing for heterogeneity in the treatment effect for components (e.g., Pogue *et al.* 2010 and 2012 [16,17]; Ristl *et al.* 2019 [18]), and methods that allow for repeated measures for nonfatal endpoints (e.g., Andersen–Gill model, negative binomial model). Another categorization of analysis methods for comparing 'novel' versus 'traditional' approaches can be based on time to first event, composite event process and pairwise hierarchical comparisons [2]. All of these methods focus on the statistical testing of the composite, while making an allowance for the differential weights that might be applied to the components. However, health economic modeling requires a specific set of weights based on the quality-adjusted life year (QALY) and health service resource cost consequences of the components, and focuses on the estimation of cost-effectiveness not the testing of hypotheses.

Furthermore, a common feature among these statistical procedures is the requirement for individual participant data. Such data are rarely available to the analyst performing HTA for all the studies pertaining to a decision problem, and as a result, these methods are not often used in practice to inform policy-making. Consequently, visual inspection of the results for the individual components remains the main evaluation approach, which can increase the level of subjectivity in the choice as to whether to pool or split the component endpoints. In addition, none of the above methods addresses the three main challenges associated with the use of composite endpoints (relative value of each composite component, heterogeneity of effects by component endpoint, and accommodation of repeat measures for recurrent events) simultaneously, which is necessary in most HTA submissions. For this reason, we do not explore these specific methods further in this paper as we consider they are not appropriate for health economic modeling.

Importance of representing uncertainty in HTA submissions

Generally, existing guidance from reimbursement authorities and other good practice guides for modeling emphasizes that all evidence relating to a decision problem should be included in an HTA. Furthermore, since HTA-relevant outcomes (e.g., healthcare resource utilization and costs per value gained) are rarely the primary endpoints of clinical studies, the focus of health economic modeling should be on the estimation of relevant outcomes with appropriate representation of uncertainty, rather than on testing of hypotheses. Where possible, structural assumptions relating to a health economic model should be parameterized, and uncertainty should be propagated through the model structure from the input parameters to the output parameters using probabilistic sensitivity analysis. This allows the representation of uncertainty not only in terms of uncertainty intervals (95% credible interval often preferred), but also in terms of the expected value of perfect information (EVPI), a Bayesian decision analysis procedure that integrates the loss function associated with incorrect decision making with the probability of making the wrong decision [19]. The EVPI can be thought of as the cost of incorrect decision making.

Where structural uncertainties cannot be parameterized (such that they are not represented within the estimated uncertainty intervals of cost-effectiveness outcomes), the preferred approach is to present scenarios for alternative modeling assumptions with associated uncertainty intervals [9]. Within the manuscript, we use a case study (described below) to illustrate how the approach to modeling composite endpoints or the components of those endpoints in health economic models can have a profound impact on the associated uncertainty of the model. Further, we consider this to be a very important form of structural uncertainty for the proposed model.
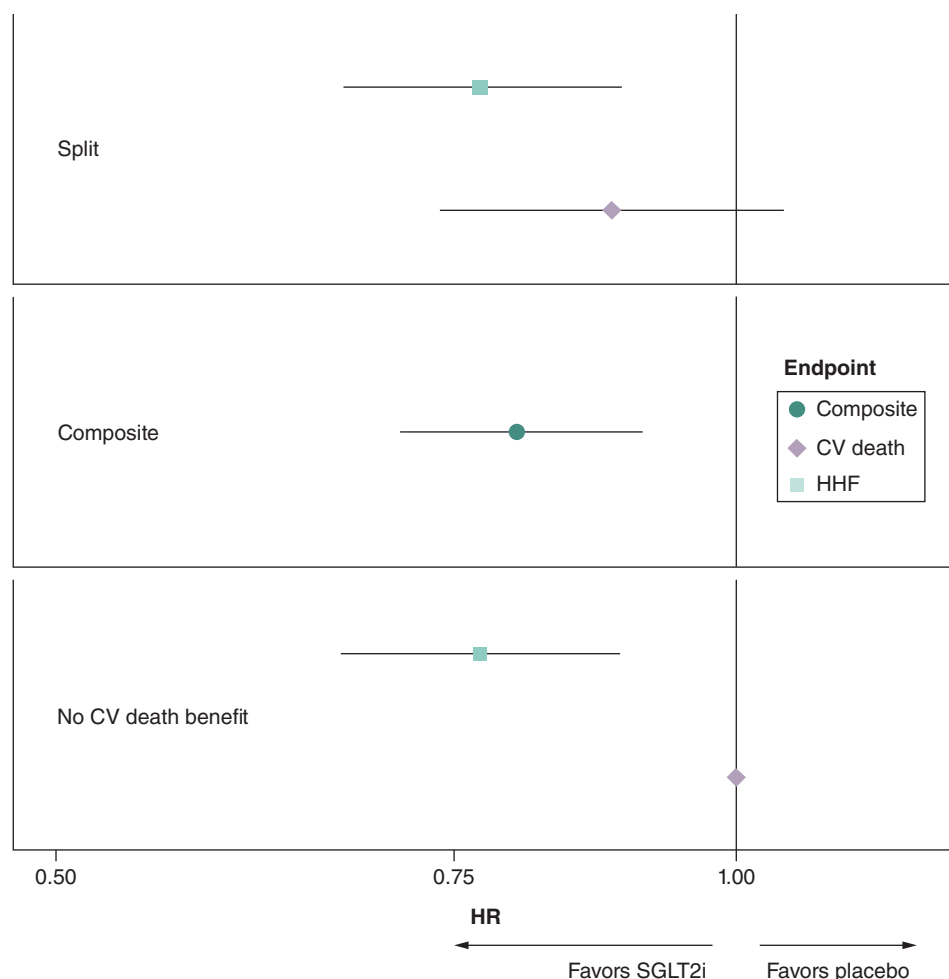
Figure 1.   Hazard ratio results from the DELIVER clinical trial in terms of the treatment effect on the composite endpoint and its components.
CV: Cardiovascular; HHF: Hospitalization for heart failure; HR: Hazard ratio; SGLT2i: Sodium-glucose co-transporter 2 inhibitor.

Case study – dapagliflozin for heart failure with preserved or mildly reduced ejection fraction (HFpEF/HFmrEF)

Background & decision making context

In this section, we describe the comparative effectiveness and cost-effectiveness analysis of dapagliflozin (an SGLT2i) plus standard of care, versus placebo plus standard of care, for the treatment of heart failure with preserved or mildly reduced ejection fraction (HFpEF/HFmrEF), a decision problem previously submitted to NICE for evaluation [20].

The key study informing this reimbursement decision was the DELIVER trial [21], a Phase III, placebo-controlled trial, in adult patients with symptomatic chronic HFpEF/HFmrEF. The primary outcome in the DELIVER trial was a composite of worsening HF (defined as either an unplanned hospitalization for HF (HHF) or an urgent visit for HF) or cardiovascular (CV) death. The primary composite endpoint used in the DELIVER trial encompasses both fatal and nonfatal events as encouraged by regulators, and thus guarantees that the component endpoints have different perceived values (whether from a clinical, patient or economic perspective) since fatal and nonfatal events are associated with different values.

Treatment with dapagliflozin added to standard of care was shown to reduce the composite endpoint when compared with placebo added to standard of care (hazard ratio [HR]: 0.82, 95% confidence interval [CI]: 0.73–0.92), shown in Figure 1 as the 'Composite' scenario. However, when the components of the composite were analyzed separately, there was uncertainty as to whether treatment with dapagliflozin reduced both events equally. For worsening heart failure, the treatment effect of dapagliflozin was statistically significant (HR: 0.79, 95% CI:

0.69–0.91) compared with placebo, but for CV death, the treatment effect was not statistically significant (HR: 0.88, 95% CI: 0.74–1.05), as shown in Figure 1 as the 'Split' scenario. Subsequently, the evidence assessment group, an independent academic group tasked with assessing company evidence submissions, suggested to present a scenario which excluded mortality effect from the model, meaning only the treatment effect of dapagliflozin on HHF remained. This scenario is represented by the 'no CV death Scenario' shown in Figure 1. The impact of such a scenario is important to consider – not only is the mortality effect removed, but such a scenario also assumes that the lack of treatment effect is known with certainty. Although this seems consistent with the quote from the SMC about presenting scenarios with nonsignificant treatment effects removed, it also runs counter to the advice from NICE [9] that structural uncertainties should be parameterized within a model. The impact of this structural uncertainty choice, to either include the possibility of a treatment effect on death with all the inherent uncertainty, or to exclude the possibility of a treatment effect on death assuming that this is known with certainty, is an important point of contrast for this manuscript.

Based on the modeling scenarios presented to the NICE technology appraisal committee, it was considered that there were substantial uncertainties in the economic modeling. As a result, the initial decision by the NICE committee concluded that the cost-effectiveness estimates for dapagliflozin are likely higher than what NICE considers a cost-effective use of National Health Service (NHS) resources [20]. Therefore, the initial draft guidance issued by the committee did not recommend dapagliflozin within its marketing authorization for the treatment of chronic HFpEF/HFmrEF in the NHS. However, following the provision of additional evidence during the consultation period, the committee issued a final appraisal document which recommended dapagliflozin within its marketing authorization as an option for treating symptomatic chronic HFpEF/HFmrEF in adults [22]. This revised recommendation was largely based on the recognition that scenarios that included a direct and/or indirect dapagliflozin treatment effect on cardiovascular and all-cause deaths were more plausible, despite a level of uncertainty still existing. When this preferred assumption was incorporated, the cost-effectiveness estimates were subsequently below the threshold of what NICE considers an acceptable use of NHS resources; thus, the intervention was recommended.

Modeling challenges & illustration

The full submission model presented to NICE involved an individual participant data analysis of the DELIVER trial informing a state transition model that, in its original form, split out the components of the composite endpoint and applied different treatment effects [23]. In the illustrative analysis described here, some minor simplifying assumptions were made to allow the direct application of the treatment effects depicted in Figure 1 to the economic model. Specifically, from the full submission model, only unadjusted HF event risk and survival models were used. For survival, the Weibull distribution was selected, and the shape and scale parameters derived from the DELIVER trial analysis were retained, but the coefficient for the treatment arm was substituted with log-transformed HRs according to the applicable analysis scenario. Similarly, in the unadjusted models to predict recurrent HF events, the intercept was kept, while the treatment arm coefficient was substituted with the log-transformed treatment HRs. In both cases, uncertainty parameters were estimated using the reported HR confidence intervals. Although not identical to the analysis submitted to NICE, the figures presented in this illustration are comparable to those in the full NICE analysis and serve the purpose of exemplifying the modeling challenges faced.

In the illustration, the cost-effectiveness model was interrogated with respect to the three different scenarios shown in Figure 1:

1. Split scenario: in this scenario, treatment effects were separately applied for both HHF and CV death based on the observed treatment effects from DELIVER.
2. Composite scenario: in this scenario, the same treatment effect was assumed for both HHF and CV death, that being the treatment effect on the composite outcome from DELIVER.
3. No CV death scenario: this scenario assumed there was a treatment effect on HHF as observed in DELIVER, but that there was no CV death treatment effect (and no uncertainty regarding the lack of treatment effect. This assumption represents a change from the results observed in the DELIVER trial.)

The results of applying these three scenarios are shown in Table 1, which summarizes the estimated ICER, the incremental net monetary benefit ([INMB], assuming a decision threshold of £20,000 per QALY), and the EVPI. The 95% credible intervals were calculated based on 1000 replicates of probabilistic sensitivity analyses. The INMB results are also presented as a forest plot in Figure 2.

Table 1. Cost-effectiveness results for three modeling scenarios.				
Scenario	ICER	INMB	(95% CI INMB)	EVPI
Split	£10,256	£1697	(-£4510 to £7904)	£609
Composite	£7636	£3306	(-£952 to £7563)	£66
No CV death	£27,152	-£426	(-£1227 to £376)	£36

CI: Confidence interval; CV: Cardiovascular; EVPI: Expected value of perfect information; ICER: Incremental cost-effectiveness ratio; INMB: Incremental net monetary benefit.
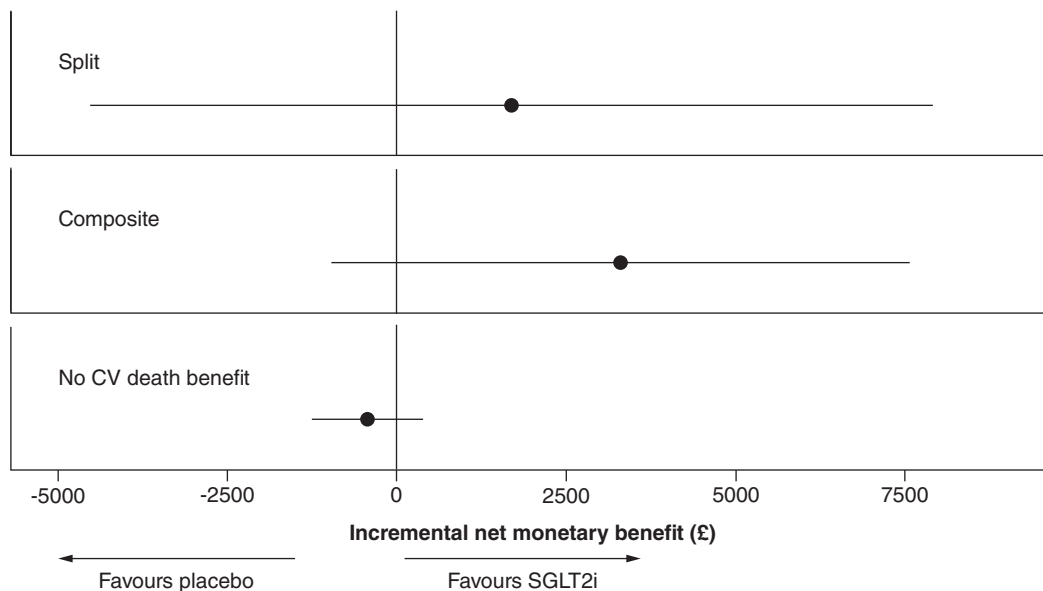

Figure 2.   Incremental net monetary benefit for three modeling scenarios.
CVD: Cardiovascular death; SGLT2i: Sodium-glucose co-transporter 2 inhibitor.

Results presented in Table 1 & Figure 2 demonstrate that the choice of how to incorporate treatment effects from a composite endpoint has important implications for both the point estimate of cost-effectiveness and the associated uncertainty. The most favorable cost-effectiveness outcomes were observed when using the composite scenario, with an ICER of approximately £7600 per QALY gained and the greatest INMB of all scenarios. The split scenario yielded a slightly increased ICER of approximately £10,000 per QALY gained, and the scenario with no CV death was associated with the least favorable cost-effectiveness outcomes, with an ICER of approximately £27,000 per QALY gained and a negative INMB when using a decision threshold of £20,000 per QALY gained (Figure 2). Considerable differences were observed between the point estimates for the three scenarios, but strikingly, although derived from the same trial data source with its inherent uncertainty, the uncertainty intervals estimated for the scenario point estimates also vary considerably. As expected, the uncertainty interval on INMB was marginally smaller for the composite scenario compared with that of the split scenario, although both are of comparable magnitude. Notably, for the no CV death scenario, the uncertainty interval on INMB was markedly reduced. This is due to the implicit assumption that the absence of a treatment effect on CV death was known with complete certainty. This assumption also influenced the EVPI results (which can be interpreted as the cost of decision uncertainty), where the EVPI similarly decreased for the no CV death scenario when compared with the other scenarios.

As previously noted, the assumption that there is an absence of treatment effect on CV death does not reflect the clinical trial results. This is illustrated in Figure 3, showing the model predictions (aggregated across treatment and no-treatment groups) under each of the three scenarios compared with the Kaplan–Meier failure curves from the DELIVER trial. As anticipated, the composite scenario slightly underestimated CV death, though the split scenario demonstrated a very close match between the model predictions and the observed DELIVER data. The no CV death scenario systematically overestimated CV death in the model.
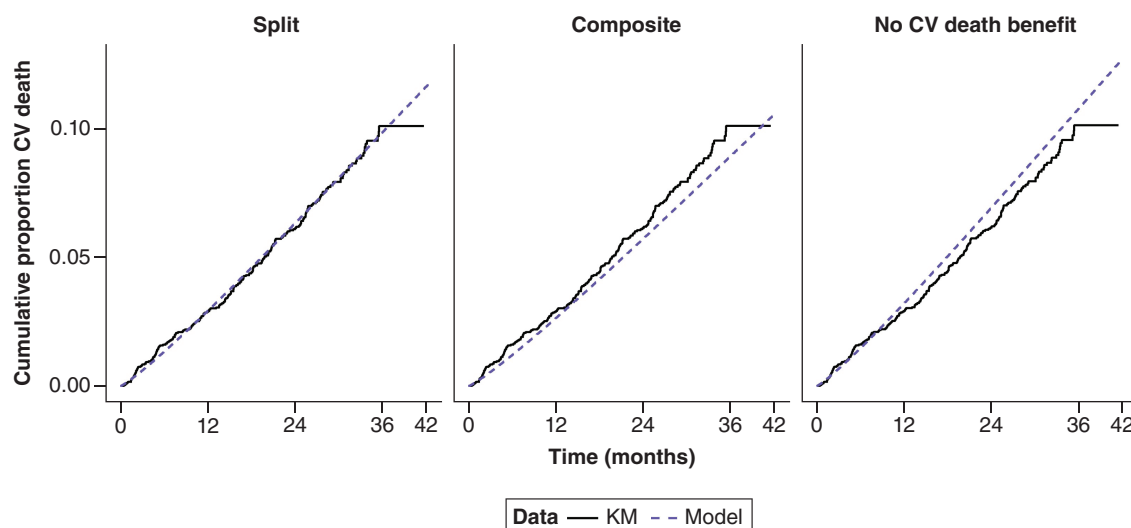
Figure 3.   Observed Kaplan–Meier estimates of cumulative cardiovascular deaths compared with model estimates of cardiovascular deaths for three modeling scenarios.
CV: Cardiovascular; KM: Kaplan–Meier.

Discussion & limitations

This study summarizes examples of reported guidance on the use of composite endpoints from drug regulatory agencies and highlights the potential practical usefulness of composite endpoints for licensing studies. However, the use of composite endpoints is less straightforward when it comes to HTAs and cost-effectiveness models to support reimbursement decisions; therefore, further guidance is needed. Using an illustration based on a recent submission to NICE in the UK, this study showed that the interpretation of the treatment effects on the components of a composite endpoint can have a profound impact on the estimated cost-effectiveness and associated uncertainty, which could influence reimbursement decisions.

A common assumption suggests that it is necessary to disaggregate composite endpoints for health economic modeling exercises, given that the components of the composite are valued differently (in terms of healthcare resource use and/or the health outcome consequences); however, this is not required in all circumstances. Health economic models synthesize the results of many input parameters into a single metric, such as an ICER or an INMB, to inform decision-makers on the estimated value-for-money of an intervention. Ultimately, therefore, all parameters of the model are synthesized together. If either the value of the components of the composite endpoint or the treatment effect on the components of the composite is equal, it would be legitimate to use the treatment effect on the composite endpoint in the health economic model. However, if neither of these conditions is met, the use of a composite endpoint in the health economic model would be biased compared with disaggregating the components and applying differential treatment effects. Nevertheless, it remains a somewhat subjective assessment of whether or not the assumption of equality of treatment effects is met, given that any formal statistical test is underpowered.

Using the example of dapagliflozin, presented in this study as an illustration, it was evident that the value of HHF and CV death outcomes was not equal in our subjective opinion. HHF is more costly in terms of healthcare resource use when compared with CV death, but CV death is associated with greater QALY loss. The results observed for the split scenario and composite scenario in Figure 1 suggest that the treatment effects on the two components (HHF and CV death) are not dissimilar. For example, both components were in the same numerical direction (favoring dapagliflozin), and both CIs overlapped the point estimate of the other component. Similar findings were observed for the cost-effectiveness analysis (Table 1 & Figure 2), where the results for the split and composite scenarios were relatively comparable and had similar CI ranges, though there was a numerical advantage for the composite scenario, which gave a slightly increased treatment effect for CV death (Figure 3). On this basis, it may be reasonable to suggest that an appropriate base case analysis for the cost-effectiveness analysis of dapagliflozin would be the split scenario; this was ultimately the preferred scenario by the NICE committee when making the final appraisal decision.

Importantly, we do not consider it appropriate, nor reasonable, to base the cost-effectiveness modeling of dapagliflozin on a scenario that assumes there is no treatment effect on CV death. The lack of significance observed on CV death arises from the decision to model the components of the composite endpoint individually, which leads to any hypothesis testing of those components being underpowered. This argument is similar to that used to argue against concluding that a subgroup treatment effect exists only for a subgroup where the CI excludes the null, while the effect is absent in a subgroup where the null is included in the CI [24,25]. This fallacious subgroup interpretation has been widely criticized in evidence-based medical literature and is implied by the well-known adage that 'absence of evidence is not evidence of absence'. Instead, the appropriate way to undertake subgroup analysis is to employ an appropriate interaction test [25,26]. Similarly, the appropriate way to conclude whether the treatment effects of the components of a composite endpoint are different would be to employ an appropriately powered statistical test of heterogeneity. However, given the lack of power that may exist for the components of a composite endpoint (assuming the trial was powered for the composite), such a statistical test can only confirm the existence of a difference and is unable to confirm the lack of a difference. Therefore, we consider it prudent to employ another popular adage that the purpose of the cost-effectiveness analysis is "*estimation not hypothesis testing*" [27]. When estimating the cost-effectiveness of a new intervention, the analyst must consider all uncertainties in the model. The no CV death scenario fails to recognize the inherent uncertainty that treatment with dapagliflozin could likely impact CV death. Indeed, this no CV death scenario suggested by the evidence assessment group contradicts the NICE guidance, which states that structural uncertainty should be parameterized in the model where possible [9]. The request to remove CV death introduces a structural certainty to the model, falsely implying that it is known that treatment cannot impact CV death. Furthermore, dismissing the potential mortality benefit of a treatment by handling the composite endpoint inappropriately, as illustrated in the case study presented in this manuscript, could misinform clinicians and have negative consequences for patients.

Conclusion

In summary, the utilization of composite endpoints in clinical research is common practice to enhance the efficiency of clinical research and development and the relevance of healthcare evaluations. However, their incorporation into the HTA process requires a nuanced understanding of their advantages and challenges, especially in the context of health economic modeling. This paper serves as a potential resource for researchers, practitioners and decision-makers, offering insights into best modeling practices that can unlock the full potential of composite endpoints in the pursuit of evidence-based healthcare decision-making.

Executive summary

- Composite endpoints, formed by aggregating multiple individual clinical endpoints into a common measure, afford statistical advantages for analysis of clinical trial data.
- Regulatory authorities emphasize the importance of their use to demonstrate the efficacy and safety of new treatments, but health technology assessment (HTA) bodies provide scant guidance.
- Analyses to inform HTA evaluation may make use of efficacy results for the composite endpoint or disaggregated by component, and if disaggregated, may further selectively consider which effects to include based on statistical measures of uncertainty.
- According to HTA good practice, all evidence relating to a decision problem should be included, and the focus should be on estimation of relevant outcomes with appropriate representation of uncertainty rather than hypothesis testing.
- A case study is presented to assess scenarios of use of a composite, its components and selective inclusion of component effects to illustrate the specific challenges and the danger of bias.
- Health economic results and uncertainties can be characterized through incremental cost-effectiveness ratios, probabilistic sensitivity analyses, net monetary benefit and the expected value of perfect information.
- Disaggregation of the composite was found to most faithfully reproduce treatment effects and uncertainty observed in the source clinical trial; in contrast, a proposed scenario of disregarding a statistically non-significant component yielded the poorest propagation of uncertainty and representation of trial results.
- Treatment effect by components of a composite should be employed unless those components are of equal value or experience the same treatment effect and consideration of uncertainty should form part of the analysis.

References

Papers of special note have been highlighted as: ● of interest; ●● of considerable interest

1. Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *Br. Med. J.* 341, c3920 (2010).

2. Mao L, Kim K. Statistical models for composite endpoints of death and non-fatal events: a review. *Stat. Biopharm. Res.* 13(3), 260–269 (2021).

●   **A topical review of relevance to the current manuscript subject matter.**

3. Weintraub WS. Statistical approaches to composite endpoints. *JACC Cardiovasc. Interv.* 9(22), 2289–2291 (2016).

4. Baracaldo-Santamaría D, Feliciano-Alfonso JE, Ramirez-Grueso R, Rojas-Rodríguez LC, Dominguez-Dominguez CA, Calderon-Ospina CA. Making sense of composite endpoints in clinical research. *J. Clin. Med.* 12(13), 4371 (2023).

5. Montori VM, Permanyer-Miralda G, Ferreira-González I *et al.* Validity of composite end points in clinical trials. *Brit. Med. J.* 330(7491), 594–596 (2005).

6. European Medicines Agency. Multiplicity issues in clinical trials – scientific guideline. (2017). Available from: https://www.ema.europa.eu/en/multiplicity-issues-clinical-trials-scientific-guideline (Accessed: 9 April 2024).

7. Food and Drug Administration. Multiple endpoints in clinical trials. *Guidance for Industry*, (2022). Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry (Accessed: 9 April 2024).

8. Pharmaceutical Benefits Advisory Committee. Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee (2016). Available from: https://pbac.pbs.gov.au/content/information/files/pbac-guidelines-version-5.pdf (Accessed: 9 April 2024).

9. National Institute of Health and Care Excellence. NICE health technology evaluations: the manual. *Process and Methods (PMG36)*, (2022). Available from: https://www.nice.org.uk/process/pmg36 (Accessed: 15 February 2024).

●● **An extensive guide published by a UK government agency that outlines their preferred modeling/analytical approaches.**

10. Canada's Drug Agency. Guidelines for the economic evaluation of health technologies: Canada [4th Edition] (2017). Available from: https://www.cadth.ca/guidelines-economic-evaluation-health-technologies-canada-4th-edition (Accessed: 16 April 2024).

11. Scottish Medicines Consortium. Guidance to Submitting companies for completion of New Product Assessment Form (NPAF) (2023). Available from: https://www.scottishmedicines.org.uk/media/6229/20230810-guidance-supplement-interim-acceptance-v1.pdf (Accessed: 9 April 2024).

12. Gavan SP, Bruce IN, Payne K *et al.* Valuing health gain from composite response endpoints for multisystem diseases. *Value Health* 26(1), 115–122 (2023).

13. Rauch G, Kieser M. An expected power approach for the assessment of composite endpoints and their components. *Comp. Stat. Data Anal.* 60, 111–122 (2013).

14. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur. Heart J.* 33(2), 176–182 (2012).

15. Ozga AK, Rauch G. Weighted composite time to event endpoints with recurrent events: comparison of three analytical approaches. *BMC Med. Res. Methodol.* 22(1), 38 (2022).

16. Pogue J, Devereaux PJ, Thabane L, Yusuf S. Designing and analyzing clinical trials with composite outcomes: consideration of possible treatment differences between the individual outcomes. *PLoS ONE* 7(4), e34785 (2012).

17. Pogue J, Thabane L, Devereaux PJ, Yusuf S. Testing for heterogeneity among the components of a binary composite outcome in a clinical trial. *BMC Med. Res. Methodol.* 10(1), 49 (2010).

18. Ristl R, Urach S, Rosenkranz G, Posch M. Methods for the analysis of multiple endpoints in small populations: a review. *J. Biopharm. Stat.* 29(1), 1–29 (2019).

19. Rothery C, Strong M, Koffijberg HE *et al.* Value of information analytical methods: report 2 of the ISPOR value of information analysis emerging good practices task force. *Value Health* 23(3), 277–286 (2020).

•   **Relevant background on value of information analysis in the context of cost-effectiveness analysis. A relevant economic characterization of uncertainty in model results.**

20. National Institute for Health and Care Excellence. Dapagliflozin for treating chronic heart failure with preserved or mildly reduced ejection fraction (2023). Available from: https://www.nice.org.uk/guidance/ta902/documents/129 (Accessed: 9 April 2024).

21. Solomon SD, McMurray JJV, Claggett B *et al.* Dapagliflozin in heart failure with mildly reduced or preserved ejection fraction. *N. Engl. J. Med.* 387(12), 1089–1098 (2022).

22. National Institute for Health and Care Excellence. Dapagliflozin for treating chronic heart failure with preserved or mildly reduced ejection fraction. *Final Draft Guidance*, (2023). Available from: https://www.nice.org.uk/guidance/ta902/documents/html-content-5 (Accessed: 9 April 2024).

•   **NICE assessment of dapagliflozin in the specified heart failure indication, the subject of the current manuscript case study.**

23. Booth D, Davis JA, McEwan P *et al.* The cost-effectiveness of dapagliflozin in heart failure with preserved or mildly reduced ejection fraction: a European health-economic analysis of the DELIVER trial. *Eur. J. Heart Fail.* 25(8), 1386–1395 (2023).

24. Altman DG, Matthews JN. Statistics notes. Interaction 1: heterogeneity of effects. *Br. Med. J.* 313(7055), 486 (1996).

•   **One of a series of seminal studies describing principles to be upheld in health technology assessment (HTA) analyses.**

25. Matthews JN, Altman DG. Statistics notes. Interaction 2: compare effect sizes not p values. *Br. Med. J.* 313(7060), 808 (1996).

•   **One of a series of seminal studies describing principles to be upheld in HTA analyses.**

26. Matthews JNS, Altman DG. Statistics notes: Interaction 3: How to examine heterogeneity. *Br. Med. J.* 313(7061), 862 (1996).

•   **One of a series of seminal studies describing principles to be upheld in HTA analyses.**

27. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br. Med. J. (Clin. Res. Ed.)* 292(6522), 746–750 (1986).